UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Analíticos Visuales en el Descubrimiento de Conocimiento de las Enfermedades Crónicas no Transmisibles en el Ecuador

Visual Analytics in Knowledge Discovery of Non Transmissible Chronic Diseases in Ecuador

Tesis presentada para optar al título de Magister de la Universidad de Buenos Aires en Explotación de Datos y Descubrimiento del Conocimiento

Eduardo Montero

Director de Tesis: Claudio Delrieux

Buenos Aires, 2019

Técnicas de Visualización

Resumen

El siguiente trabajo presenta un aplicativo de visualización de datos y analíticos visuales que brinda funcionalidades de análisis exploratorio y analíticos visuales al análisis de la encuesta nacional de salud y nutrición *ENSANUT* desarrollada en el 2012 en Ecuador, permitiendo entre otras cosas detectar nuevos patrones en la vigilancia del estado de salud y nutrición en Ecuador, con especial énfasis en las enfermedades crónicas no transmisibles (diabetes e hipertensión).

Para la construcción de este aplicativo se utilizó como insumo principal los resultados de la encuesta ENSANUT publicados en la página del Instituto Nacional de Estadísticas y Censeos INEC. Se utilizaron las bases de datos relacionadas con las enfermedades crónicas no transmisibles, el consumo alimentario, el estado de los macro y micronutrientes; así como los aspectos bioquímicos y antropométricos de Ecuador.

Mediante la metodología CRISP se construyó un proceso de extracción, transformación y carga utilizando SQL Server y R Studio. Se aplicaron técnicas tales como análisis de componentes principales, análisis de correspondencias, análisis de conglomerados, árboles de decisión, entre otras. Mismas que permitieron generar la estructura de datos necesaria para la creación del aplicativo de analíticos visuales utilizando Qliksense y aplicando técnicas como *overview, zoom and filter, brushing, details on demand.*

El trabajo se enfocó en realizar un análisis a nivel de variable y a nivel poblacional. Los resultados obtenidos a nivel de variable arrojaron, que las variables más importantes en la diabetes son la glucosa, homa, triglicéridos, la edad, la cintura, ferritina, resistencia a la insulina, estado nutricional, colesterol, peso y presión diastólica. Mientras que para la hipertensión son la presión sistólica y diastólica, peso, cintura, edad, IMC, triglicéridos, ferritina, talla, colesterol, hemoglobina, glucosa.

Los resultados obtenidos a nivel poblacional indicaron que existe una alta correlación entre las variables asociadas a la diabetes e hipertensión y la ubicación geográfica. Puesto que, considerando los aspectos antropométricos, las provincias de la Amazonía tienen a poseer los niveles más bajos de imc, cintura, presión sistólica y diastólica. Mientras que los valores más altos de estas variables se encuentran en las provincias restantes, diferenciadas entre sí por los contrastes entre los valores de imc y cintura versus presión sistólica y diastólica.

Por otro lado, considerando los aspectos bioquímicos se encontró que todas las provincias de la región amazónica tienen a poseer los valores más bajos de colesterol, presión, insulina, glucosa y triglicéridos. Mientras que los valores más altos de estas se encuentran en las provincias restantes, diferenciadas entre sí por cuan altos son estos valores.

Palabras Clave: visualización, minería de datos, árboles de decisión, visual analytics,

Visualization Techniques

Abstract

The following work presents a data visualization app and visual analytics which brings exploratory analysis and visual analytics to the health and nutrition national survey (ENSANUT) developed in 2012 in Ecuador. Allowing, among other things, detecting new patterns in the vigilance of health and nutrition status in Ecuador with special focus on non-transmissible chronic diseases (diabetes and hypertension.)

The results of ENSANUT survey published on the National Institute of Census and Statistics INEC page were used as the main income in the construction of this app. At the same time, databases related with non-transmissible chronic diseases, food consumption, macro and micronutrients state and biochemical and anthropometrics aspects were considered as well.

An ETL process was constructed guided by CRISP methodology and using SQL Server, R Studio and techniques such as principal components analysis, correspondence analysis, clustering analysis and decision trees. The goal of this process was generate the data structure which was used to construct the data visualization app and visual analytics using Qliksense and applying visualization techniques such as overview, brushing, zoom and filter, details on demand.

This work was focused on developing an analysis at variable and population level. The principal results at variable level indicated that the most important variables in diabetes were glucose, homa index, triglycerides, year, waist diameter, ferritin, insulin resistance, nutritional state, cholesterol, weight and diastolic pressure. On the other hand, the most important variables related with hypertension were systolic and diastolic pressure, weigh, waist diameter, year, IMC, triglycerides, ferritin, height, cholesterol, hemoglobin, glucose.

The results obtained at population level revealed that there is a high correlation between anthropometric variables and geographic location since the Amazonian provinces (except Sucumbios) tend to have the lowest levels of imc, waist diameter, systolic pressure and diastolic pressure. However, the other provinces have the highest values of these variables differentiated between them by the level of imc and waist diameter against systolic pressure and diastolic pressure.

Finally, considering the biochemical aspects was found that all Amazonian provinces tend to have lowest values of cholesterol, systolic pressure, diastolic pressure, insulin, glucose and triglycerides. However, the other provinces have the highest values of these variables differentiated between them by how high these levels are.

Keywords: Visualization, Data Mining, Decision Trees, Visual Analytics,

Indice

Listado de Figuras

Listado de Tablas

1. Introducción

1.1. Contexto del Ecuador

Ecuador es un país sudamericano que limita al norte con Colombia, al sur y este con Perú y al oeste con el océano Pacífico. Es mayoritariamente católico muy diverso y heterogéneo envuelto en contexto social complejo de subdesarrollo y pobreza. Tiene una extensión de 256.370 kilómetros cuadrados.

Adicionalmente, Ecuador es uno de los países con la más alta concentración de ríos y de mayor biodiversidad por km^2 [1] [2]. Está dividido en cuatro regiones muy diversas y claramente diferenciadas, costa, sierra, oriente y región insultar, en las que se distribuyen 24 provincias [3] (Anexo 1).

Ecuador es un país pluriétnico y multicultural. Las etnias que lo conforman son los pueblos indígenas, montubio, mestizos, afroecuatorianos y migrantes. En la época de la conquista española los pueblos indígenas de la Sierra fueron sometidos rápidamente en el siglo XVI. Aunque diezmados por la violencia, esclavitud y enfermedades lograron subsistir.

Algunos pueblos de la Costa fueron sometidos mientras que otros se adentraron a tierras a las cuales los conquistadores accedieron varios siglos después. Por otro lado, los pueblos amazónicos nunca fueron conquistados y su contacto con la sociedad dominante fue reducido [1].

La costa es un territorio conformado por llanuras, colinas, cuencas sedimentarias y elevaciones de poca altitud. Está compuesta de seis provincias Guayas, Manabí, Esmeraldas, El Oro, Los Ríos y Santa Elena. Todas las provincias mencionadas, a excepción de Los Ríos tienen salida al mar. En esta región se encuentra la red fluvial más grande del país: la Cuenca del río Guayas que consta de 12 afluentes [3].

La sierra es la región compuesta por la cordillera de los Andes que atraviesa el país de sur a norte dando lugar a hoyas y valles a lo largo del callejón interandino. La sierra abarca 11 provincias que tienen importantes elevaciones montañosas, tales como Chimborazo, Cotopaxi, Cayambe, Antisana, Altar, entre otros. En esta región se encuentra Quito, la capital del país [3] (Anexo 1).

Por otro lado, el oriente ecuatoriano está localizado en la parte oriental del país y se extiende desde el extremo oriental de los Andes hasta las llanuras del río Amazonas. Está compuesta por seis provincias caracterizadas por la diversidad de su flora y fauna y por sus extensos ríos, entre los cuales se destacan el Putumayo, el Napo y el Pastaza. Hay dos regiones geográficas Alta Amazonía y la llanura Amazónica [3]. (Anexo 1).

La región insular o Islas Galápagos es un archipiélago conformado por 13 islas principales, se encuentra ubicado a casi mil kilómetros del continente, reconocida a nivel mundial por la

cantidad de especies endémicas que en ella habitan, las mismas que inspiraron a Charles Darwin a desarrollar su teoría de selección natural [3] [4].

Ciudades Principales

El desarrollo del país está mayormente determinado por dos ciudades. Por un lado Quito, la capital política y administrativa del país, localizada en la provincia de Pichincha. Las provincias aledañas a esta son Santo Domingo de los Tsáchilas, Cotopaxi, Napo, Sucumbíos e Imbabura. Por otro lado está Guayaquil, considerada la capital económica, localizada en la provincia del Guayas las provincias circundantes a esta son Santa Elena, Manabí, Los Ríos, Cañar, Azuay y El Oro. Algunas características de estas ciudades se detallan a continuación.

Quito

Quito, capital de la Republica, está ubicada al centro norte del país. En la actualidad es la ciudad más poblada del país. Sin embargo cuando se realizó la Encuesta Nacional de Salud y Nutrición (2012) la ciudad de Guayaquil ocupaba este lugar [7]. Quito fue denomina la "ciudad eclesiástica de América" (Gómez, 1997: 47), debido a la significativa concentración de iglesias y órdenes religiosas y al poder e influencia que han ejercido y ejercen en esta sociedad. [5]

Localizada en una meseta ubicada en las faldas del volcán Pichincha, la posición de la ciudad es muy cercana a la línea ecuatorial. Clima subtropical de altura que se caracteriza por temperaturas frías y constantes durante todo el año y dos estaciones, una húmeda y otra seca. Temperatura media anual de 13.7 C. [6]

Guayaquil

Guayaquil la capital económica del país, se encuentra ubicada al sur oeste, en la costa del Océano Pacífico. Además es cabecera cantonal y la segunda ciudad más poblada del Ecuador [49]. El puerto de la ciudad es uno de los más importantes de la costa del océano Pacifico Oriental [8], localizado a 10 kilómetros del mar, en la orilla oeste del Río Guayas. Su posición cerca del ecuador significa que el clima de la ciudad es tropical cálido y húmedo, con temperaturas constantes durante todo el año y con una estación lluviosa y otra seca. Temperatura media anual 25.6 ℃ [9].

1.1. Enfermedades Crónicas no Transmisibles (ECNT)

Acorde al Anexo 3 (Propuesta de Investigación Convocatoria para Proyectos de Investigación PUCE 2017), las enfermedades crónicas no transmisibles (ECNT), causan anualmente 40 millones de muertes a nivel mundial, y aumentan cada año. El 70% de las muertes que se producen en el mundo, son a causa de la ECNT. La población entre 30 y 69 años son las más afectadas, aproximadamente 15 millones de personas en este grupo etario

las padecen. Adicionalmente más del 80% de estas muertes ocurren en países de ingresos económicos bajos y medianos. (OMS, 2017).

Las enfermedades cardiovasculares, el cáncer, las enfermedades respiratorias y la diabetes, ocasionan un número elevado de muertes cada año. El consumo de tabaco, la inactividad física, el uso nocivo del alcohol y las dietas malsanas, aumentan el riesgo de morir a causa de alguna de las ECNT. (OMS, 2017)

En Ecuador, según el Ministerio de Salud Pública (MSP) en el 2016, se registraron como causas de mortalidad: en primer lugar, al infarto de miocardio con 6106 defunciones, en tercer lugar, a la diabetes mellitus con 2360 defunciones, la octava causa fue la enfermedad cardiaca hipertensiva con 1485 defunciones y en décimo primer lugar la hipertensión arterial esencial con 1033 defunciones, de un total de 67506 notificadas en este año.

En el sistema de vigilancia de enfermedades crónicas del Ministerio de Salud Pública del Ecuador, se evidenció para el 2016 alrededor de 106.008 personas con diabetes, y 220.638 personas con hipertensión arterial. Además, entre las 20 causas de mortalidad y morbilidad se encontraron enfermedades como consecuencia de la hipertensión arterial y la diabetes mellitus, consideradas como factores de riesgo y enfermedades crónicas no transmisibles. (Ministerio de Salud de Publica de Ecuador, 2017)

1.2. Investigaciones Relacionadas a Nivel Nacional e Internacional

Durante los próximos diez años, gracias a los adelantos científicos, se pretende alcanzar una nueva meta global para la prevención de enfermedades crónicas, esta es la reducción del 2% anual en las tasas de mortalidad por enfermedades crónicas; sin embargo, algunos países deben lidiar con sus recursos económicos limitados, así como la doble carga de la enfermedad, es decir la presencia de enfermedades infecciosas y crónicas. [10]

La estrategia Global de la Organización Mundial de la Salud (OMS) para la dieta, la actividad física, y la salud, dentro del marco de control del uso del tabaco, describe las acciones necesarias para la reducción de las malas prácticas, así como la adopción de prácticas saludables, para mejorar la calidad de vida de las poblaciones. (OMS, 2019)

Es necesario realizar acciones urgentes y efectivas para manejar el progresivo aumento de las enfermedades crónicas, lo que demanda analizar las evidencias necesarias para la identificación y el manejo adecuado de los recursos en base al contexto encontrado, unas de las alternativas son los sistemas de vigilancia, planteados localmente y a nivel mundial como es el caso del instrumento STEPS de la OMS (herramienta utilizada para recopilar datos y medir los factores de riesgo de las enfermedades crónicas. Este Instrumento comprende tres niveles diferentes o *Steps* de evaluación de los factores de riesgo: Step 1, Step 2 y Step 3). (OMS, 2019)

En los países de mediano y bajo ingreso, cuatro de cada cinco muertes son debido a las ECNT, los factores de riesgo asociados a las mismas se han extendido, estas cifras no solo se evidencian en las áreas urbanas, las áreas rurales están siendo cada día más afectadas, debido al progreso, y a los procesos de urbanización mal planificados; la globalización ha llegado al área rural y esto ha cambiado el pensamiento de que había una mejor calidad de vida en el área rural. Esta afirmación se la ha podido realizar gracias a las evidencias obtenidas de la utilización del STEPS. (OMS, 2019) [11] [12]. La generación de un sistema funcional y permanente bajo la vigilancia comunitaria de los factores de riesgo asociados a las enfermedades crónicas no transmisibles puede ser parte de la solución en el proceso de identificar las otras áreas relacionadas estas enfermedades. [10]

En el 2008, en el Ecuador las enfermedades crónicas no transmisibles (diabetes mellitus, enfermedades cerebrovasculares, enfermedades hipertensivas, cardiopatía isquémica, insuficiencia cardiaca y cirrosis), junto con los accidentes de transporte terrestre y las agresiones, fueron las principales causas de muerte en la población general [13]. La prevalencia de hipertensión arterial es de 9.3% en la población de 18 a 59 años de edad; la pre hipertensión (120-139 / 80-89) es de 37.2%; la prevalencia de diabetes en la población de 30 a 59 años es del 17.6%, siendo el grupo etario de 50 a 59 años, el más afectado con una prevalencia del 10.3%, de acuerdo a ENSANUT (Encuesta Nacional de Salud y Nutrición). [5]

1.3. Diabetes

La OMS describe a la diabetes de la siguiente manera: [15]

> "La diabetes es una enfermedad crónica que aparece cuando el páncreas no produce insulina suficiente o cuando el organismo no utiliza eficazmente la insulina que produce. La insulina es una hormona que regula el azúcar en la sangre.
> El efecto de la diabetes no controlada es la hiperglucemia (aumento del azúcar en la sangre), que con el tiempo daña gravemente muchos órganos y sistemas, especialmente los nervios y los vasos sanguíneos."

Diabetes de tipo 1

La diabetes de tipo 1 (también llamada insulinodependiente, juvenil o de inicio en la infancia) se caracteriza por una producción deficiente de insulina y requiere la administración diaria de esta hormona. Se desconoce aún la causa de la diabetes de tipo 1 y no se puede prevenir con el conocimiento actual. Sus síntomas consisten, entre otros, en excreción excesiva de orina (poliuria), sed (polidipsia), hambre constante (polifagia), pérdida de peso, trastornos visuales y cansancio. Estos síntomas pueden aparecer de forma súbita.

Diabetes de tipo 2

La diabetes de tipo 2 (también llamada no insulinodependiente o de inicio en la edad adulta) se debe a una utilización ineficaz de la insulina. Este tipo representa el 90% de los casos mundiales. Y se debe en gran medida a un peso corporal excesivo y a la inactividad

física. Los síntomas pueden ser similares a los de la diabetes de tipo 1, pero a menudo menos intensos. En consecuencia, la enfermedad puede diagnosticarse solo cuando ya tiene varios años de evolución y han aparecido complicaciones. Hasta hace poco, este tipo de diabetes solo se observaba en adultos, pero en la actualidad también se está manifestando en niños.

Consecuencias frecuentes de la diabetes.

- Con el tiempo, la diabetes puede dañar el corazón, los vasos sanguíneos, ojos, riñones y nervios.
- La diabetes aumenta el riesgo de cardiopatía y accidente vascular cerebral (AVC). Según un estudio realizado en varios países, un 50% de los pacientes diabéticos muere de enfermedad cardiovascular (principalmente cardiopatía y AVC).
- La neuropatía de los pies combinada con la reducción del flujo sanguíneo incrementa el riesgo de úlceras de los pies, infección y, en última instancia, amputación.

1.4. Enfermedades Cardiovasculares

De acuerdo a la OMS [16], las enfermedades cardiovasculares (ECV) son un grupo de desórdenes del corazón y de los vasos sanguíneos, entre los que se incluyen:

- La cardiopatía coronaria: enfermedad de los vasos sanguíneos que irrigan el músculo cardiaco.
- Las enfermedades cerebrovasculares: enfermedades de los vasos sanguíneos que irrigan el cerebro. Las arteriopatías periféricas: enfermedades de los vasos sanguíneos que irrigan los miembros superiores e inferiores.
- La cardiopatía reumática: lesiones del músculo cardiaco y de las válvulas cardíacas debidas a la fiebre reumática, una enfermedad causada por bacterias denominadas estreptococos, entre otras.

Los ataques al corazón y los accidentes vasculares cerebrales (AVC) suelen ser fenómenos agudos que se deben sobre todo a obstrucciones que impiden que la sangre fluya hacia el corazón o el cerebro. La causa más frecuente es la formación de depósitos de grasa en las paredes de los vasos sanguíneos que irrigan el corazón o el cerebro. Los AVC también pueden deberse a hemorragias de los vasos cerebrales o coágulos de sangre. Los ataques cardíacos y accidentes cerebrovasculares (ACV) suelen tener su causa en la presencia de una combinación de factores de riesgo, tales como el tabaquismo, las dietas malsanas y la obesidad, la inactividad física, el consumo nocivo de alcohol, la hipertensión arterial, la diabetes y la hiperlipidemia.

Principales factores de riesgo

Las causas más importantes de cardiopatía y AVC son una dieta malsana, la inactividad física, el consumo de tabaco y el consumo nocivo de alcohol. Los efectos de los factores de

riesgo comportamentales pueden manifestarse en las personas en forma de hipertensión arterial, hiperglucemia, hiperlipidemia y sobrepeso u obesidad. Estos "factores de riesgo intermediarios", que pueden medirse en los centros de atención primaria, son indicativos de un aumento del riesgo de sufrir ataques cardíacos, accidentes cerebrovasculares, insuficiencia cardíaca y otras complicaciones.

Está demostrado que el cese del consumo de tabaco, la reducción de la sal de la dieta, el consumo de frutas y hortalizas, la actividad física regular y la evitación del consumo nocivo de alcohol reducen el riesgo de ECV. Por otro lado, puede ser necesario prescribir un tratamiento farmacológico para la diabetes, la hipertensión o la hiperlipidemia, con el fin de reducir el riesgo cardiovascular y prevenir ataques cardíacos y accidentes cerebrovasculares. Las políticas sanitarias que crean entornos propicios para asegurar la asequibilidad y disponibilidad de opciones saludables son esenciales para motivar a las personas para que adopten y mantengan comportamientos sanos.

También hay una serie de determinantes subyacentes de las enfermedades crónicas, es decir, "las causas de las causas", que son un reflejo de las principales fuerzas que rigen los cambios sociales, económicos y culturales: la globalización, la urbanización y el envejecimiento de la población. Otros determinantes de las ECV son la pobreza, el estrés y los factores hereditarios.

Síntomas comunes de las enfermedades cardiovasculares

Síntomas de cardiopatía y AVC

La enfermedad subyacente de los vasos sanguíneos a menudo no suele presentar síntomas, y su primera manifestación puede ser un ataque al corazón o un AVC. Los síntomas del ataque al corazón consisten en dolor o molestias en el pecho, dolor o molestias en los brazos, hombro izquierdo, mandíbula o espalda.

Además puede haber dificultad para respirar, náuseas o vómitos, mareos o desmayos, sudores fríos y palidez. La dificultad para respirar, las náuseas y vómitos y el dolor en la mandíbula o la espalda son más frecuentes en las mujeres.

El síntoma más común del AVC es la pérdida súbita, generalmente unilateral, de fuerza muscular en los brazos, piernas o cara. Otros síntomas consisten en la aparición súbita, generalmente unilateral, de entumecimiento en la cara, piernas o brazos; confusión, dificultad para hablar o comprender lo que se dice; problemas visuales en uno o ambos ojos; dificultad para caminar, mareos, pérdida de equilibrio o coordinación; dolor de cabeza intenso de causa desconocida; y debilidad o pérdida de conciencia. Quienes sufran estos síntomas deben acudir inmediatamente al médico.

Reducción de la Carga de Enfermedades Cardiovasculares.

La OMS ha identificado una serie "inversiones óptimas" o intervenciones muy costoeficaces para prevenir y controlar las ECV, cuya aplicación es viable incluso en entornos con escasos recursos. Existen dos tipos de intervenciones: las poblacionales y las individuales; se recomienda utilizar una combinación de las dos para reducir la mayor parte de la carga de ECV.

A nivel individual, las intervenciones sanitarias de prevención de los primeros ataques cardíacos y accidentes cerebrovasculares, deben centrarse primordialmente en las personas que, si se tienen en cuenta todos los factores, presentan un riesgo cardiovascular medio a alto o en los individuos que presentan un solo factor de riesgo —por ejemplo, diabetes, hipertensión o hipercolesterolemia— con niveles superiores a los umbrales de tratamiento recomendados. La primera intervención (basada en un enfoque integral que tiene en cuenta todos los riesgos) es más rentable que la segunda y tiene el potencial de reducir sustancialmente los episodios cardiovasculares. Se trata de un enfoque viable dentro de los servicios de atención primaria en entornos de escasos recursos, que puede ser puesto en práctica incluso por trabajadores sanitarios que no son médicos.

Para la prevención secundaria de enfermedades cardiovasculares en pacientes con diagnóstico definitivo, por ejemplo, de diabetes, es necesario administrar tratamientos con los siguientes fármacos: ácido acetilsalicílico; betabloqueantes; inhibidores de la enzima convertidora de la angiotensina; estatinas.

Los efectos de estas intervenciones son en buena parte independientes, aunque si se combinan con el cese del consumo de tabaco, se puede prevenir cerca del 75% de los episodios cardiovasculares recurrentes. Hoy por hoy, la aplicación de estas intervenciones presenta grandes deficiencias, sobre todo en el nivel de la atención primaria.

Por otro lado, se requieren a veces costosas operaciones quirúrgicas para tratar las ECV, tales como: derivaciones coronarias; angioplastia con globo (introducción de un pequeño globo en una arteria obstruida para reabrirla); reparaciones y sustituciones valvulares; trasplante cardíaco; implantación de corazones artificiales.

También se necesitan dispositivos médicos para tratar algunas ECV, por ejemplo: marcapasos, válvulas protésicas y parches para cerrar comunicaciones entre las cavidades del corazón.

1.5. Encuesta Nacional de Salud y Nutrición 2012. ENSANUT

La Encuesta Nacional de Salud y Nutrición explica lo siguiente:

> "El Ministerio de Salud, en su compromiso de proteger la salud de la población del país, debe tener como respaldo la evidencia científica, y la información más confiable y actualizada sobre la situación de salud de los ecuatorianos. Con este propósito, el Ministerio de Salud, conjuntamente con el Instituto Nacional de Estadística y Censos, se comprometió llevar a efecto una encuesta nacional que actualizará los datos de la situación nutricional de

> la población, que recogiera datos sobre la salud sexual y reproductiva, y que investigará los factores de riesgo de las enfermedades crónicas más prevalentes. Esta información servirá de base para la adopción de estrategias, diseño de políticas públicas y programas que protejan a toda la población." (MSP ENSANUT ECU, 2014)

En base a este objetivo, el ministerio de salud realizó esta encuesta desde el 2011 hasta el 2013. Su diseño muestral permite extrapolar los datos a nivel nacional, subregional, por zonas de planificación por condición social, por rango de edad, por etnia y por sexo y ofrece un panorama de la dimensión de los problemas de salud y nutrición. El mismo que constituye el insumo base para el reforzamiento de políticas de salud.

Los resultados de esta encuesta además de permitir profundizar en el análisis de problemas de salud y nutrición conforman la base fundamental para generar otras investigaciones que respondan al constante cambio epidemiológico y nutricional de la población.

1.6. Justificación

La identificación de la prevalencia de los factores de riesgo asociados a las ECNT en una comunidad rural, ha sido poco estudiada a nivel de Latinoamérica. Existen estudios en el continente Africano, en la India y en Asia, en los cuales utilizaron la metodología STEPS de la OMS a nivel de comunidades rurales y encontraron una prevalencia comparable con las áreas urbana; de ahí que nace el interés de trabajar en las áreas rurales, donde se piensa que por ser agricultores, por tener una alimentación diferente a la de la ciudad, no existen o hay en menor frecuencia, factores de riesgo para desarrollar ECNT.[17] [18] (OMS, 2019)

La evidencia en otros contextos en los cuales la urbanización, el cambio el estilo de vida, el aumento de la expectativa de vida y el fenómeno de la globalización, han incrementado en la última década, en las áreas rurales los niveles de prevalencia de los factores de riesgo para ECNT. [12][19][20]

El desarrollo de un proceso de vigilancia comunitaria, en el cual la comunidad cuida de la comunidad utilizando herramientas de la epidemiologia comunitaria es muy importante y necesario ya que permitirá identificar los factores de riesgo para diabetes e hipertensión, y establecer una comunicación eficaz y efectiva entre la comunidad y los actores y responsables de la salud local y distrital; lo que ayudará a mejorar la calidad de vida de las comunidades rurales. [21] [22] [23]

1.7. Objetivos de esta tesis

Objetivo General

Desarrollar un aplicativo de visualización de datos y analíticos visuales que brinde funcionalidades de análisis exploratorio y analíticos visuales al análisis de la encuesta nacional de salud y nutrición *ENSANUT* desarrollada en el 2012 en Ecuador, permitiendo entre otras cosas detectar nuevos patrones en la vigilancia del estado de salud y nutrición en

Ecuador, con especial énfasis en las enfermedades crónicas no transmisibles (diabetes e hipertensión).

Objetivos Específicos

- Utilizar el proceso de descubrimiento del conocimiento *KDD* y técnicas de visualización para resolver el problema planteado en el objetivo principal
- Facilitar con estas técnicas las tareas de análisis para identificar las variables que más influyen en las enfermedades crónicas no transmisibles, tales como diabetes e hipertensión.
- Realizar un análisis a nivel poblacional para identificar patrones y características más relevantes de las diferentes regiones, subregiones y provincias del país, basado en el consumo de alimentos, aspectos antropométricos, bioquímicos, demográficos y geográficos.
- Aportar al reforzamiento de políticas públicas, servicios de salud y educación.

2. Estado del Arte

2.1. Visualización de la información

La Visualización de la Información (VI) puede pensarse como la articulación de visualización (definida en este contexto como *formar en la mente una imagen visual de un concepto abstracto*), e información (entendida *conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada*). De ese modo una posible definición de VI podría ser *la comunicación de conocimiento a través de representaciones visuales.* Otros autores (Card, Mackinlay, & Shneiderman, 1999) la definen como *"el uso de representaciones visuales e interactivas de datos abstractos soportadas por el uso de computadoras, para amplificar el conocimiento"* [24]. Si consideramos la primera definición de Visualización de la Información que propusimos en el apartado anterior, podemos encontrar los orígenes de la misma en épocas muy remotas de la historia de la humanidad. El hombre siempre ha buscado formas de comunicar su conocimiento en forma visual, y si bien hay casos notables más antiguos, podría decirse que el siglo XVIII es el que la mayoría de los autores nombran al referirse a los comienzos de la Visualización de la Información. El trabajo más comentado en la bibliografía de Visualización de la Información como origen de esta disciplina, es el de William Playfair (1786) quien parece haber sido el primero en utilizar la línea y el área de un gráfico para representar datos visualmente (**Error! Reference source not found.**~~Figura~~ 1). Muchos de los reportes de negocios que vemos en la actualidad son variantes de los gráficos desarrollados por Playfair en aquel entonces.

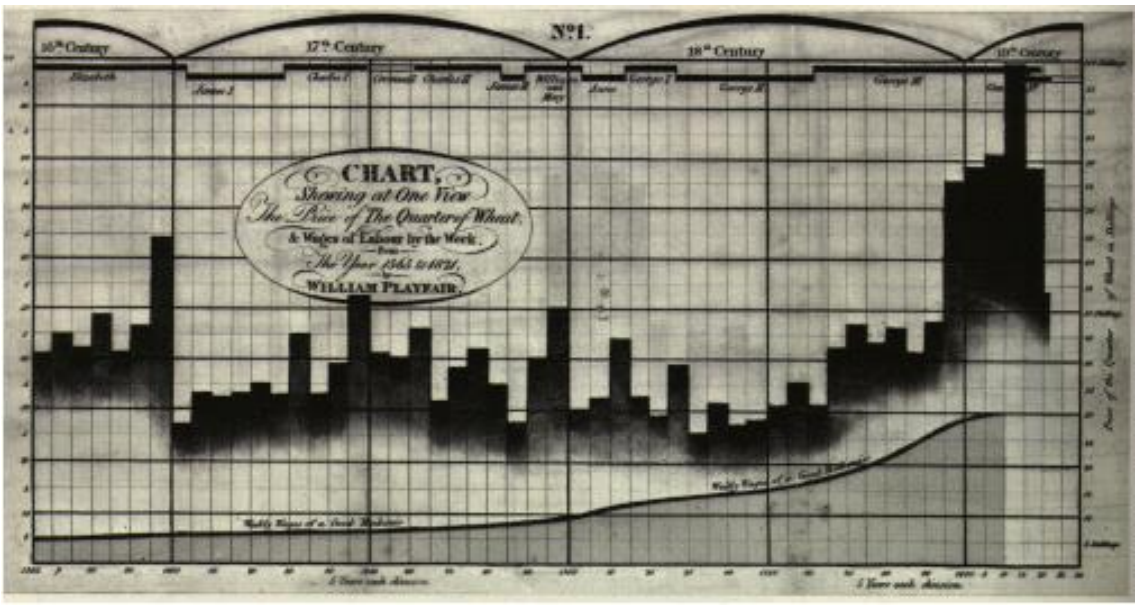


Figura 1. William Playfair – Precio del Trigo y los Salarios (Playfair, 1786)

La segunda mitad del 1800 es conocida como la Edad de Oro en la Visualización de la Información, por la cantidad de innovaciones y mejoras introducidas en la cartografía y gráficos estadísticos. Dentro de esta era encontramos el famoso gráfico de John Snow (1855), representando la epidemia de cólera que sufrió la ciudad de Londres en el año 1854. Snow combinó el mapa de la ciudad de Londres junto con gráficos de barras para identificar visualmente la causa de la epidemia. Así descubrió que la mayoría de los muertos por esta enfermedad se encontraban en las inmediaciones de la bomba de agua de la calle Broad, refutando la teoría de que el cólera se trasmitía por aire.

Florence Nightingale es otra de las grandes aclamadas de la Era de Oro de la Visualización. Se la conoce como la autora de los diagramas de área polar o *coxcombs* (Nightingale, 1857) Estos gráficos fueron utilizados para demostrar la diferencia entre las causas de mortalidad en hospitales británicos durante la guerra de Crimea (**Error! Reference source not found.**~~Figura~~ 2). En esta misma época encontramos el mapa que representa la Campaña del Ejército de Napoléon en Rusia de 1812-1813, de Charles J. Minard, muchas veces citado como el mejor gráfico estadístico que jamás se haya logrado (Figura 2) [25].

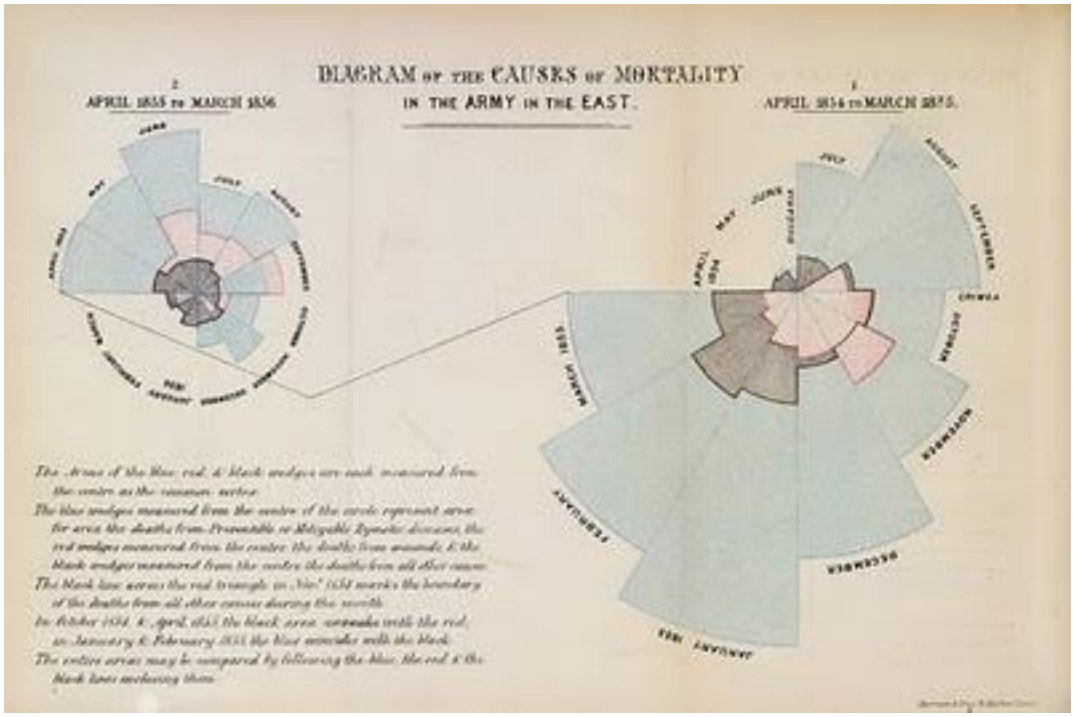


Figura 2. Diagrama de Causas de Mortalidad en el Ejército en el Este. Florence Nightingale (1857)

En la segunda mitad del siglo XX comienza otra era de creatividad en el área de visualización, especialmente de la mano de dos autores: Jacques Bertin y John Tukey. Jacques Bertin, cartógrafo francés, publica en 1967 "*Semiologie Graphique"* (Bertin, 1967) un libro cuyo contenido sigue vigente hasta el día de hoy [26]. En su trabajo determina que toda visualización está formada por una serie de componentes que tienen distinto poder expresivo y que cada uno de ellos funciona mejor dadas ciertas condiciones. Bertin sugiere seis variables básicas para la creación de gráficos: tamaño, valor, textura, color, orientación y forma. Para cada una señala en qué casos funcionan mejor y cómo utilizarlas.

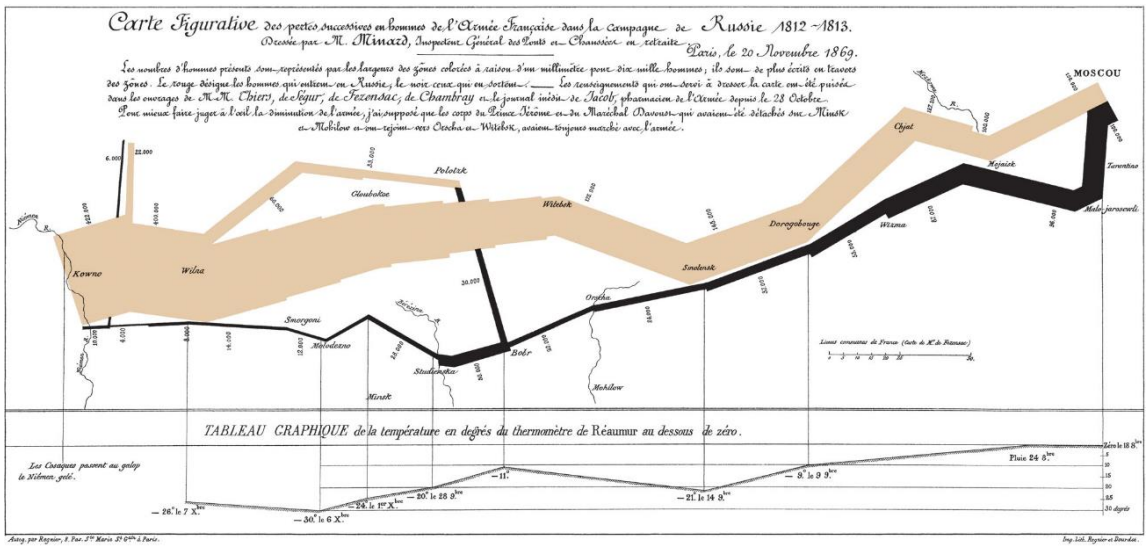


Figura 3. Gráfico que ilustra las pérdidas sucesivas de soldados franceses en la campaña a Rusia 1812-13 (Minard, 1861)

Si bien el trabajo de Tukey *Exploratory Data Analysis* (EDA) se publicó recién en 1977 (Tukey, Exploratory Data Analysis, 1977), ya en 1962 el autor proponía en su trabajo *The Future of Data Analysis* (Tukey, 1962), la legitimación del análisis de información estadística como una rama distinta a la estadística matemática. En EDA el énfasis está puesto en lograr rápidas compresiones de información estadística a través del uso de imágenes. En este trabajo introduce el ya famoso gráfico de *"box and whisker"* que le permite al analista comprender rápidamente los cuatro indicadores principales de la distribución de los datos (cuartiles, mediana, mínimo y máximo).

Con la aparición de las computadoras, se da otra era de explosión creativa en la disciplina y los desarrollos desde entonces son innumerables. Con el crecimiento de la informática, es cada vez más fácil generar y almacenar grandes cantidades de información, por lo que comienzan a ser necesarias nuevas técnicas de visualización específicamente diseñadas para grandes volúmenes de datos. A partir del año 2000, las técnicas de Visualización de la Información se expanden a la mayoría de las áreas del saber. Con la globalización del conocimiento a través del uso de internet, los trabajos en el área crecen a un ritmo exponencial. Los eventos y conferencias sobre Visualización también se multiplican y se subdividen cada vez más según las distintas ramas de aplicación. Tal es así que la conferencia de la IEEE, que comenzó en 1990, tiene hoy tres subsecciones: VAST, dedicada exclusivamente a los trabajos en el área de Visual Analytics (concepto que desarrollaremos más adelante y principal foco de este trabajo); InfoVis, para trabajos más genéricos de Visualización de la Información; y SciVis, la sección exclusiva para los trabajos de Visualización Científica.

Una de las áreas de conocimiento donde se introdujo la ciencia de la Visualización, es la Minería de Datos. Si bien ambas disciplinas han evolucionado en forma separada,

comparten en la actualidad varios puntos en común. Muchos investigadores han trabajado en la integración de ambas y se encuentran aun haciéndolo. Dado que el foco de este trabajo es precisamente tal combinación, describiremos la evolución de la conjunción de ambas disciplinas en un apartado posterior.

Con la evolución de la disciplina, surgieron también varios enfoques sobre el proceso de construcción de visualizaciones. Diferentes referentes de Visualización de la Información enumeran diversas etapas e interacciones en el proceso de visualización de la información. Uno de los autores más citados es (Ware, Information Visualization. Perception for Design, 2004) quien enumera las siguientes etapas en el proceso de Visualización (**Error! Reference source not found.**~~Figura~~ 4) [26]:

- Recolección y almacenamiento de la información
- Preprocesamiento prediseñado para transformar los datos en algo comprensible por la mente humana
- El hardware y algoritmos gráficos necesarios para producir imágenes en una pantalla
- El sistema de percepción y cognitivo del humano

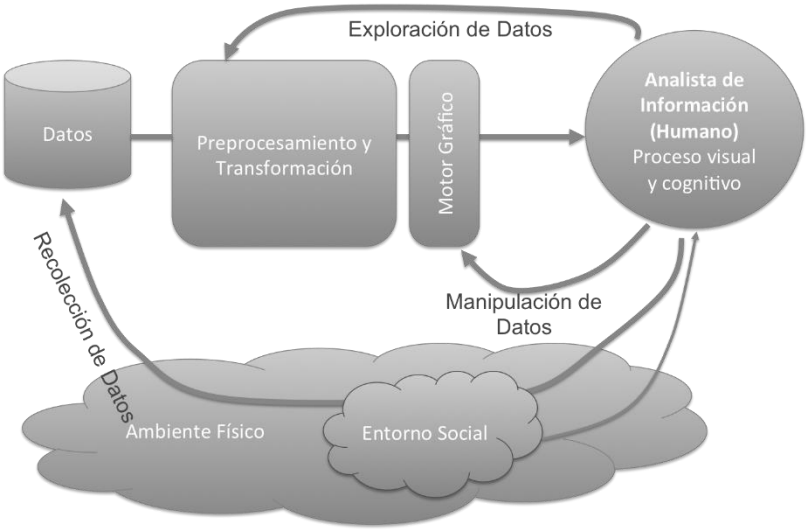


Figura 4. Diagrama esquemático del Proceso de Visualización (adaptado de Ware, 2004)

En el esquema propuesto por Ware, la mayor parte del proceso está relacionada con la recolección de datos. Esta recolección está influenciada por un ambiente físico (fuentes de datos disponibles) y un entorno social que determina qué datos buscar en función de la percepción del analista. El proceso es altamente interactivo ya que es el humano quien determina cuándo se da fin al ciclo de visualización. En cambio, (Fry, Processing: A

Programming Handbook for Visual Designers and Artists, 2007) considera que el proceso de visualización de datos consiste en una serie de pasos para poder contestar una pregunta que da origen a la necesidad de visualización; [27]

- **Adquirir**. Obtener los datos ya sea de un archivo, un disco o un sitio en la red.
- **Analizar**. Proveer cierta estructura a los datos y ordenarlos en categorías.
- **Filtrar**. Remover los datos que no sean de interés.
- **Extraer**. Aplicar métodos de estadística o minería de datos para encontrar patrones o ubicar la información en un contexto matemático.
- **Representar**. Elegir un modelo visual básico.
- **Refinar**. Mejorar el modelo básico para hacerlo más claro o más visualmente atractivo.
- **Interactuar**. Agregar métodos para manipular la información o determinar qué debe verse.

Otros autores que presentaron conceptualizaciones similares fueron Haber y McNabb, Daniel Keim, Tamara Munzner, por citar los más recientes.

La relación entre VI y minería de datos (MD) es muy importante, y uno de los aspectos de su convergencia es lo que se conoce actualmente como analíticos visuales (visual analytics). La MD puede pensarse como un paso dentro de un proceso más amplio de descubrimiento del conocimiento. Existen mayormente dos enfoques para estudiar la evolución de la MD. El primero parte de la evolución del concepto como resultado del desarrollo de los sistemas de bases de datos. Según un segundo enfoque, la MD es el resultado de la evolución de las tres disciplinas principales que la componen: la estadística, el aprendizaje automático y la inteligencia artificial.

Tradicionalmente las técnicas de MD pueden dividirse según el fin que tienen: extracción o identificación de patrones, agrupamiento de datos (clustering), clasificación o categorización predictiva. Cada uno de estos grupos ha experimentado una evolución diferente. La técnica más conocida de extracción de patrones es la de Reglas de Asociación propuesta por Agrawal (Agrawal, Imielinski, & Swami, 1993) para la detección de patrones de compra en tickets de supermercados [28]. Una regla de asociación es una regla probabilística que determina que si ciertos conjuntos de datos ocurren en forma simultánea, entonces otros conjuntos de atributos también tienen posibilidades de ocurrir. Esta técnica ha sido muy utilizada en la comunidad de MD y a la vez referenciada en la mayor parte de los trabajos de la disciplina. Luego de este primer trabajo, el algoritmo se extendió a otras aplicaciones. Hoy el mismo es utilizado por muchos sistemas de recomendación para sugerir ítems de interés a las personas de acuerdo a su comportamiento de compra.

Otra de las grandes categorías de algoritmos utilizados en MD es la de agrupamiento o clustering. Esta técnica consiste en agrupar los datos en grupos pero sin

conocimiento previo de los mismos. A diferencia de la mayoría de los algoritmos de MD, se trata de una técnica exploratoria. La división más común en clustering es separar los algoritmos entre métodos jerárquicos y métodos de particionamiento. Los primeros construyen los grupos gradualmente, mientras que los segundos aprenden los grupos directamente. Las técnicas de agrupamiento han sido ampliamente utilizadas en aplicaciones de MD y continúan desarrollándose para poder resolver las principales limitaciones de estos algoritmos. Entre ellas podemos mencionar el tipo de atributos que pueden manejar, la escalabilidad a grandes bases de datos, la habilidad para trabajar con datos con muchas dimensiones, la capacidad para encontrar grupos en formas irregulares (no basadas en elipses), el manejo de casos extremos (outliers), la complejidad del tiempo, la dependencia de orden de datos, el apoyo en el conocimiento a priori de los usuarios y la interpretabilidad de los resultados.

El último grupo de algoritmos de MD es el de clasificación o categorización. Estas técnicas tienen como objetivo construir clasificadores que pueden ser aplicados a nuevos datos para categorizar éstos en grupos. La diferencia con las técnicas de clustering, es que para clasificar necesitamos datos pre-categorizados que servirán para entrenar el algoritmo. Debido a esto, las técnicas de clasificación suelen ser llamadas algoritmos de aprendizaje supervisado mientras que las de agrupamiento son referenciadas como algoritmos de aprendizaje no supervisado. Dentro de este grupo se encuentran los árboles de decisión, técnicas de regresión y redes neuronales.

Las técnicas de regresión también han sido utilizadas, y continúan siéndolo, para resolver problemas de clasificación en MD. En casos simples, se puede utilizar la regresión lineal para predecir valores futuros de datos. Sin embargo, los problemas del mundo real suelen no ser proyecciones lineales de datos pasados, por lo que es necesario aplicar técnicas de regresión más avanzadas, como por ejemplo regresión logística. La regresión logística es una generalización de la regresión lineal. Es utilizada principalmente para predecir variables binarias y ocasionalmente variables multi-clase. Esta metodología se enmarca dentro del conjunto de Modelos Lineales Generalizados (GLM). A diferencia de los árboles de decisión, la regresión logística devuelve la probabilidad de que un valor pertenezca a determinada clase. Dado que tiene sus orígenes en la rama de estadísticos de la MD, muchas veces es preferida como técnica de clasificación por sobre los árboles de decisión o redes neuronales.

La complejidad creciente en varios de los campos de aplicación y los avances en la tecnología ha impuesto nuevos desafíos para la profesión. Entre ellos podemos mencionar diferentes formatos de datos, ubicaciones dispares, avances en redes, nuevos contextos de negocio, por mencionar sólo algunos. La integración de MD con técnicas de VI también es un gran reto y es lo que pretende resolver la investigación en Analíticos Visuales (AV). Según una de las primeras definiciones (Wong, 2004), AV "*es la ciencia de razonamiento analítico facilitado por interfaces humano-máquina interactivas*". Sin embargo,

rápidamente la misma fue mejorada, definiéndose hoy AV como (Keim, Kohlhammer, Ellis, & Mansmann, 2010) *"la disciplina que combina técnicas de análisis automático con visualizaciones interactivas para un efectivo entendimiento, razonamiento y toma de decisiones basadas en conjuntos de datos grandes y complejos"*. [29][30]

En la figura 5 puede observarse el modelo propuesto por Keim. El primer paso consiste en la preparación y transformación de los datos para su posterior exploración. Luego de esta tarea, el analista puede elegir entre el análisis visual o el automático. Si se utilizan primero métodos automáticos, se aplican algoritmos de Minería de Datos para el modelado de los mismos. Una vez que se crea el modelo, el analista debe evaluar y refinar el mismo, lo cual puede ser realizado de mejor forma a través de la interacción con los datos. Las visualizaciones permiten a los analistas interactuar con los métodos automáticos al modificar parámetros o seleccionar otros algoritmos. Luego se pueden visualizar los modelos para evaluar los conocimientos generados por los mismos.

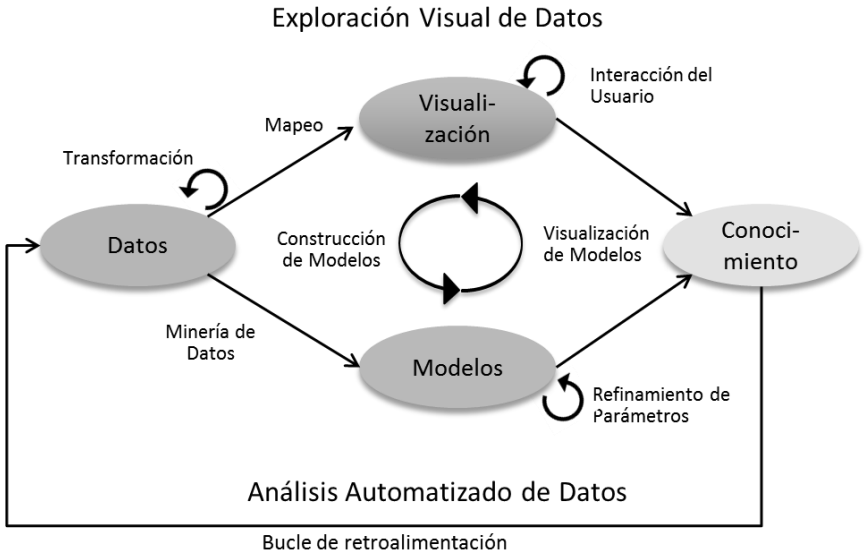


Figura 5. El proceso de AV (traducción de Keim, Kohlhammer, Ellis, & Mansmann, 2010)

En lo que resta de este capítulo presentaremos algunos antecedentes referidos a bases de datos e investigaciones de VI, MD, AV aplicados a nuestro objetivo principal. Acorde al artículo Data-Mining Technologies for Diabetes: A Systematic Review [31] de la base de datos de MEDLINE (una de las más completas en investigación médica) se destacan 17 artículos que describen los métodos de MD usados en investigaciones relacionadas con diabetes, en los cuales se enfatizan identificar los tipos de diabetes, métodos y software de MD y cuyos resultados están enfocados en el desarrollo científico y a su vez, mejorar la salud de los pacientes.

2.2. Búsqueda del Mejor Predictor

Huang y colaboradores, aplicando Naive Baiyes, clasificador IB1, y árbol de decisión c4.5 sobre una base de datos de 2064 registros, donde 1148 son de género masculino encontraron que los principales factores que incrementan los niveles de glucosa en la sangre son, la edad, la duración de diagnóstico, si necesita o no insulina y la dieta del paciente. Myiaki y colaboradores, se enfocaron en los predictores de complicaciones cardiovasculares y usando árboles de decisión sobre una base de datos de 165 pacientes encontraron que las variables más influyentes eran el peso y la edad.

Por otro lado Sigurdardottir y colaboradores, aplicando árbol de decisión c4.5 (WEKA) encontraron que la variable "Intervención en la educación diabética" es altamente significativa en cuanto a los niveles de glucosa en la sangre se refiere. Este estudio sugiere adicionalmente que los factores tales como duración, contenido de la educación e intensidad de la educación no tienen impacto en los cambio de la glucosa en la sangre.

2.3. Análisis de datos Genómicos asociados a la Diabetes

Covani y compañeros estudiaron la asociación entre la periodontitis, acorde a Mediplus "es una inflamación e infección de los ligamentos y huesos que sirven de soporte a los dientes", y la diabetes de tipo 2 ambas comparten 4 genes líderes, el método del gen líder agrupa genes de una lista de acuerdo a su peso y cantidad de relaciones usando clustering jerárquico o k medias. Finalmente los genes pertenecientes a los clúster mejor ranqueados son considerados como genes líderes. La sinusitis "inflamación de los senos paranasales" y la periodontitis no comparten genes líderes. Este experimento carece de validez científica ya que es puramente teórico, sin embargo permite generar nuevas hipótesis que servirán en investigaciones posteriores.

2.4. Otras Investigaciones

Las siguientes investigaciones están relacionadas con el seguimiento del cuidado de la salud, limpieza de datos, análisis de efectos secundarios en ciertas drogas, detección de fraude en el seguro de salud y predicción de mortalidad temprana.

Concarol y coautores, utilizando reglas de asociación basadas en secuencias de eventos sobre una base de datos de 101339 registros llegaron a la conclusión de que el 56% en el incremento de la glucemia para una persona con sobrepeso que se encuentra bajo una terapia de hipertensión muestra valores normales de glucemia y altos de hemoglobina glicosilada

Según el artículo~~,~~ *Applied visual analytics for exploring the National Health and Nutrition Examination Survey* [40], se desarrolló un sistema de visualizaciones para estudiar el estado de salud y nutrición en los Estados Unidos usando una base de datos nacional de salud y nutrición con sus siglas en inglés (*NHANES*).

La misma que consta de una extensión del dispersograma tradicional ya que en la parte superior de la diagonal muestra una matriz de conglomerados N – dimensional que se ajusta dependiendo de la selección que se haga en filtros tales como grupo etario, sexo, etnia y el número de clusters. En la parte inferior de la diagonal se encuentra la matriz de diagramas de dispersión tradicional.

El objetivo de esta creación es reutilizar el espacio redundante de la parte superior del dispersograma. Estos *clusters* fueron creados utilizando el método k –medias y distancia euclidiana como medida de similaridad. Adicionalmente consta de dos gráficos de barras, el primero con índices de alimentación saludable con sus siglas en inglés (*HEI*) y el segundo en base a una Guía Nutricional para americanos con sus siglas en inglés (*DGA*) para contrastar y encontrar patrones en los conglomerados utilizando los filtros mencionados (Figura 6).

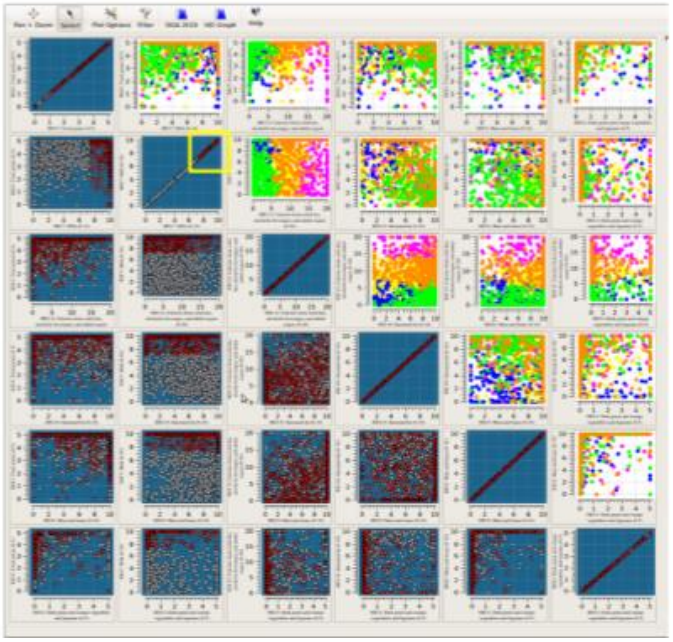


Figura 6. Análisis NHANES

2.5. Grafos y Complicaciones Asociadas a la Diabetes

Acorde a un estudio realizado en México por investigadores la Universidad de Guadalajara y la Universidad de Papaloapan [43], se utilizó la teoría de grafos para estudiar y describir el comportamiento del desarrollo de las úlceras en el pie diabético incorporando información fisiológica de las úlceras. Un problema frecuente en los pacientes diabéticos es el pie diabético, esto debido a las alteraciones vasculares observadas que muestran características especiales.

El pie diabético, se define como infección, ulceración y destrucción de los tejidos profundos, asociadas con anormalidades neurológicas de diversa gravedad en las extremidades inferiores (OMS, 2019). La neuropatía de los pies (lesión microvascular de vasos sanguíneos que irrigan los nervios) combinada con la reducción del flujo sanguíneo incrementa el riesgo de úlceras de los pies, y en última instancia, amputación. La neuropatía diabética se debe a lesión de los nervios a consecuencia de la diabetes, y puede llegar a afectar a un 50% de los pacientes.

De todas las amputaciones relacionadas con diabetes, 70-80% son precedidas por úlceras crónicas y hasta dos tercios experimentarán una segunda amputación (OMS, 2019). Se calcula que el tratamiento y atención básica de la diabetes permitirían prevenir hasta el 80% de las amputaciones de pies diabéticos (OMS, 2019). Esto es, las amputaciones de pie diabético se pueden disminuir mediante la modificación de factores como la neuropatía y el control glucémico, siendo este último el principal factor que causa un desorden en el proceso de angiogénesis para la generación de nuevos vasos sanguíneos.

2.5.1. Representación de las Ulceras en un Grafo

El pie es particularmente vulnerable a daños circulatorios neurológicos, y el menor trauma puede causar úlceras o, incluso, infecciones (Figura 7). La figura 8 indica las principales zonas del pie en las que aparecen con mayor frecuencia las úlceras. Las úlceras localizadas en la zona plantar tienen forma ovalada, son profundas, con bordes callosos y base granulada. Por lo contrario, las que aparecen en los dedos tienen bordes planos e irregulares y forma redondeada y es posible la afectación ósea. Existe un sistema de clasificación desarrollado por Wagner para la estadificación de las úlceras de pie diabético, que ha sido ampliamente aceptado (Figura 9) (Wagner, 1979):

Figura 7. Úlcera neuropática en una posición típica bajo el metatarso y rodeada de callosidad

Figura 8. Puntos más susceptibles a la formación de úlceras en el pie diabético

Para modelar un grafo y poder visualizar la dinámica de las úlceras en el cual se pueda incluir la información fisiológica, definimos un grafo.

"Definición 1. Un grafo G es una terna ordenada (V(G),E(G), ψG) que consiste de un conjunto no vacío V(G) de vértices, de un conjunto E(G) de aristas y de una función de incidencia ψG , que para cada arista se cumple: ψG asocia la arista e, a un par de elementos de V(G), ψG(e) = { u,v }".[19]

El grafo G tiene vértices etiquetados y todas sus aristas dirigidas, esta dirección representa el cambio de estado de la úlcera. Cada vértice v es etiquetado con la información fisiológica del estado de la úlcera, por la función $V(G) \rightarrow (R+ \cup \{ 0 \}) p$, donde p es el número de parámetros fisiológicos considerados para el desarrollo de úlceras. Representación de las úlceras en un grafo (Figura 10).

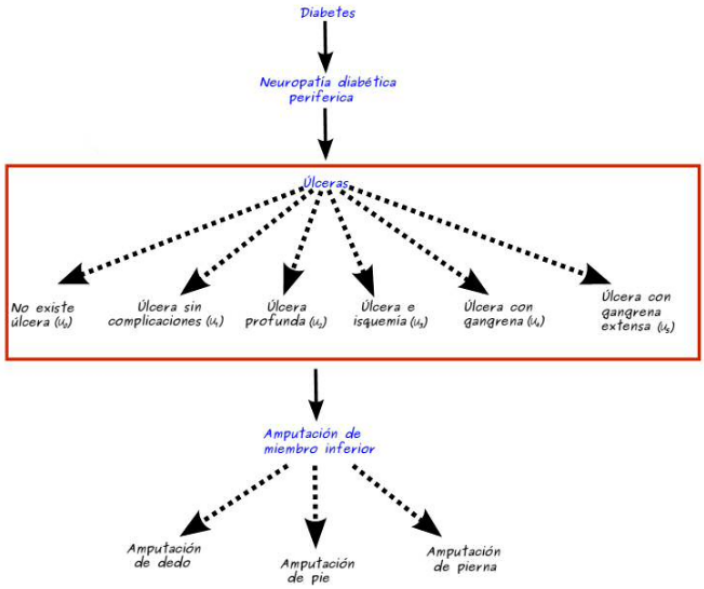


Figura 9. Clasificación de úlceras

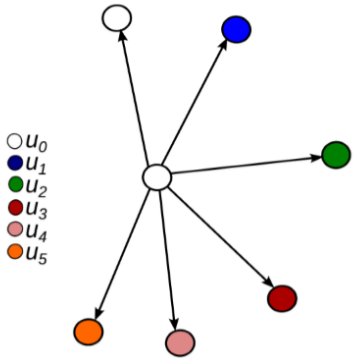


Figura 10. Grafo de esquema de úlceras

2.5.2. Resultados

"El cambio de estado de la úlcera depende de la concentración de glucosa C_{gluc} y el consumo de oxígeno R_{oxy} en sangre (Buchwald, 2011). Esto es, la concentración de oxígeno determina el estado de la úlcera (Lepantalo & et al., 2011). Por otra parte, el tiempo en el que se puede dar el cambio de estado depende de varios factores que se tienen que considerar, pero no se cuenta con la suficiente información experimental al respecto para considerar el factor tiempo". [31]

Con base en datos experimentales de monitoreo durante un año en pacientes con úlceras, se determinaron los intervalos de concentración de oxígeno (Lepantalo & et al., 2011):

"u0 ∈ (0.114, 0.143) M m-3, u1 ∈ (0.084, 0.114) M m-3, u2 ∈ (0.057, 0.084) M m-3, u3 ∈ (0.043, 0.057) M m-3, u4 ∈ (0.03, 0.043) M m-3 y u5 ∈ (0.026, 0.03) M m-3" [31].

En la figura 6 se describe la dinámica de la evolución de las estados ui (i = 0,1,2,3,4,5) que se definen como sigue: Estado u0, puede cambiar a cualquiera de los 5 estados; Estado u1, puede cambiar a los estados u2, u3, u4, u5 o regresar al estado u0; Estado u2, puede cambiar al estado u3, u4, u5 o regresar al estado u0 o u1; Estado u3 puede cambiar al estado u4, u5 o bien regresar al estado u1 o u2; Estado u4, se puede cambiar al estado u5 y no se regresa a ningún otro estado.

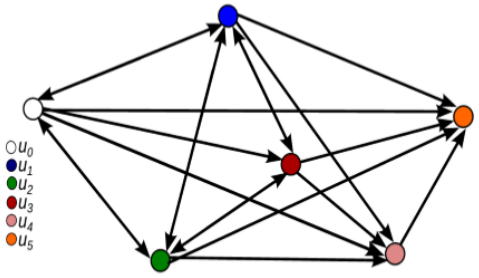


Figura 11. Dinámica de las úlceras

El estado crítico de las úlceras es el u3, ya que además de la úlcera existe isquemia y por lo tanto no se puede regresar a una piel sana, es decir, no es reversible hasta u0, tal como se muestra en la figura anterior. Existe un conjunto de posibles caminos en función de los parámetros fisiológicos, entonces de acuerdo a la dinámica que existe entre los diferentes estados de las úlceras, se determina el siguiente árbol que describe algunas de los posibles escenarios en los que puede estar el paciente con pie diabético (Figura 12).

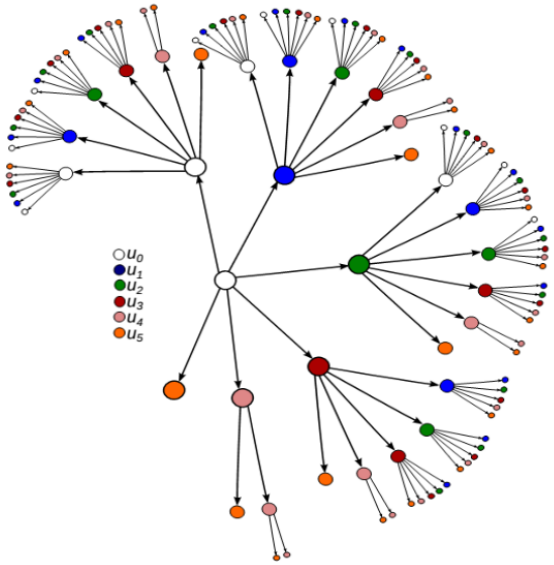


Figura 12. Posibles Caminos

En conclusión este trabajo muestra que el desarrollo de las úlceras en el pie diabético se puede representar mediante la teoría de grafos, en donde los vértices simbolizan cada uno de los estados, y las aristas dirigidas indican los posibles cambios entre ellos. En base a esto, se analizó cada uno de los estados de las úlceras para determinar su reversibilidad, obteniendo que el estado crítico es el u3 de las úlceras. Finalmente, mediante un grafo se puede modelar la interacción entre los diferentes estados de las úlceras en el pie diabético.

3. Técnicas de Análisis de Datos

3.1. Análisis de Componentes Principales

El análisis de componentes principales es una técnica exploratoria mediante la cual se transforma un conjunto de variables correlacionadas en un conjunto menor de variables no correlacionadas llamadas componentes principales, conservando la mayor variabilidad del conjunto de datos. El principal objetivo de esta técnica es eliminar variables y/o información redundante y conseguir una representación gráfica de información multidimensional. Esta técnica se aplica única y exclusivamente a variables cuantitativas.

La obtención de las componentes principales se puede dar mediante varios métodos: [33]

- Buscando la combinación lineal de las variables que máxima la variabilidad (Hotteling).
- Buscando el subespacio de menor ajuste por el método de mínimos cuadrados (Pearson).
- Minimizando la discrepancia de las distancias euclídeas entre los puntos calculados en el espacio original y el subespacio de dimensionalidad reducida (Coordenadas Principales, Gower).

La siguiente fórmula explica las características de la componente principal

$$CP1 = A11V11 + A12V12 + A13V13 + A14V14 + \ldots + A1nV1n$$

$$CP2 = A21V21 + A22V22 + A23V23 + A24V24 + \ldots + A2nV2n$$

En donde Ann se denomina *carga* o *load* y Vnn es la variable de estudio. Las cargas son índices de correlación de la variable con respecto a la componente principal.

Las componentes principales pueden ser de dos tipos, componente de tamaño y componente de forma. La componente principal de tamaño es aquella que posee todas sus cargas positivas. Si una observación tiene alta esta componente, entonces posee valores altos en todas las variables estudiadas, mientras más altos sean los valores de las cargas asociadas más altos serán los valores de las variables. [33]

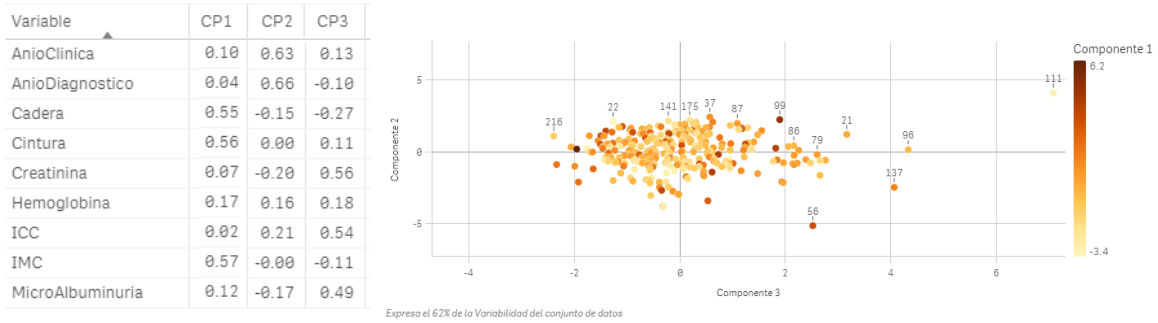
Por otro lado, la componente de forma indica contrastes entre las variables. Si una observación tiene alta esta componente, entonces los contrastes entre las variables también serán altos. De igual manera si una observación tiene valores bajos en esta componente, quiere decir que los contrastes entre las variables son bajos. La intensidad de los contrastes estará determinada por los valores de las cargas. [33]

Ejemplo:

En la siguiente ilustración se puede apreciar un gráfico de dispersión que muestra cómo se encuentran distribuidos las observaciones usando las tres primeras componentes principales.

La componente principal 1 es una componente de tamaño. La observación 111 tiene valores de CP1 muy bajos lo cual implica que tiene valores bajos también en la microalbuminuria, el índice cadera cintura, la creatina, hemoglobina, el año de diagnóstico, año de ingreso a la clínica y especialmente bajos en cadera, cintura e índice de masa corporal.

La componente principal 2 es una componente de forma. Siguiendo con el caso de la observación 111, ésta posee, de entre todas las observaciones, el valor más alto de esta componente. Indicando, de esta manera, que existen grandes contrastes entre el año de diagnóstico, año de ingreso a la clínica, la hemoglobina y el índice cadera cintura contra los valores de cadera, creatinina y la microalbuminuria.



Variable	CP1	CP2	CP3
AnioClinica	0.10	0.63	0.13
AnioDiagnostico	0.04	0.66	-0.10
Cadera	0.55	-0.15	-0.27
Cintura	0.56	0.00	0.11
Creatinina	0.07	-0.20	0.56
Hemoglobina	0.17	0.16	0.18
ICC	0.02	0.21	0.54
IMC	0.57	-0.00	-0.11
MicroAlbuminuria	0.12	-0.17	0.49

Figura 13. Análisis de Componentes Principales

3.2. Análisis Factorial de Correspondencias

Al igual que el análisis de componentes principales, el análisis de correspondencias es una técnica descriptiva cuyo principal objetivo es reducir la cantidad de variables de un conjunto de datos con la menor pérdida de información posible. La diferencia con respecto al análisis de componentes principales es que esta técnica se aplica única y exclusivamente a variables cualitativas. [34]

En otras palabras [34]

> "El análisis de correspondencias busca una representación de coordenadas de filas y columnas de una tabla de contingencia, de modo tal que los patrones de asociación presentes en la tabla se reflejen en dichas coordenadas.
>
> Una tabla de contingencia es un arreglo matricial de números positivos donde en cada casilla se representa la frecuencia absoluta observada para esa combinación de variables"

Dicho de otra manera, esta técnica descriptiva permite encontrar relaciones de similaridad y no similaridad entre las modalidades de diferentes variables cualitativas o de una misma

variable cualitativa. También permite identificar un comportamiento o patrón común dentro del conjunto de datos analizado.

Cuando se aplica sobre dos variables se denomina Análisis Factorial de Correspondencias mientras que cuando intervienen más de dos variables lleva el nombre de Análisis Factorial de Correspondencias Múltiple. Al igual que en el caso del Análisis de Componentes Principales, no debe existir independencia entre las variables para que la aplicación de esta técnica tenga sentido.

Ejemplo:

En el siguiente gráfico de dispersión muestra la aplicación de esta técnica sobre ciertos aspectos antropométricos tales como el estado nutricional, el diámetro de la cintura y la presión arterial.

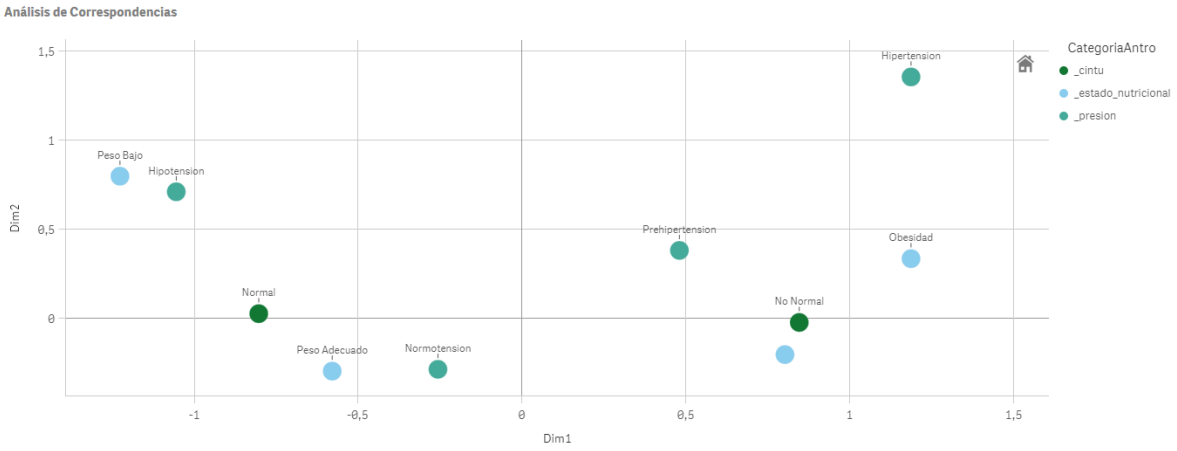


Figura 14. Análisis Factorial de Correspondencias

Este sería un análisis factorial de correspondencias múltiples en el cual intervienen tres variables, la cintura, el estado nutricional y la presión arterial, con sus modalidades correspondientes. La cercanía entre las modalidades de variables distintas indica correlación entre ellas, por lo tanto se puede apreciar que hay una relación alta entre un *Peso Bajo* y la *Hipotensión*, también se puede concluir que un *Peso Adecuado* conlleva tener un diámetro de *Cintura Normal* y *Normotensión*, mientras que al tener problemas de *Sobre Peso* se tienen a tener un *Diámetro de Cintura* fuera de los índices normales y esto puede llevar a que la persona sea *Obesa*.

Finalmente, se puede ver también que la *Hipertensión* se encuentra bastante alejada del resto de modalidades lo cual permite concluir que hay pocos casos de *Hipertensión* respecto al total de casos analizados, y también que no hay una clara correlación entre la *Hipertensión* y el resto de modalidades, salvo con la *Obesidad* y la *Pre Hipertensión* que son los más cercanos a esta. A su vez, también se aprecia que la *Normotensión* es la

modalidad más cercana a la intersección de los ejes, lo cual significa que la mayoría de los casos tiene valores de presión normales.

3.3. Gráfico de Cajas (Boxplot)

El gráfico de cajas o mayormente conocido como *Boxplot* permite representar, de forma sencilla y clara, los aspectos más importantes de la distribución de variables numéricas, basado en medidas de posición tales como la mediana, la media, cuartiles y rango intercuartilico. [35]

Los extremos de la *caja* están definidos por el primer y tercer cuartil. El segundo cuartil o mediana se encuentra dentro de la caja. Los *bigotes* que son las líneas horizontales se encuentran a 1,5 distancias intercuartilicas respecto del primer y tercer cuartil. Las observaciones que se encuentra por arriba del *bigote* superior y por debajo del *bigote* interior son consideradas valores atípicos. Dependiendo de cuán alejados se encuentren podrían llegar a ser atípicos extremos (Figura 15).

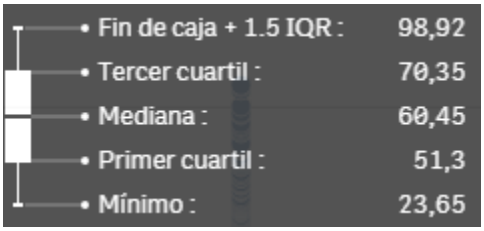


Figura 15. Gráfico de Cajas (Boxplot)

Otra de las utilidades del gráfico de cajas, además de mostrar la distribución de las variables numéricas y de identificar la existencia de valores atípicos es la de comparar la distribución de diferentes variables o de la misma variable en grupos distintos. La siguiente ilustración muestra la distribución de tres variables. En el primer boxplot se puede ver que la mayor parte de los datos se encuentra concentrados en el extremo superior. Este tipo de distribución se conoce como distribución asimétrica negativa o hacia la izquierda.

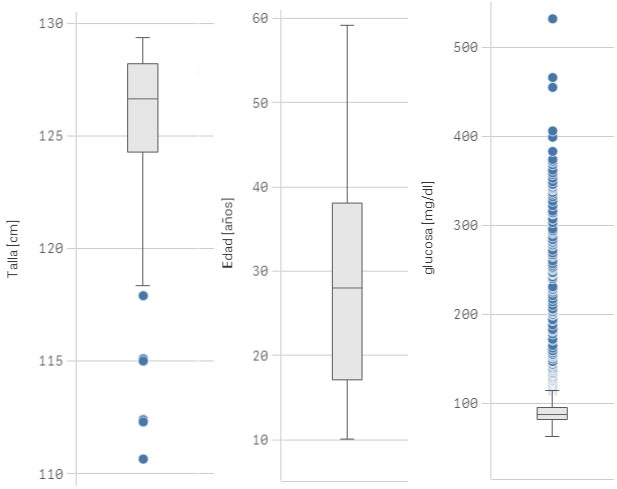


Figura 16. Gráfico de Cajas (Boxplot) de cintura, edad y glucosa

En la figura anterior el boxplot del centro muestra una distribución que tiende a ser simétrica ya que la mediana se encuentra cerca del centro de la caja, tampoco se evidencia la presencia de valores extremos. Finalmente en el tercer boxplot los datos están concentrados en el extremo inferior. Este tipo de distribución se conoce como distribución asimetría positiva o hacia la derecha, también se puede ver la presencia de varios valores extremos.

3.4. Arboles de Decisión

Un árbol de decisión es una representación de una función multivariada la cual está formada por nodos y ramas. Es un método de aprendizaje supervisado, puesto que se conoce el valor de la clase o variable dependiente. Puede ser utilizado tanto para clasificación como para regresión. Se construye a partir de un conjunto de datos de entrenamiento y el resultado que genera es un conjunto de reglas simples y fáciles de interpretar.

El interés del uso práctico de los árboles de decisión tuvo su origen en las ciencias sociales gracias al trabajo de Morgan (Sonquist & Morgan , 1964) en la creación del método AID *Automatic Interaction Detection* este fue uno de los primeros métodos de ajuste de datos basados en árboles de decisión. A partir de este momento, los árboles de decisión trascendieron, de ser únicamente una representación ilustrativa para la toma de decisiones en una herramienta útil y sencilla de utilizar. [37]

Posteriormente Breiman y colaboradores (Breiman, Friedman, Stone, & Olshen, 1984) en su obra *"Classification and regression trees"* aportaron con un método práctico de inducción para la construcción de árboles de decisión de forma recursiva, este método es conocido como CART. Después, se desarrolló el algoritmo ID3 "*Interactive Dichotomiser 3*" que utiliza la entropía de la información para generar árboles (Quinlan J. R., 1986). Subsiguientemente esta obra fue mejorada por su propio autor y denominada C4.5 (1993). A su vez, se introdujo un algoritmo recursivo de clasificación no binario llamado CHAID "*Chi – square automatic detection*" (Kass, 1980)

Ejemplo:

La siguiente ilustración muestra un árbol de decisión que consta de siete nodos finales, las hojas indican si la persona tiene o no diabetes. Salta a la vista que en los nodos 3, 4 y 5 es en donde se concentra la mayor proporción de personas con diabetes. Indicando que a partir de valores de homa (índice de resistencia a la insulina) superiores a 4.5 la tendencia a padecer diabetes es mayor que para valores inferiores a estos.

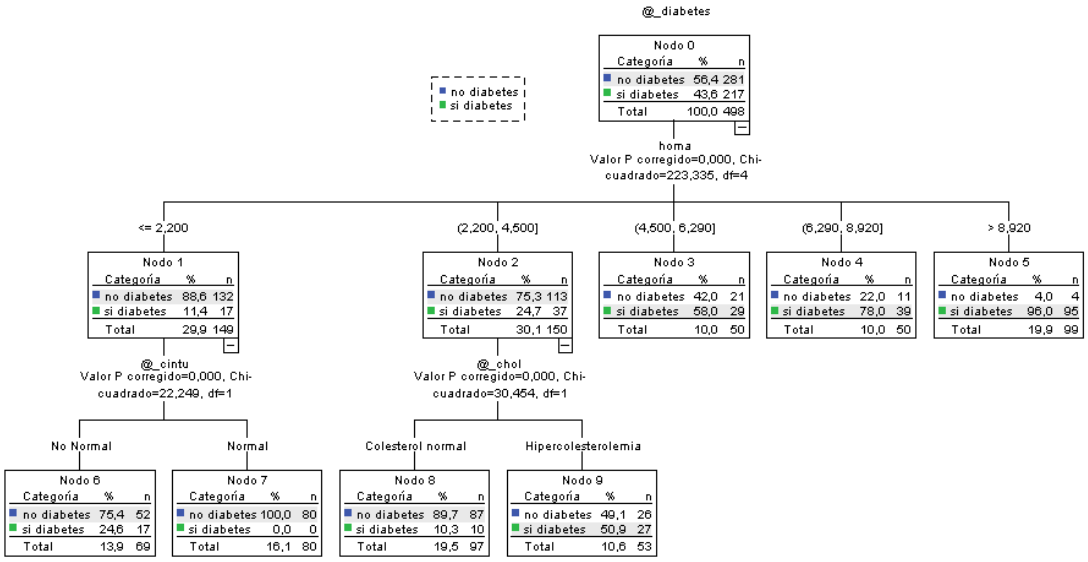


Figura 17. Árbol de decisión de la diabetes

Los árboles de decisión permiten realizar a su vez:

Segmentación: Establecer cuáles son los grupos más importantes para clasificar a una observación.

Clasificación: Asignar a una observación a uno de los grupos generados.

Predicción: Establecer reglas para hacer predicciones de ciertos eventos.

Reducción de dimensiones: Identificar cuáles son las variables más importantes en el análisis de un fenómeno determinado.

Identificación – Interrelación: Identificar las variables y las relaciones más importantes.

Recodificación: Discretizar variables o establecer criterios cualitativos perdiendo la menor cantidad posible de información relevante.

CHAID

Es un método de crecimiento de un árbol de decisión, basado en la detección automática de chi – cuadrado. Este valor es un estadístico que indica independencia de variables, mientras más alto es el estadístico mayor es la dependencia entre las variables. En cada paso el árbol CHAID elige la variable independiente o predictora que presenta la interacción más fuerte con respecto la variable dependiente [37].

3.5. Medidas de Similitud y Disimilitud

Para definir cuán similares son los objetos se utiliza diferentes medidas de proximidad o similaridad en función de la distancia que hay entre ellos. Distancia es una métrica o

función matemática que se aplica a un par de elementos que satisfacen las siguientes condiciones:

No negatividad: La distancia entre los objetos es positiva.

Simetría: La distancia de A hacia B, es la misma que de B hacia A.

Desigualdad triangular:

$$d: XxX \rightarrow R$$

$$d(a,c) \leq d(a,b) + d(b,c) \ \forall\ a,b,c \in X$$

La matriz de similitud $\boldsymbol{S} = (\boldsymbol{d_{ij}})$ y de disimilitud $\boldsymbol{D} = (\boldsymbol{d_{ij}})$ debe cumplir con estas tres propiedades. En algunos casos se puede transformar una matriz $\boldsymbol{S}$ en una matriz $\boldsymbol{D}$

$$d(a,b) = 1 - s(a,b)\ Si\ el\ dominio\ de\ la\ similitud\ es\ [0,1]$$

$$d(a,b) = 1 - \frac{s(a,b)+1}{2}\ Si\ el\ dominio\ de\ la\ similitud\ es\ [-1,1]\ y\ s = -1\ se\ corresponde\ con\ la\ mayor\ distancia\ normalizada$$

3.6. Análisis de Conglomerados (agrupamiento o clustering)

El objetivo fundamental del análisis de conglomerados se puede resumir en clasificar a las observaciones (personas, animales o cosas) en grupos en los que los integrantes tengan características parecidas entre sí y considerablemente distintas al resto de grupos. El análisis de conglomerados es una técnica exploratoria de datos no supervisada enfocada en resolver problemas de clasificación. A diferencia de los árboles de decisión no se conoce el valor de la clase. [38]

En otras palabras, el análisis de conglomerados consiste en agrupar a los individuos de forma tal que el grado de asociación o similitud entre los miembros del mismo conglomerado sea más fuerte que el grado de asociación o similitud entre los miembros de diferentes conglomerados. Los conglomerados en el análisis de conglomerados son equivalentes a las clases en los árboles de decisión. Los algoritmos de clasificación a su vez, pueden dividirse en jerárquicos y no jerárquicos.

Conglomerados Jerárquicos y No Jerárquicos

La diferencia fundamental entre ellos radica en que en el caso de los jerárquicos la clase resultante tiene un número creciente de clases anidadas, mientras que en el segundo las clases no son anidadas. A su vez, los algoritmos jerárquicos pueden ser de tipo divisivos o aglomerativos.

En el primer caso se parte de una clase que engloba todo el conjunto de datos y posteriormente se va dividiendo en clases. Mientras que para el caso de los aglomerativos se parte de la premisa de que cada objeto representa una clase y en los pasos sucesivos se van obteniendo clases de objetos similares. Se utiliza el dendograma para representar la clasificación jerárquica [35]. Ejemplo:

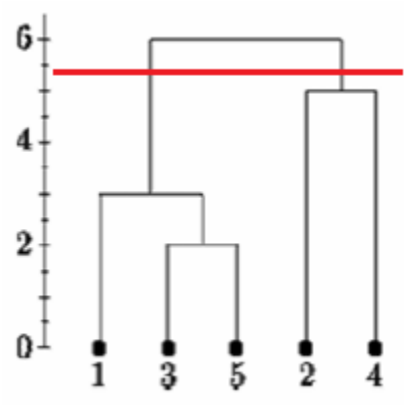


Figura 18. Conglomerados Jerarquicos

En la figura anterior se puede ver un dendograma que establece dos conglomerados, por un lado un conglomerado conformado por los objetos 2 y 4, y otro conglomerado que contiene a los objetos 1, 3 y 5 donde los últimos dos objetos forman parte a su vez, de un subconglomerado.

K medias

Es el método de clúster más utilizado en donde se especifica de antemano la cantidad de conglomerados que se obtendrá. Una vez que se haya escogido la cantidad de conglomerados el algoritmo define aleatoriamente los puntos medios o centroides de cada uno. Posteriormente se asigna cada objeto al centroide del clúster más cercano. Luego se recalcula el centroide o punto medio del clúster y se vuelve a asignar cada objeto al centroide más cercano. El algoritmo termina cuando no convenga realizar más asignaciones [38].

Los métodos de clustering o conglomerados, buscan satisfacer una función de optimización. Este método busca minimizar la suma del cuadrado de las distancias euclideas entre los miembros de un clúster y su respectiva media.

Ejemplo:

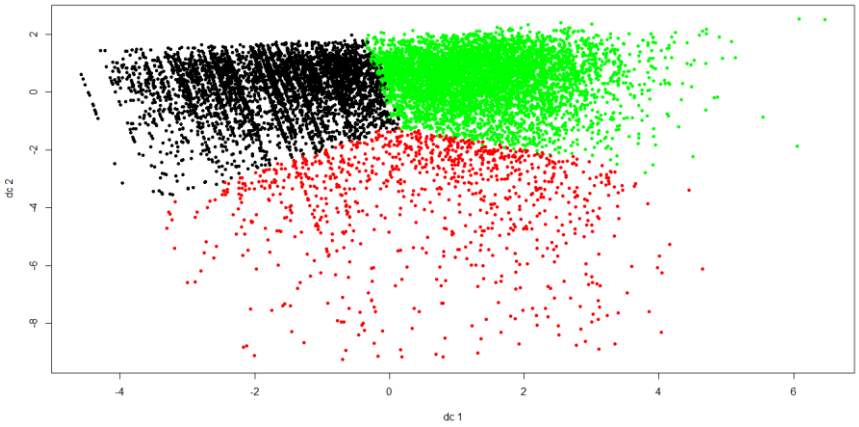


Figura 19. Método k medias

En la figura anterior se observa el resultado de la aplicación de este método de conglomerados sobre un conjunto de datos relacionados con los micronutrientes en una población determinada. Se utiliza las dos primeras componentes principales para representar los datos.

Clustering Difuso

Este método de conglomerados se basa en la lógica difusa. La lógica difusa también conocida como lógica heurística se basa en lo relativo de lo observado como posición diferencial o de análisis. Este tipo de lógica toma dos valores aleatorios pero contextualizados y referidos entre sí. Por ejemplo una persona que mide 2 metros es considerada como una persona alta si previamente se ha establecido que la altura de una persona baja es de 1 metro. Ambos valores están contextualizados a personas y referidos a una medida lineal.

Para el caso del clustering difuso se define una función de membresía para cada objeto y cada clúster que variará entre 0 y 1:

$$z = (z_{ik}) \qquad 0 \leq z_{ik} \leq 1$$

Y de una matriz de membresía $N * K$ de uno a N objetos y de 1 a K clusters. Para determinar esta matriz se debe minimizar el criterio de calidad del clustering difuso:

$$W_F\ (z, c) = \sum_{k=1}^{K} \sum_{i\ \in\ I}^{c} Z_{ik}^{\alpha}\ d(y_i, c_k)$$

Donde $\propto$ modifica la forma de la función de membresía. Se determina de forma empírica y el valor típico es 2. [38]

PAM

Con sus siglas en inglés (*Partition Around Mediods*) es un método de conglomerados por partición similar a k – medias. La diferencia radica en que PAM busca *medoides* o prototipos, es decir, objetos representativos del clúster que tienen una distancia mínima al resto de los miembros de su conglomerado. De forma similar que en el caso de k medias, hay que especificar de antemano la cantidad de conglomerados que obtendremos. En otras palabras, k medias está basado en la media aritmética mientras que PAM en la mediana.

El objetivo fundamental de este método es minimizar la suma de las disimilitudes entre los miembros del conglomerado y su medoide. Esta diferencia en cuanto a la función de optimización con respecto a k medias hace que sea más robusto al trabajar con conjuntos de datos con valores atípicos. Las etapas del algoritmo se resumen de la siguiente manera. [38]

> "**Etapa de construcción:**
>
> Elige k objetos al azar y los utiliza como medoides.
>
> Calcula la matriz de disimilitud
>
> Asigna a cada objeto su medoide más cercano. Para cada conglomerado calcula la sumatoria de las disimilitudes entre los miembros y medoides. Finalmente calcula la sumatoria para todos los conglomerados.
>
> **Etapa de intercambio**
>
> Para cada conglomerado intercambia el medoide con alguno de sus otros miembros y determina si este cambio baja la suma de las disimilitudes del grupo. Si esto ocurre, reemplaza el medoide.
>
> Si se modificó el medoide de alguno de los conglomerados vuelve al paso 3, caso contrario termina.". [38]

Validación de los Conglomerados

Los métodos de agrupamiento siempre van a encontrar grupos, inclusive cuando no existen patrones entre los objetos, por ello es importante considerar varios aspectos tales como:

- Determinar si los datos tienen a ser agrupados o no.
- La cantidad correcta de conglomerados.
- Evaluar cuan bueno es el ajuste del agrupamiento de forma no supervisada.
- Comparar los resultados de los clusters entre sí.

Criterios y medidas para evaluar el agrupamiento

Cohesión (SSE): Mide cuán cerca se encuentran los miembros de un mismo conglomerado respecto al prototipo, centroide para el caso de *k medias* y *medoide* para el caso de *PAM.*

$$\boldsymbol{SSE} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} \sum_{\boldsymbol{x \in c_i}} \boldsymbol{dist}(\boldsymbol{c_i}, \boldsymbol{x})^{\boldsymbol{2}}$$

Separación (SSB): Mide la proximidad entre miembros de diferentes conglomerados o entre prototipos de grupos y el prototipo general.

$$TSE = SSE + SSB$$

Donde TSE es la suma total de cuadrados.

Si las distancias son Euclideas:

$$TSE = \sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} \sum_{\boldsymbol{x \in c_i}} (\boldsymbol{x} - \boldsymbol{c_i})^{\boldsymbol{2}} + \sum_{i=1}^{K} |c_i|(c - c_i)^2$$

Una buena selección de la cantidad de conglomerados se realiza cuando la separación entre los clusters es alta y cuando la cohesión entre los miembros de un clúster es mínima.

Silhouette

El coeficiente silhouette mide la tendencia de los datos a ser agrupados. Este coeficiente varía entre -1 y +1 cuanto mayor sea este mayor será la tendencia de los datos a ser clusterizados. El proceso que se realiza para obtener este coeficiente es el siguiente. [38]

- Para cada objeto i se calcula la distancia promedio a todos los otros objetos de su clúster. A ese valor se lo denomina a_i
- Para el objeto i y todos los clusters que no lo contienen, calcular las distancias promedios a todos los objetos de cada clúster. Al mínimo de todos ellos se lo denomina b_i

Finalmente coeficiente silhouette del objeto i es

$$\boldsymbol{s_i} = \frac{(\boldsymbol{b_i} - \boldsymbol{a_i})}{\boldsymbol{max}(\boldsymbol{a_i} - \boldsymbol{b_i})}$$

Coeficiente Dunn

El coeficiente Dunn tiene como objetivo identificar grupos densos y bien separados. Se define como la relación entre la mínima distancia entre los grupos y la máxima distancia dentro de los grupos o entre miembros de un mismo grupo. Para cada miembro de grupo el índice Dunn se calcula de la siguiente manera:

$$D = \frac{min_{1 \le i < j \le n}\, d(i,j)}{max_{1 \le k \le n} d'(k)}$$

Donde $d(ij)$ representa la distancia entre grupos $i\ y\ j$ y $d'(k)$ mide la distancia intra - grupo del grupo k. La distancia inter – grupo $d(ij)$ entre dos grupos pueden ser cualquier número de medidas de distancia, como la distancia entre los centroides de los grupos. De modo parecido la distancia intra – grupo $d'(k)$ puede ser medida de diferentes formas, como la distancia máxima entre cualquier par de elementos en el grupo k. El criterio interno busca grupos con alta semejanza inter – grupo, algoritmos que producen grupos con coeficiente Dunn altos son los más aceptables [39].

4. Materiales y Métodos

En esta sección se detallan los datos, pasos y herramientas utilizados durante todo el proceso de desarrollo de la herramienta visual. Se utilizó la metodología CRISP para el desarrollo de proyectos de minería de datos, técnica muy utilizada en diversos organismos y empresas el cual constituye un modelo guía de seis fases, *Business Undestanding, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment* [40]

Adicionalmente se trabajó con algunas técnicas de visualización de datos de acuerdo a las propuestas de Daniel Keim para la creación de un sistema de visualizaciones que permita encontrar el equilibrio entre la vista sumarizada o general del conjunto de datos (*overview*) y la vista detallada o particular del conjunto de datos (*details on demand*) en base a una navegación interactiva y en simultaneo del usuario utilizando filtrado y acercamiento de la información (*zooming and filtering*) [41].

4.1. Herramientas de Análisis

SQL Server 2012

Utilicé *SQL Server Integration Services* para realizar todo el proceso de extracción, transformación y carga de información (ETL). Los datos fueron extraídos de la Encuesta Nacional de Salud, Salud Reproductiva y Nutrición de Ecuador (ENSANUT) y depositados en una base de datos *SQL Server*. Esta nueva estructura de datos constituyó el insumo, tanto, para la generación de las visualizaciones como para la aplicación de técnicas de minería de datos en R.

R 3.5

Trabajé con la versión 3.5 de R para aplicar técnicas y algoritmos estadísticos y de minería de datos tales como estandarización de variables, análisis de correlaciones, técnicas de reducción de dimensiones, métodos de segmentación de la población y métodos de clasificación sobre las tablas almacenadas en el motor de base de datos. Posteriormente generé archivos (.csv) con los resultados respectivos.

Qliksense Desktop

Usé la versión de escritorio de Qliksense tanto en el armado del modelo de datos como en la visualización de la información, generando una estructura de datos tal, que permita realizar un análisis univariado, multivariado y predictivo con especial énfasis en las enfermedades crónicas no transmisibles (*ECNT*). Complementando y contribuyendo, de esta manera, al análisis de ENSANUT en la detección de patrones en la vigilancia del estado de salud y nutrición en Ecuador.

4.2. Entendimiento del Tema (*Business Understanding*)

En esta sección señalo las características principales de las bases de datos con las cuales trabajé.

Según el "Tomo 1 Encuesta Nacional de Salud y Nutrición ENSANUT – ECU 2012" el Ministerio de Salud de Ecuador, conjuntamente con el Instituto Nacional de Estadísticas y Censos con el afán de proteger y conocer las condiciones de salud y nutrición del país efectuó un censo a nivel nacional de esta manera obtuvo y actualizó información referente a los siguientes aspectos:

La salud reproductiva materna e infantil, las enfermedades crónicas no transmisibles, la situación nutricional y de consumo alimentario, el estado de lo micronutrientes, el acceso a programas de complementación alimentaria y suplementación profiláctica, la actividad física, el acceso a los servicios de salud y también el gasto en salud de la población ecuatoriana. La población encuestada tenía un rango etario de entre 0 a 59 años.

Posteriormente, el equipo de investigación desarrolló un análisis descriptivo univariado de cada uno de estos aspectos de forma independiente y considerando el rango etario, quintil económico, etnia, niveles de escolaridad, área geográfica, género, región, zona de planificación, subregión, provincia, ciudad, etc. [42].

La finalidad de este análisis fue, entre otras cosas conocer la proporción de personas con problemas nutricionales, encontrar que porcentaje de la población presenta deficiencias de micronutrientes y macronutrientes (vitamina A, vitamina B12, carbohidratos, proteínas, grasas, entre otras). Y, finalmente, determinar la prevalencia de las enfermedades crónicas no transmisibles en la población ecuatoriana, qué según la OMS el 70% de las muertes que se producen en el mundo, son a causa de la ECNT.

El Objetivo principal del presente trabajo fue complementar el análisis de la encuesta ENSANUT - ECU 2012, a través de un sistema de visualizaciones que permita detectar nuevos patrones en la vigilancia del estado de salud y nutrición en Ecuador con especial énfasis en las ECNT y sus complicaciones asociadas a partir, principalmente de los siguientes aspectos:

- Socios demográficos.
- Antropométricos.
- Económicos.
- Bioquímicos.
- Consumo de Alimentos

Este análisis podría permitir, reforzar políticas públicas de salud y educación así como generar informes a la comunidad.

4.3. Entendimiento de los Datos (*Data Understanding*)

Las bases de datos se usaron para el desarrollo de la tesis son de acceso libre y fueron descargadas del sitio oficial del Instituto Ecuatoriano de Estadísticas y Censo (INEC) http://www.ecuadorencifras.gob.ec/category/ensanut/. Las mismas que fueron generadas a partir de la Encuesta de Salud y Nutrición – ENSANUT- ECU realizada desde el 2011 hasta el 2013.

Su diseño muestral permite extrapolar los datos a nivel nacional, subregional, por zonas de planificación, por condición social, por rangos de edad, por etnia y por sexo, a su vez, ofrece un panorama de la dimensión de los problemas estudiados y sus determinantes, con lo cual facilita analizar las respuestas sociales que deben plantearse a cada uno de los problemas investigados.

Si bien ENSANUT- ECU recopila información de múltiples aspectos relacionados con la salud y nutrición, tomé en consideración (con ayuda del especialista en el tema) únicamente las variables relacionadas con las ECNT y las complicaciones asociadas a estas. Las bases de datos utilizadas se encuentran detalladas a continuación.

Nombre	Descripción	Cantidad de Registros	Cantidad de Variables	Campo Clave
ensanut_f1_personas	Aspectos geográfica y demográfica	92502	284	idpers
ensanut_f10_antropometria	Aspectos antropométrica	60629	44	idpers
ensanut_f12_bioquimica	Aspectos bioquímicos	21479	69	idpers
ensanut_f11_consumo_parteb	Aspectos referentes a la alimentación	759700	51	ciudad

Tabla 1. Bases de datos utilizadas

4.4. Pre procesamiento de Datos (*Data Preparation*)

Para la preparación de datos generé una estructura básica de inteligencia de negocios que permitió centralizar y estandarizar los datos que posteriormente fueron minados [43]. Este proceso consistió, fundamentalmente, en crear dos repositorios de información *DATA* y *APP*.

El primer repositorio permitió almacenar los datos tal y como se encuentran en los archivos fuente y el segundo tuvo como objetivo almacenar los datos después del proceso de limpieza. Todo este proceso se resume en la (Figura 20).

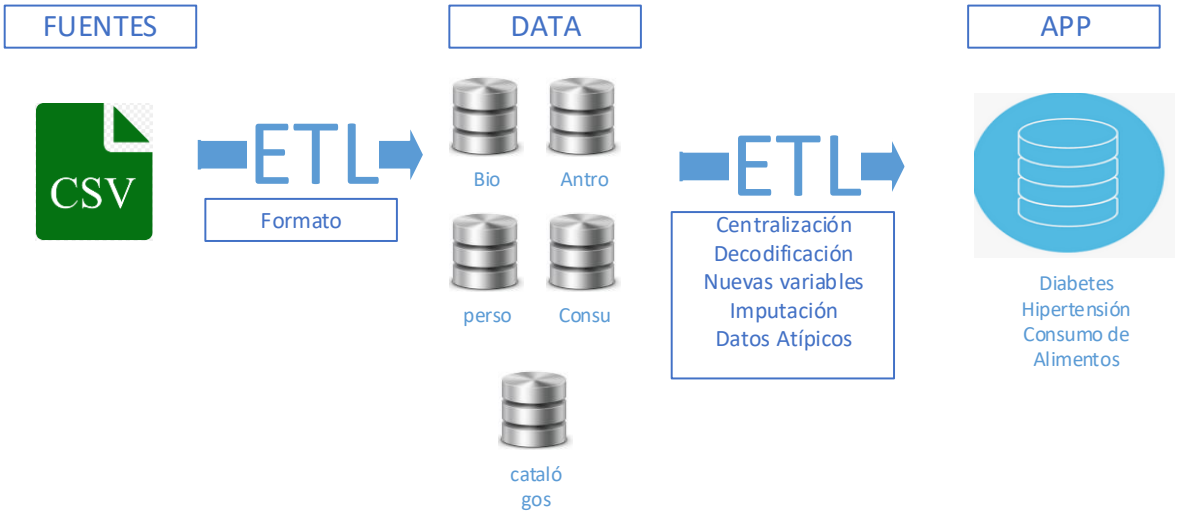


Figura 20. Pre procesamiento de datos

DATA

La creación del repositorio *DATA* se resume de la siguiente manera. Utilizando SQL Server Integration Services se diseñó un proceso de Extracción Transformación y Carga de Datos (*ETL)* por cada una de las fuentes de datos especificadas en la Tabla1, con la finalidad de cargar los datos en la tabla respectiva sin realizar ninguna modificación sobre los mismos.

A su vez, se construyó otro de proceso que permita cargar, en tablas distintas los códigos y las descripciones de las variables codificadas las mismas que fueron extraídas de la base de datos ensanut_f1_personas (Tabla 2).

Nombre	Descripción	Cantidad de Registros
DM_ETNIA	Código y descripción de la etnia	4
DM_QUINTIL	Código y descripción del quintil económico	5
DM_AREA	Código y descripción del área	2
DM_SUBREGION	Código y descripción del subregión	9
DM_PROVINCIA	Código y descripción de la provincia	26
DM_CIUDAD	Código y descripción de la ciudad	1517
DM_NBI	Código y descripción de las necesidades básicas insatisfechas	5
DM_SEXO	Código y descripción del género	2
DM_GRUPO_ALIMENTO	Código y descripción del grupo de alimento	15
DM_CONDICION_ALIMENTO	Código y descripción del grupo de la condición del alimento	7
DM_CATEG_EDAD	Código y descripción de la categoría de edad	12

Tabla 2. Tablas de Aspectos Geográficos, Demográficos y Consumo de Alimentos

APP

En esta etapa se desarrolló todo el proceso de limpieza de datos que incluye transformación, estandarización y creación de nuevas variables, detección de datos atípicos, imputación de datos faltantes, etc. Creé un *ETL* por cada uno de los aspectos a analizar los cuales describo a continuación:

Aspectos Geográficos y Demográficos

Se optó por excluir del conjunto de datos a las mujeres que saben que están embarazadas ya en análisis de los aspectos antropométricos realizados por el Ministerio de Salud sobre ENSANUT también son excluidas [44]. Posteriormente se decodificó variables tales como etnia, quintil, área, subregión, provincia, ciudad, necesidades básicas insatisfechas, género y zona de planificación. Finalmente, creé una tabla llamada *PERSONA* con los aspectos geográficos y demográficos a nivel de persona.

Aspectos Antropométricos

Como primer paso, se decodificó las variables, ciudad y género. Después, debido a que hay tres mediciones de peso, talla, diámetro de la cintura y presión se calculó el promedio de cada una de ellas. También creé una nueva variable que permite identificar si la persona toma o no medicación para la presión. Adicionalmente creé la variable *IMC* dividiendo el peso en kilogramos para la altura al cuadrado [42].

$$IMC = \frac{Peso\ [Kg]}{Altura\ [m^2]}$$

Subsecuentemente, con la ayuda el experto, eliminé valores extremadamente atípicos, e.g. cintura mayor a 700 cm, peso mayor a 400 Kg y menor de 23 Kg, índice de masa corporal mayor a 50 [Kg/m^2] y menores a 10 [Kg/m^2], a su vez, excluí algunos registros por tener valores de presión errados, e.g. presión sistólica mayor a 55 mmHg y presión diastólica menos a 40 mmHg. Tomé en cuenta únicamente a las personas mayores de 10 años.

Posteriormente, creé las variables que indican normalidad de los aspectos antropométricos, tales como, el índice de masa corporal, presión sistólica y diastólica, diámetro de la cintura, adicionalmente generé rangos de edad de diez en diez. [42]. Finalmente creé una tabla llamada *ANTROPOMETRIA* con los aspectos antropométricos a nivel de persona.

Aspectos Bioquímicos

Primeramente se decodificó la variable sexo y se cargó únicamente las personas mayores a 10 años. Acto seguido se construyó las variables nuevas que indican normalidad en los aspectos bioquímicos tales como, insulina, colesterol, triglicéridos, glucosa, colesterol de alta densidad, colesterol de baja densidad, índice homa, glucosa, ferritina y hemoglobina [42].

De manera similar que en el caso de los aspectos antropométricos se eliminó varios registros con errores atribuibles a fallas en la de digitación e.g. glucosa superior a 750 mg/dl, índice homa mayor a 20, colesterol mayor a 760 mg/dL, hdlc superior a 150 mg/dL, ldlc superiores a 300 mg/dL, triglicéridos mayores a 1200 mg/dL, niveles de insulina superiores a 100 μU/mL, hemoglobina entre 6.6 g/dL y 24.4 g/dL [43].

Finalmente se construyó una tabla llamada *BIOQUIMICA* con los aspectos bioquímicos a nivel de persona.

Aspectos Alimenticios

Se inició el proceso excluyendo la ciudad San Antonio, debido a que las características del conjunto de datos no me permitieron identificar a qué provincia pertenece. Posteriormente, se decodificó las variables grupo alimenticio, condición del alimento, categoría de edad y ciudad. Finalmente se construyó una tabla llamada *CONSUMOALIMENTOS_TXT* con estos datos.

Adicionalmente, se calculó los valores medios por provincia debido a que ese es el nivel jerárquico más bajo para las variables carbohidrato, hierro, grasa, proteína, vitamina B12, vitamina A y zinc. Obteniendo la tabla *CONSUMOALIMENTOS_NUM* con estos datos.

Diabetes e Hipertensión

Como punto inicial, se unificó las tablas *PERSONA, ANTROPOMETRIA, BIOQUIMICA* generando una tabla llamada *DIABETES_HIPERTENSION* con 13687 registros. Esta tabla registra 182 casos de personas con diabetes y 709 casos de personas con hipertensión.

Vale la pena aclarar que los casos de diabetes fueron determinados evaluando los niveles de glucosa en ayunas, de ser mayor a 126 mg/dL catalogué a la persona como diabética. Sin embargo seguramente en la base de datos hay casos de diabetes pero con niveles normales de glucosa debido al tratamiento, lamentablemente no hay forma de determinar cuáles son esos casos. Por otro lado para la hipertensión, si la persona tenía una presión sistólica mayor de 140 mmHg o más de 90 mmHg de presión diastólica, dicha persona fue catalogada como hipertensa.

Sin embargo, esto no quiere decir que esos sean los únicos casos de diabetes e hipertensión en el conjunto de datos, ya que puede haber casos en los cuales algunas personas padezcan estas enfermedades pero al estar siendo tratadas sus niveles de glucosa y presión son normales.

Para el tratamiento de valores faltantes, se imputó el promedio por ciudad. Este proceso lo realicé sobre las siguientes variables, presión sistólica y diastólica, quintil, cintura, insulina, colesterol, triglicéridos, glucosa, colesterol de alta densidad, colesterol de baja densidad, índice homa, glucosa y ferritina.

Finalmente, se creó una tabla llamada *DIABETES_HIPERTENSION_PROVINCIA* con los promedios de las variables bioquímicas, antropométricas y consumo de alimentos a nivel de las 24 provincias.

La siguiente tabla describe cada una de las variables utilizadas, originales y las creadas en el proceso.

Variable	Descripción	Aspecto	Tipo Variable
presa	Presión Sistólica	Antropométrico	Numérica
presb	Presión Diástolica	Antropométrico	Numérica
peso	Peso	Antropométrico	Numérica
talla	Talla	Antropométrico	Numérica
cintu	cintura	Antropométrico	Numérica
_presion	Indica si tiene hipertension, prehipertensión, normotensión o hipotensión	Antropométrico	Categórica
_imc	Indice de masa corporal	Antropométrico	Categórica
_estado_nutricional	Estado nutricional	Antropométrico	Categórica
_cintu	Valores normales de cintura	Antropométrico	Categórica
medicacionPresion	Usa o no medicación para la presión	Bioquímico	Categórica
glucosa	glucosa en la sangre	Bioquímico	Numérica
insulina	niveles de insulina	Bioquímico	Numérica
chol	colesterol	Bioquímico	Numérica
hdlc	Colesterol lipoproteínico de alta densidad	Bioquímico	Numérica
ldlc	Colesterol lipoproteínico de baja densidad	Bioquímico	Numérica
trig	trigliceridos	Bioquímico	Numérica
ferritin	ferritina	Bioquímico	Numérica
hb	hemoglobina	Bioquímico	Numérica
_chol	Indica valores normales de colesterol	Bioquímico	Categórica
_insulina	Indica valores normales de insulina	Bioquímico	Categórica
_trig	Indica valores normales de trigliceridos	Bioquímico	Categórica
_diabetes	Indica si tiene o no diabetes	Bioquímico	Categórica
_hdlc	Indica valores normales de Colesterol lipoproteínico de alta densidad	Bioquímico	Categórica
_ldlc	Indica valores normales de Colesterol lipoproteínico de baja densidad	Bioquímico	Categórica
_resistencia_insulina	Indica si presenta o no resistencia a la insulina	Bioquímico	Categórica
_glucosa	Indica valores normales de glucosa	Bioquímico	Categórica
grupoAlimentos	Grupo de alimento	Consumo Alimentos	Categórica
condicionAlimentos	Condición del alimento	Consumo Alimentos	Categórica
carbohidrato	Carbohidratos	Consumo Alimentos	Numérica
hierro	Hierro	Consumo Alimentos	Numérica
grasa	Grasa	Consumo Alimentos	Numérica
proteina	Proteína	Consumo Alimentos	Numérica
vitaB12	Vitamina B12	Consumo Alimentos	Numérica
vitaA	Vitamina A	Consumo Alimentos	Numérica
zinc	Zinc	Consumo Alimentos	Numérica
etnia	Etnia a la que pertenece	Demográfico	Categórica
quintil	Quintil económico	Demográfico	Categórica
nbi	Necesidades Básicas Insatisfechas	Demográfico	Categórica
edad	Edad	Demográfico	Categórica
sexo	genero	Demográfico	Categórica
_grupo_edad	Grupo de edad	Demográfico	Categórica
area	Area geográfica donde vive	Geográfico	Categórica
subregion	Subregión geográfica	Geográfico	Categórica
ciudad	Ciudad	Geográfico	Categórica
zonaPlanificacion	Zona de planificación	Geográfico	Categórica
provincia	Provincia	Geográfico	Categórica

Tabla 3. Características de las variables utilizadas

Por otro lado la siguiente tabla resume los criterios utilizados para la inclusión de los casos a analizar:

Condición	Unidades
edad >= 10	[años]
23 <= peso <= 399	[kg]
10 <= imc <= 50	[kg/m2]
presb > 40	[mmHg]
presa > 55	[mmHg]
peso >23	[kg]
talla >100	[cm]
glucosa < 750	[mg/dl]
homa < 20	[µU/mL]
insulina < 91.6	[µU/mL]
chol < 760	[mg/dL]
hdlc < 150	[mg/dL]
trig < 1200	[mg/dL]
6.6 <= hb <= 24.4	[g/dL]
insulina < 100	[µU/mL]
ldlc < 300	[mg/dL]

Tabla 4. Criterios de inclusión de casos a analizar

4.5. Modelado Visualización y Evaluación de Resultados (*Modelling, Visualization and Evaluation*)

Culminado el pre procesamiento, para el minado de datos, se decidió tomar dos enfoques de análisis; el primero a nivel de variable y el segundo a nivel de población.

Para llevar a cabo el primer estudio se realizó un análisis univariado de cada una de las variables que intervienen de alguna manera con la diabetes y la hipertensión. Finalmente se construyó de forma secuencial árboles de decisión con la finalidad de hallar las diez variables más influyentes en las ECNT considerando la ganancia de información a través del valor de *chi cuadrado*, generado con el método *CHAID*.

Para llevar a cabo el segundo estudio se realizó un análisis descriptivo univariado y multivariado por cada uno de los aspectos señalados en la Tabla 1. Con la finalidad de satisfacer los *objetivos del negocio* (Ver punto 4.2); y para facilitar la interpretación de los resultados, se decidió hacer que el análisis multivariado sea a nivel de provincia.

A su vez, se realizó tanto, la estandarización de las variables como un análisis de correlaciones. Luego se aplicó las siguientes técnicas de reducción de dimensiones: análisis de correspondencias, y análisis de componentes principales. Posteriormente, se utilizó los siguientes métodos de segmentación de la población: k – medias, fuzzy, PAM y se usó el coeficiente silhouette para identificar la tendencia al *cluster* de los datos.

Finalmente, se usó tanto, los resultados obtenidos en el proceso de minado, como las tablas finales de Diabetes, Hipertensión y Consumo de Alimentos (ver punto 4.4.) para construir un sistema de visualizaciones que permita complementar el análisis de ENSANUT. El siguiente gráfico resume este procedimiento.

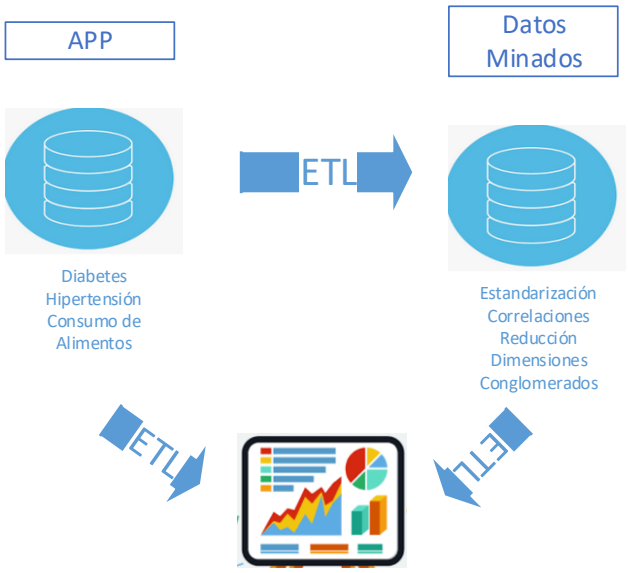


Figura 21. Carga al Sistema de Visualizaciones

5. Resultados

5.1. Herramienta de Visualización

En la figura 22 se puede apreciar el modelo de datos generado, la tabla central se llama "DiabetesHipertension"; tablas tales como "ConsumoAlimentos_TXT" y "Provincia", se asocian a ésta de forma directa mediante el campo "provincia". Por otro lado, las tablas "ScaladasProvincia", "ClustersProvincia", "CPProvincia", "ConsumoAlimentos_NUM" se asocian a la tabla "DiabetesHipertension" usando la tabla de "Provincia" como tabla intermedia.

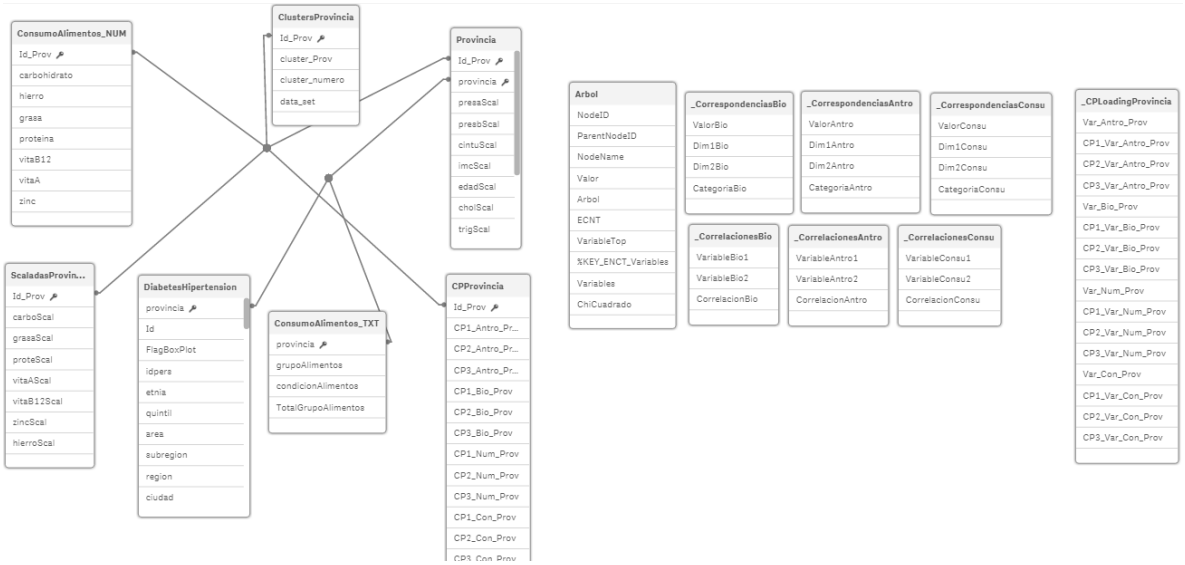


Figura 22. Modelo de Datos, Aplicación de Diabetes e Hipertensión

Adicionalmente, las tablas que se encuentran al costado derecho se conocen como tablas *islas* ya que si bien, forman parte del modelo de datos, no están relacionadas entre sí. Estas tablas tienen información relacionada con los árboles de decisión, análisis de correspondencias y análisis de correlaciones.

La aplicación de visualización de datos desarrollada, consta de 14 *hojas* o lienzos en los cuales se realizaron tanto los análisis a nivel de variable como a nivel poblacional utilizando las técnicas de minería de datos descritas en el punto 4 (Figura 23). Gracias al modelo de datos asociativo de Qliksense las selecciones o filtros se conservan al cambiar de una hoja a otra.

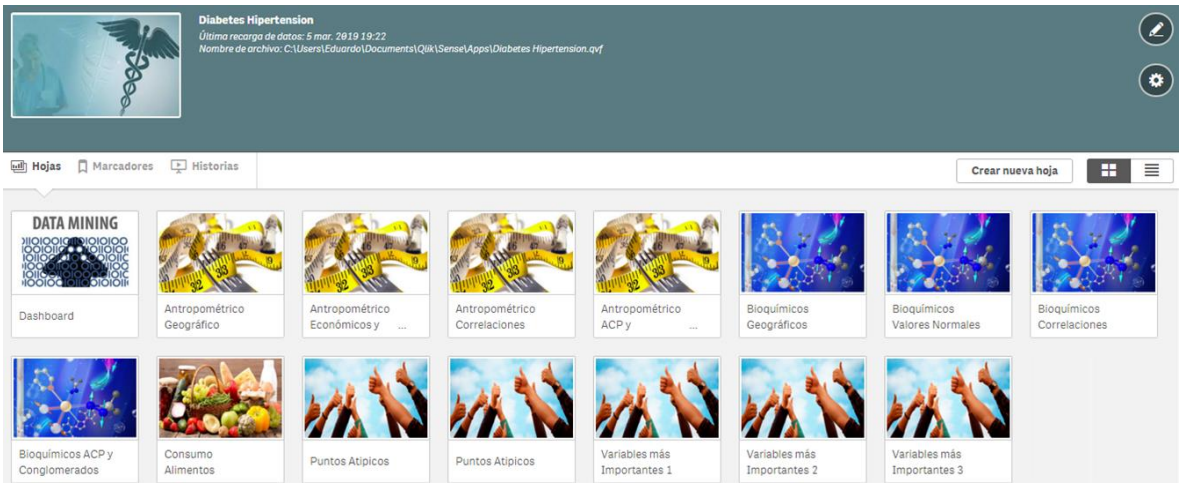


Figura 23. Aplicación de Diabetes e Hipertensión

Resulta importante explicar dos conceptos de análisis de datos, las *métricas* y las *dimensiones*, las métricas son las variables que se desea visualizar, mientras que las dimensiones con los campos o valores mediante los cuales las métricas serán visualizadas. E.g. queremos ver el peso en kilogramos agrupado por ciudad, la ciudad sería la dimensión mientras que el peso sería la métrica.

En la figura 24 se puede apreciar la hoja titulada "Antropométrico Geográfico" donde se realizó un análisis geográfico de los aspectos antropométricos, destacando las selecciones actuales en la parte superior, los filtros o dimensiones (área, zona, región, subregión, provincia, ciudad, sexo, edad, etnia, estado nutricional) en el extremo superior derecho y las métricas (imc, talla, peso, personas) en la parte superior izquierda. Cabe resaltar que los filtros, métricas y gráficos pueden estar en cualquier parte de la hoja.

Esta hoja consta de cuatro gráficos, cuatro botones y cuatro *KPIs* (Key Performance Indicators). Uno de los gráficos es el mapa geográfico, el mismo que muestra los niveles del aspecto antropométrico seleccionado (imc, talla, peso, personas) en cada una de las provincias del Ecuador; usa intervalos de colores que van desde azul oscuro a marrón oscuro. El gráfico ubicado en la parte superior central muestra el estado nutricional por área geográfica, usando un gráfico llamado *dependency wheel.* La selección de la métrica o variable a analizar se la hace dando click en los botones ubicados en la parte superior izquierda de la hoja.

Esta hoja también consta de un gráfico de dispersión que permite analizar el peso estandarizado en el eje "y", la talla estandarizada en el eje "x" y la población encuestada que se representa con el tamaño de las burbujas que, a su vez representan a las provincias del país. Adicionalmente la región a la que pertenecen las provincias está representada por con un color distinto. Por último se desarrolló un gráfico de barras que muestra la métrica

seleccionada (en este caso el peso), por una de las dimensiones geográficas seleccionadas (que pueden ser área, zona, región, subregión, provincia, ciudad).

E.g. la figura 24 muestra el peso, la talla, el imc y la población correspondiente a la etnia indígena con problemas de obesidad y peso bajo, en donde las personas con mayor peso se encuentran en las provincias de Santa Elena y Santo Domingo de los Tsáchilas, mientras que los menos pesados están en Guayas y Esmeraldas. Las *selecciones actuales* se pueden ver en la esquina superior izquierda de la hoja y en los valores de los filtros con fondo verde. Si no está seleccionado ningún filtro la hoja muestra todos los datos disponibles. Adicionalmente, esta herramienta permite confirmar la selección realizada dando click en el botón verde (ver filtro etnia). Los botones "izquierda" y "derecha" ubicadas al extremo superior derecho permiten navegar entre hojas.

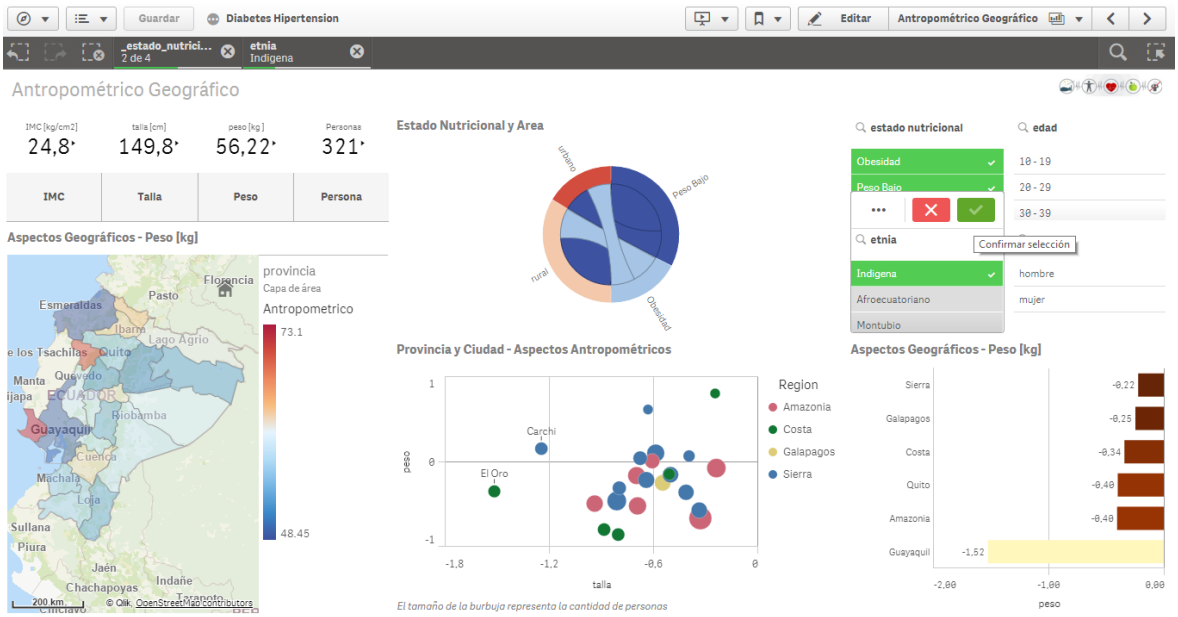


Figura 24. Hoja de Aspectos Antropométricos – Geográficos

La figura 25, muestra la hoja titulada "Económicos y Demográficos" donde se desarrolló un análisis de correspondencias de los aspectos económicos y demográficos de la población ecuatoriana, para el ejemplo se excluyó de la selección el sexo o el género. En la parte derecha se encuentra el filtro de las categorías.

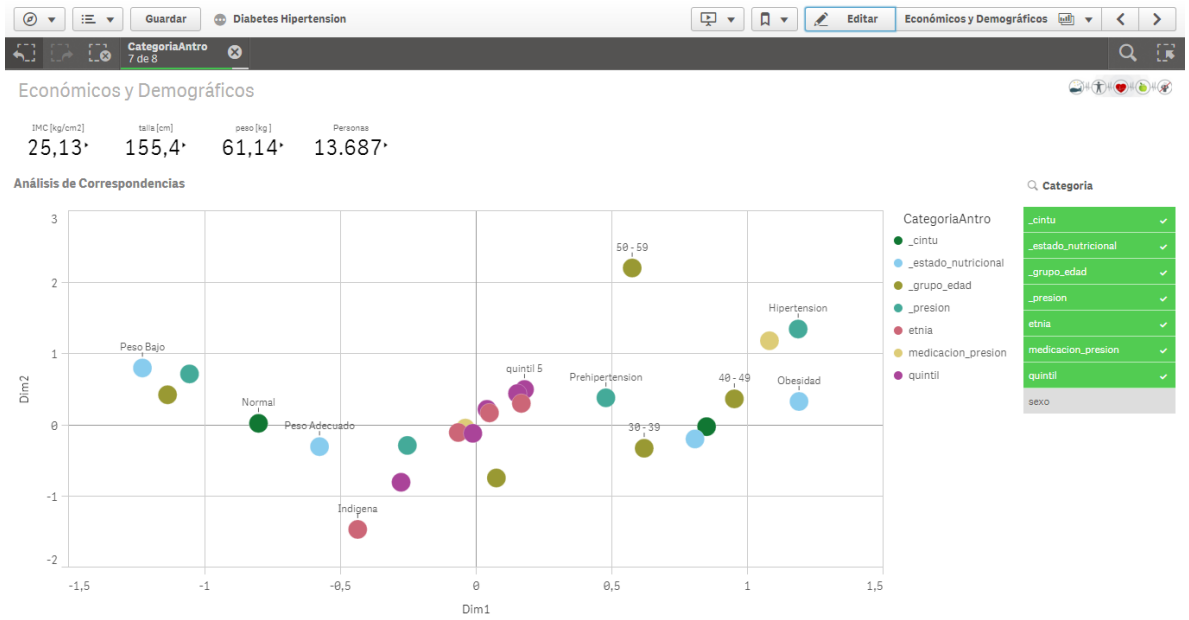


Figura 25. Hoja de Aspectos Antropométricos – Geográficos

La figura 26 muestra en la hoja titulada “Antropométrico Correlaciones” un mapa de calor que indica cuán corelacionadas se encuentra las variables antropométricas utilizando el coeficiente de correlación de pearson. La forma de navegar en esta hoja es la misma que en los casos anteriores, la matriz de correlaciones se vera afectada por la selección de las variables ubicadas a la izquieda de la hoja. Si no se seleccionada ninguna de las variables el mapa de calor muestra todas las correlaciones.

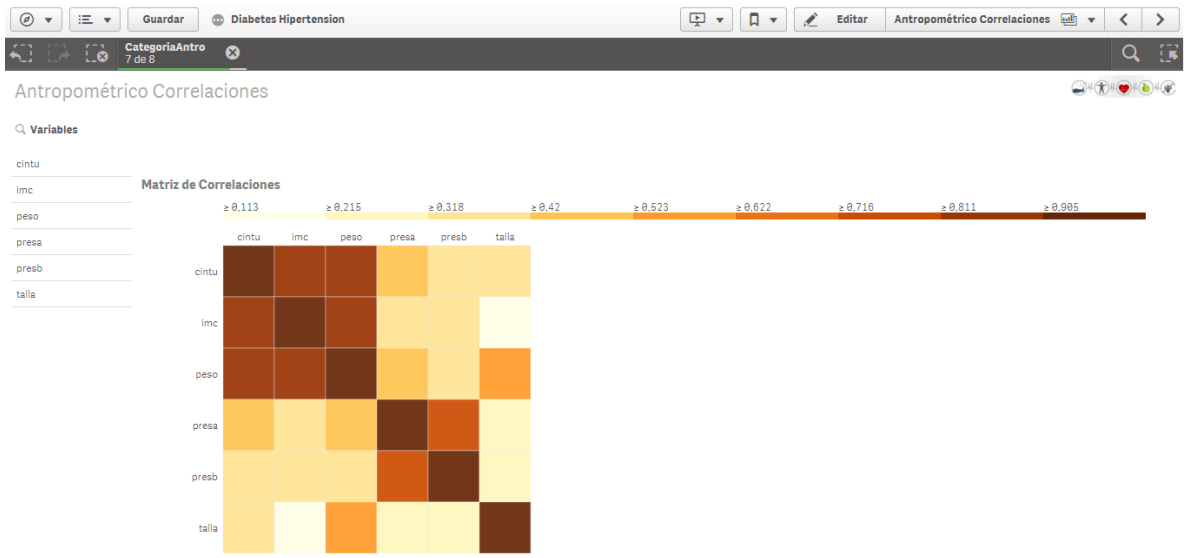


Figura 26. Hoja de Matriz de Correlaciones Aspectos Antropométricos

En la figura 27 se puede apreciar la hoja titulada “Antropometricos ACP y Conglomerados” en donde se realiza un análisis de componentes principales y conglomerados de los

aspectos antropométricos. Por defecto las visualizaciones indican los resultados de aplicar k medias igual a 2. En esta hoja, los filtros se encuentra ubicados en la parte superior izquierda. Se puede escoger el método de clustering de k=2 o k=3, también el cluster o clusters que queremos mostrar (1, 2 o 3) junto con la región que se precise analizar. Adicionalmente, se puede seleccionar las provincias simplemente dando click en una o varias de ellas en el mapa ubicado a la derecha de la hoja.

En esta hoja hay un gráfico de dispersión que muestra los valores de las componentes principales en cada una de las provincias y el color de las mismas esta determinado por el número de cluster al que pertenence (lo mismo sucede en el caso del mapa de la derecha). Adicionalmente, hay *boxplots* por cada una de las variables mostrando su distribución en cada uno de los clusters. E.g. En la figura 27 se muestra el resultado de aplicar k medias con k igual a 2 y excluyendo a las Islas Galapágos de la selección.

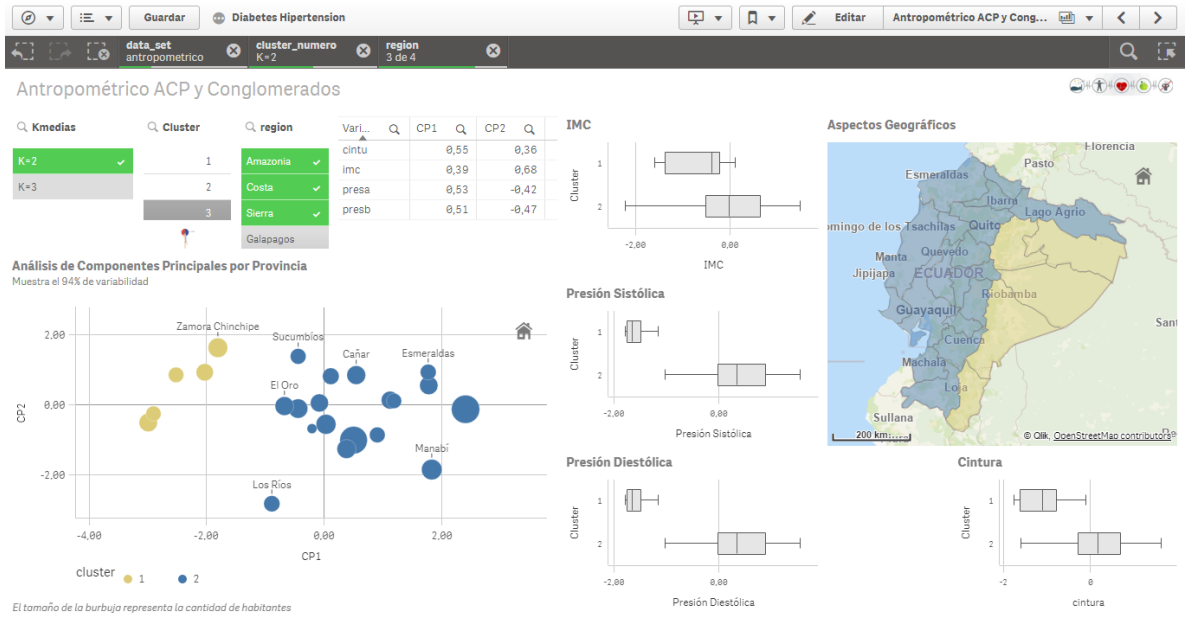


Figura 27. Hoja de ACP, Conglomerados de Aspectos Antropométricos

En la figura 27 se aprecia la hoja titulada “Bioquímicos Geográficos” donde se realizó un análisis geográfico de los aspectos bioquímicos. Los gráficos, botones y filtros están dispuestos de forma similiar que en la hoja “Antropométrico Geográfico”, los botones ubicados en la parte superior izquierda definen la métrica o variable a analizar, tanto en el mapa geográfico (ubicado en el extremo inferior izquierdo), como en el gráfico de barras (al costado derecho) y en el *dependency wheel* en la parte superior.

El gráfico de dispersión permite analizar el colesterol lipoproteínico de alta densidad (hdlc) en el eje “x”, el colesterol lipoproteínico de baja densidad (ldlc) en el eje “y”, y los niveles de colesterol representados por el tamaño de las burbujas, mismas que representan cada una de las provincias del país. El color de las burbujas indica la región a la que pertenece cada

provincia. E.g. en la figura 28 se puede apreciar los niveles de colesterol, ldlc, hdlc de las personas con problemas de sobrepeso, de etnia afroecuatoriana e indigena que tienen entre 20 a 39 años, agrupados por provincia, área geográfica y valores normales de colesterol.

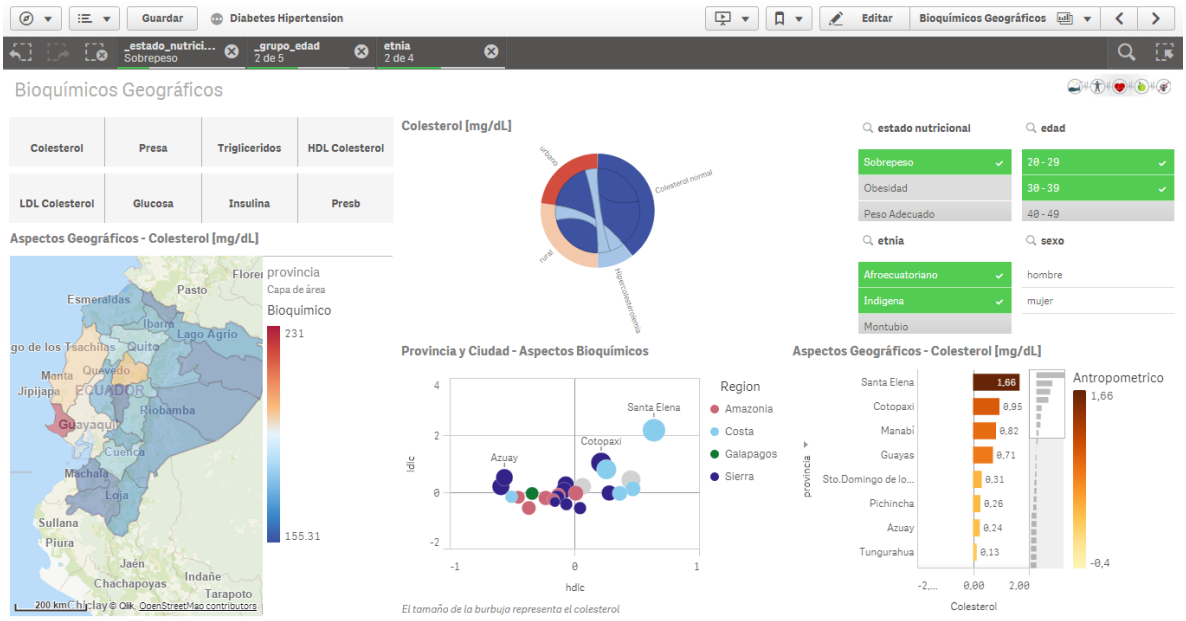


Figura 28. Hoja de Aspectos Bioquímicos – Geográficos

En la figura 29 se puede apreciar la hoja “Bioquímicos Valores Normales” en la que se realizó un análisis de correpondencias de los valores normales de los aspectos bioquímicos de la población ecuatoriana. Las variables utilizadas y sus valores normales se encuentran en la tabla 4. Al costado derecho de la hoja se puede seleccionar las variables a ser estudiadas, al dar click en el cuadro de contorno negro ubicado al extremo superior derecho del gráfico de dispersión se brinda la oportunidad de seleccionar un rango de puntos, tal y como muestra la figura 29.

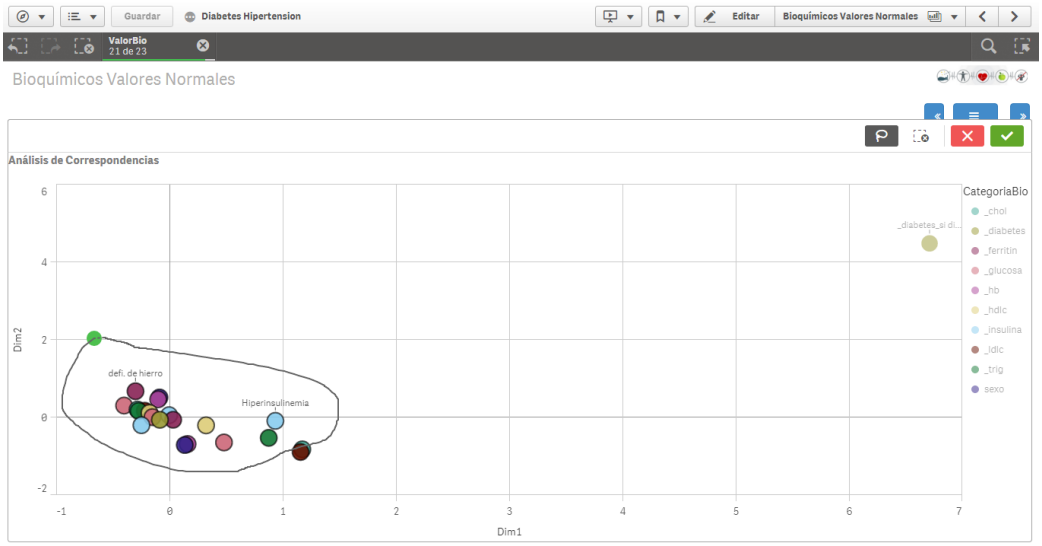


Figura 29. Hoja de Aspectos Bioquímicos – Valores Normales

La figura 30 muestra en la hoja titulada “Bioquímicos Correlaciones” un mapa de calor que indica cuán corelacionadas se encuentran las variables bioquímicas, incluyendo la presión sistólica y diastólica utilizando el coeficiente de correlación de pearson. La forma de navegar en esta hoja es la misma que en los casos anteriores, la matriz de correlaciones se verá afectada por la selección de las variables ubicadas a la izquieda de la hoja. Si no se seleccionada ninguna de las variables el mapa de calor muestra todas las correlaciones. E.g. en este caso se visualizan todas las variables disponibles a excepción de la presión diastólica.

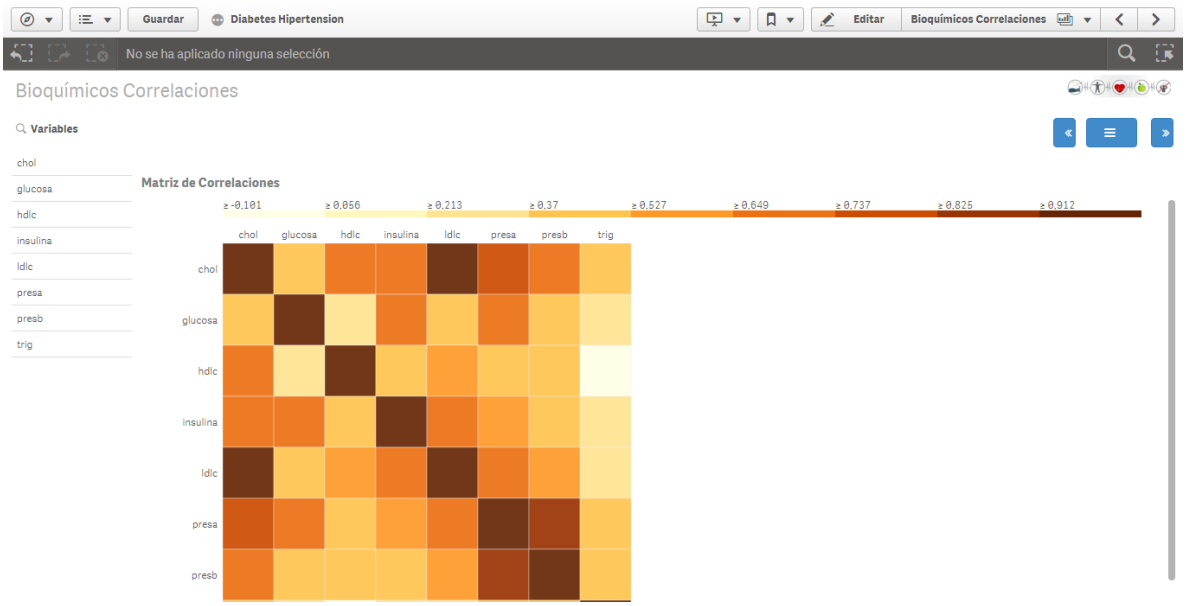


Figura 30. Hoja de Matriz de Correlaciones Aspectos Bioquímicos

En la figura 31 se puede apreciar la hoja titulada “Bioquímicos ACP y Conglomerados” en donde se realiza un análisis de componentes principales y conglomerados de los aspectos bioquímicos. Por defecto, las visualizaciones indican los resultados de aplicar k medias igual a 2. En esta hoja, los filtros se encuentra ubicados en la parte superior izquierda. Se puede escoger el método de clustering de k=2 o k=3, también el cluster o clusters que queremos mostrar (1, 2 o 3) junto con la región que se precise analizar. Adicionalmente, se puede seleccionar las provincias simplemente dando click en una o varias de ellas en el mapa ubicado a la derecha de la hoja.

En esta hoja hay un gráfico de dispersión que muestra los valores de las componentes principales en cada una de las provincias y el color de las mismas esta determinado por el número del cluster al que pertenence (lo mismo sucede en el caso del mapa de la derecha). Adicionalmente, hay *boxplots* por cada una de las variables mostrando su distribución en cada cada clusters. E.g. En la figura 31 se muestra el resultado de aplicar k medias con k igual a 2 y excluyendo la provincia del Guayas de la selección.

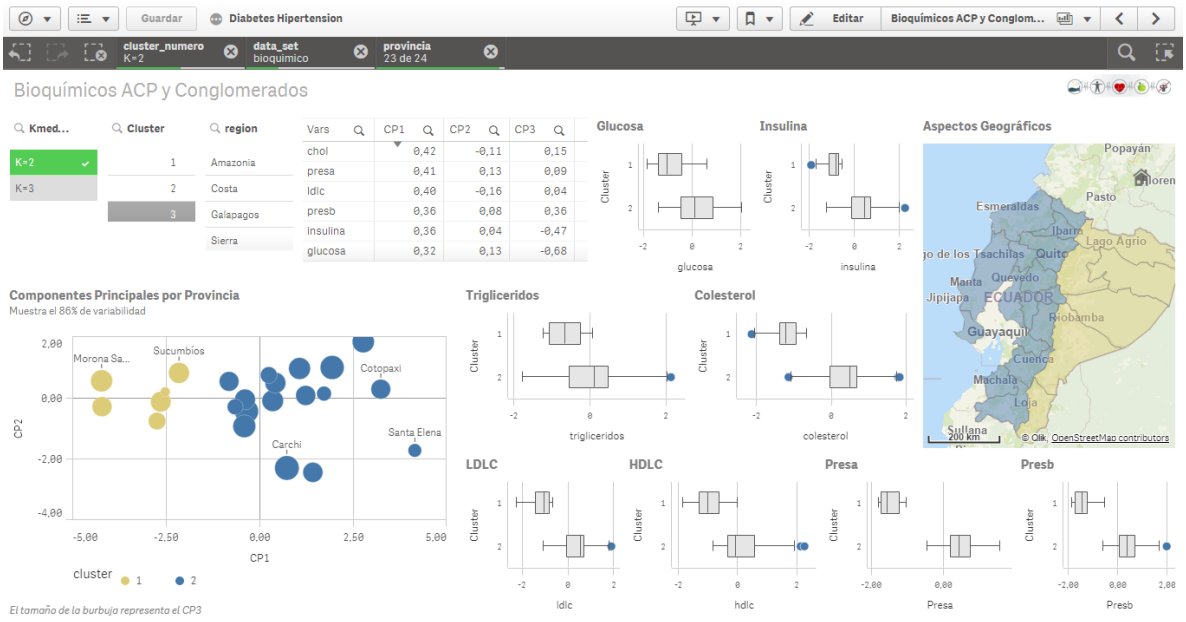


Figura 31. Hoja de ACP, Conglomerados de Aspectos Bioquímicos

En la figura 32 se puede apreciar la hoja titulada “Consumo Alimentos” en donde se hace un análisis del consumo de grasas, proteinas y carbohidratos de las provincias y ciudades del país así como la proporción de personas con diabetes e hipertensión. Los filtros se encuentran ubicados en el extremo superior izquierdo. En el gráfico de dispersión ubicado en el extremo inferior izquierdo de la hoja se representa a las provincias los niveles de proteína en el eje “y”, las grasas en el eje “x” y el consumo de carbohidratos representado por el volumen de las burbujas las mismas que representan cada una de las provincias del país.

El color de las burbujas indica la región geográfica a la que pertenecen. A la derecha del gráfico de dispersión se puede apreciar dos gráficos de barras que indican la proporción de personas con enfermedades crónicas no transmisibles agrupadas por provincia y ciudad. En estos gráficos se puede realizar un *drill down* por provincia y ciudad, es decir, al seleccionar una de las provincias el gráfico mostrará las ciudades de esta provincia. E.g. en la figura 32 se puede ver el consumo de grasas, proteínas y carbohidratos de las provincias, así como la proporción de personas con enfermedades crónicas no transmisibles agrupadas por provincia, en base a la selección realizada, es decir, para la región Costa, Sierra y Amazonía.

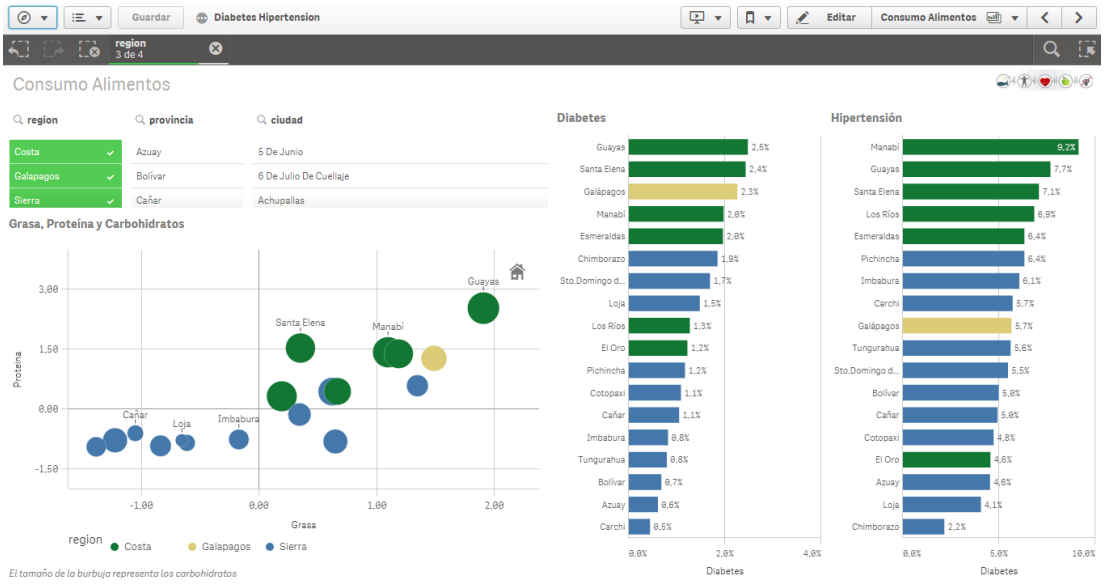


Figura 32. Hoja de Consumo de Alimentos

En la figura 33 se puede apreciar las hojas llamadas “Puntos Atípicos” en donde se indica la distribución de cada una de las variables numéricas usando *boxplots.* En la parte izquierda de las hojas se encuentran los filtros de selección, los cuales son región, provincia, étnia y edad, mientras que en la parte derecha están los *KPIs* de la cantidad de personas seleccionadas, la cantidad de personas con diabetes y las que tiene hipertensión. Finalmente, a la derecha de los filtros se encuentran los boxplots de las variables numéricas respectivas. E.g. en la figura 33 se puede ver la distribución de las variables númericas para las personas que habitan en la amazonía y en la costa, que son de etnia afrodescendiente, indigena y montuvio; y que a su vez, tienen entre los 10 y los 39 años de edad.

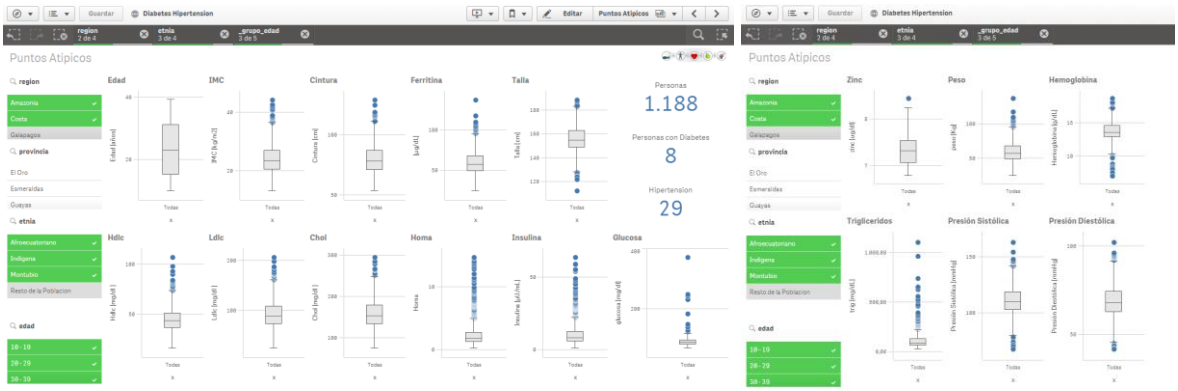


Figura 33. Hoja de Puntos Atípicos

Las siguientes tres hojas hacen referencia al análisis de las variables más importantes para la diabetes y la hipertensión (Figura 34), en el extremo superior izquierdo se encuentra un *radio button* que permite seleccionar una de las dos ECNT. Por defecto, en la hoja “Variables más Importantes 1” aparecerá seleccionado “Diabetes” , mostrando, de esta

manera dos cosas. A la izquierda de la hoja, un gráfico de barras que muestra las variables más importantes en base al valor del chi cuadrado, y, a la derecha se encuentran los árboles de decisión corrrespondientes a las primeras cuatro variables más importantes de la ECNT seleccionada. Las dos hojas restantes de las "variables más Importantes" funcionan de la misma forma que la primera, la diferencia es que la hoja "Variables más Importantes 2" muestra la quinta, sexta, septima y octava variable más importante, mientras que la hoja "Variables más Importantes 3" muestra las dos variables restantes.

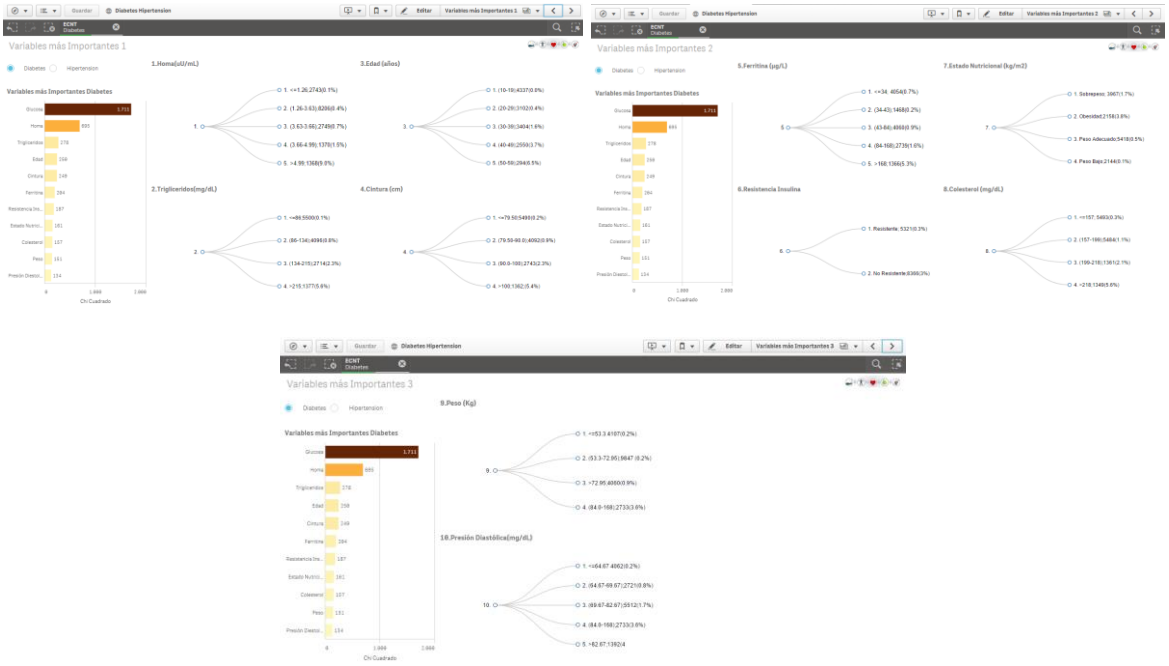
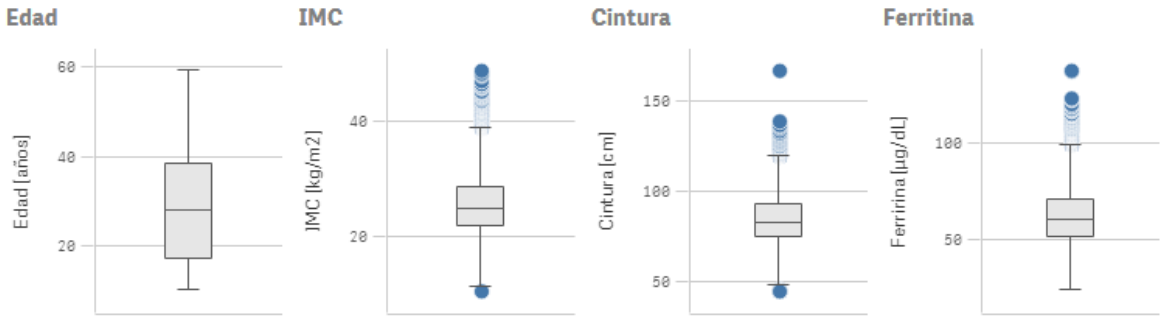


Figura 34. Hojas Variables más Importantes

5.2. Análisis a nivel de Variable

5.2.1. Análisis Univariado

En esta sección se utilizó *gráficos de caja* para cada una de las variables con la finalidad de analizar la distribución de los datos y la presencia de valores atípicos. Gracias a este análisis y con la ayuda del especialista en medicina se excluyó los valores atípicos (ver punto 4.4.).La distribución final de los datos para cada una de las variables se expresa en las siguientes ilustraciones:



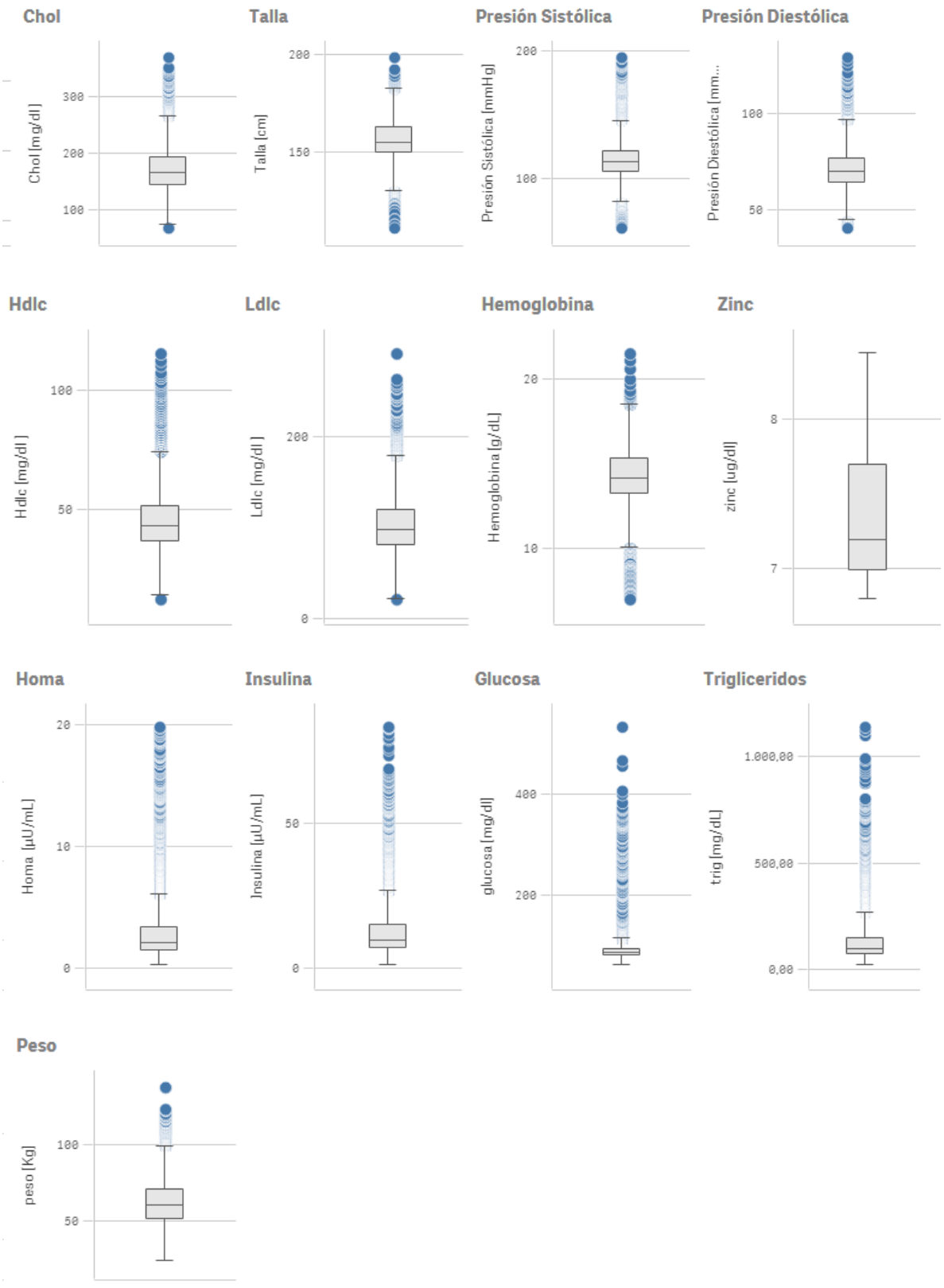


Figura 35. Gráfico de cajas de las variables asociadas a la diabetes y la hipertensión

En el gráfico de cajas de la *edad* se puede ver claramente que no hay valores extremos ya que el conjunto de datos tiene personas de entre 10 a 59 años. La mayor parte de la población analizada se encuentra entre los 20 y 40 años.

Para el índice de masa corporal a pesar de que el *boxplot* muestra valores atípicos, estos, no son considerados errores. Ya que son valores posibles tanto para los valores altos, en el caso de personas con sobre peso y obesidad, como para valores bajos en el caso de personas con peso bajo. Esta variable tiene una distribución simétrica.

El diámetro de la cintura muestra la presencia de un valor extremo que es el que sobresale del resto de valores altos. A pesar de eso esta variable tiene una distribución simétrica también. Es decir, no se destaca por tener valores muy altos o muy bajos.

Para el caso de la ferritina hay una ligera concentración de datos a la izquierda de la distribución y presenta un valor extremo que sobresale del resto. Sin embargo es un valor complemente posible de darse. Tomando en cuenta el *boxplot* del colesterol se puede concluir que tiene una distribución normal. Sin embargo presenta valores altos.

De la misma manera la talla de las personas tiene una distribución normal con ciertos valores extremos, tanto por encima del bigote superior, como por debajo del bigote interior. Valores completamente normales ya hay personas de 10 a 59 años en el conjunto de datos. Lo mismo pasa con la presión sistólica y diastólica y algo parecido acontece con hdlc, ldlc, hemoglobina, zinc y peso.

Para la insulina, homa, glucosa y triglicéridos se puede ver claramente que tiene una distribución asimétrica positiva, es decir los datos se encuentra mayormente concentrados en la parte inferior del boxplot, esto quiere decir que hay pocos casos en los que hay valores altos en estas variables. Estas variables tienen habitualmente ese comportamiento en la población ecuatoriana.

Variables más Importantes para la Diabetes

Se aplicó consecutivamente árboles de decisión usando el algoritmo *CHAID* sobre todo el conjunto de datos para identificar las variables que más influyen en la diabetes; con base en el nivel de ganancia de información que generan utilizando el valor del estadístico *chi cuadrado* como métrica de análisis. Los parámetros que se utilizaron fueron los siguientes:

Nivel de Significacion	0,05
Estadistico Chi Cuadrado	Pearson
Numero maximo de iteraciones	100
Correguir los valores de significacion	Metodo Bonferroni

Tabla 5. Parámetros, árbol de Decisión CHAID

Después de encontrar la primera variable, se la excluyó del estudio y se volvió a generar el árbol de decisión y se encontró la segunda variable más influyente, es decir el primer nodo del árbol. Este proceso se continuó realizando hasta encontrar las diez variables más importantes. La siguiente figura muestra el top de las diez variables más influyentes en base al estadístico *chi cuadrado*. Debido a que los valores de glucosa definen si la persona padece de diabetes se decidió incluirla en el análisis y añadir una variable más al proceso de selección.

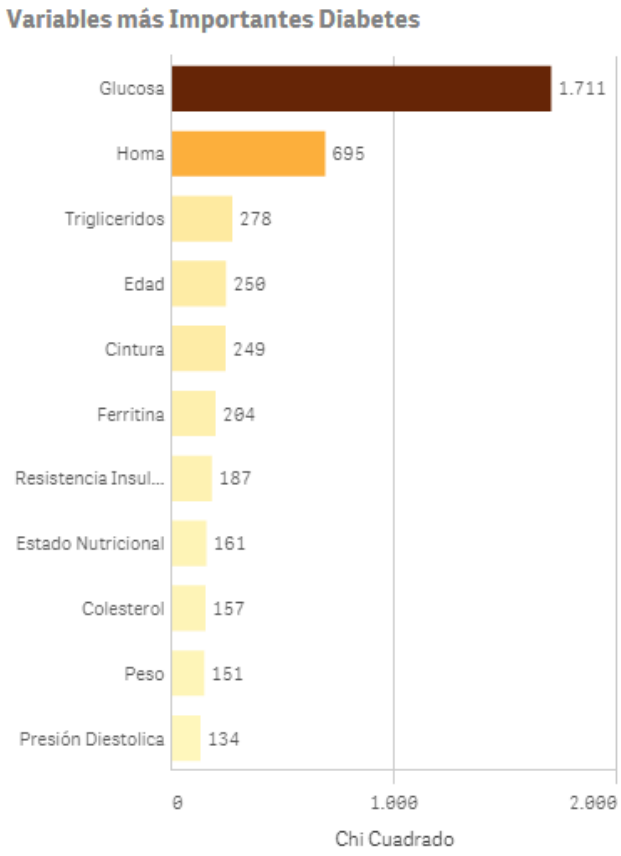


Figura 36. Gráfico de las variables más importantes para la diabetes

El primer árbol muestra a la glucosa como la variable más correlacionada con la diabetes, esto es evidentemente cierto ya que los valores de esta definen si la persona padece o no esta enfermedad, si son superiores a 126 mg/dl la persona tiene diabetes. Este árbol de decisión obtuvo dos nodos, el primero albergó a las personas con glucosa menores a 100 mg/dl, teniendo cero casos de personas con diabetes. Mientras que el segundo nodo albergó a todas las personas con diabetes siendo el 13.7% del total de casos que satisfacen la condición de este nodo.

El segundo árbol de decisión evidenció que el índice homa es la variable más importante. Vale la pena resaltar que este índice sirve para cuantificar la resistencia de una persona a la insulina, hormona que regula los niveles de azúcar en la sangre. Este árbol de decisión

obtuvo cinco nodos en donde se pudo ver claramente que el último nodo es el que agrupa a las personas con un índice homa superior a 5.000 tiene 9% de personas diabéticas de un total de 1368 personas. En contraste con este, ninguno de los nodos restantes supera el 1.5% con personas diabéticas. Es más, el segundo nodo, que es el que alberga a 8206 personas no llega a tener más de 0.4%.

El tercer árbol de decisión revela a los triglicéridos como la segunda variable más importante en la diabetes obteniendo 4 nodos, de los cuales el primero alberga a 5500 casos, los mismos que tienen sus niveles de triglicéridos inferiores a 86 mg/dL y solamente el 0.1% de ellos tienen diabetes. Recordemos que los triglicéridos son un tipo de grasa, que en determinados casos podría llegar a taponar las arterias y eventualmente provocar un infarto. Adicionalmente, el nodo cuatro es el que registra más casos de diabetes llegando al 5.6% de 1377, siendo la característica principal de este grupo el tener los triglicéridos por encima de los 215 mg/dL.

El cuarto árbol de decisión indica que la edad es la tercera variable más influyente, en donde las personas de más de 40 años de edad son más propensos a tener diabetes, el nodo 4 que alberga a las persona de entre 40 a 49 años, es el que registra más casos de personas con glucosa superior a 126 mg/dL. Recordemos que las edades de las personas encuestadas varían de entre 10 a 59 años. Los nodos dos (con rango de 20 a 29 años) y tres (de 30 a 39) también registran algunos casos de personas con diabetes 04% de un total de 3102 y 1.6% de 3404 respectivamente. El nodo 1 que alberga a 4337 no registra casos de diabetes.

El quinto nodo muestra al “diámetro de la cintura” como la siguiente variable más importante en la diabetes. Es una característica de las personas con diabetes el tener un diámetro de la cintura considerablemente grande puesto que también poseen problemas asociados al sobrepeso. El quinto nodo que alberga a las personas con diámetro de la cintura superior a los 100 centímetros, en total 1362 personas tiene 5.4% con diabetes. En contraste con el primer nodo que registra solamente el 0.2% de personas con diabetes.

En el quinto lugar se encuentra la ferritina. Recordemos que la ferritina es la proteína que transporta el hierro y la ausencia de la misma podría indicar deficiencia de hierro en la persona. Si bien no hay una conexión directa entre esta variable y la diabetes es interesante encontrar a esta variable en este listado pues podría ser la variable que conecta información entre los macro y micronutrientes y la diabetes. El árbol de decisión indica que el 5.3% de las personas con ferritina superior a 168 [µg/L] tienen diabetes. Mientras que solamente el 0.2% de las personas que tienen entre ferritina 34 a 43 [µg/L] padecen esta enfermedad.

El séptimo árbol de decisión evidencia que la Resistencia a la Insulina es la sexta variable más influyente en la diabetes. Recordemos que la diabetes aparece cuando el páncreas no produce insulina suficiente o cuando el organismo no utiliza eficazmente la insulina que produce. Por lo tanto es lógico pensar que la mayor parte de los casos de personas con

diabetes sean resistentes a la insulina. Esto se puede ver en el nodo 1 del árbol el cual alberga a las personas que presentan resistencia a la insulina y que a su vez constituyen el 3.0% de un total de 8366.

Como séptima variable se encuentra el estado nutricional. Esta es una categorización que se calcula en función del índice de masa corporal de la persona (ver punto 4.4.) para personas mayores de 19 años un IMC menor a 18.5 representa un "Peso Bajo". Para valores entre 18.5 y 24.9 es "Peso Adecuado" de 25 a 29.9 es "Sobrepeso" y para valores mayores a 29.9 es "Obesidad". En el octavo árbol de decisión se puede apreciar que la mayor parte de las personas con diabetes tienen problemas de obesidad y sobrepeso (3.8% de 2158 y 1.7% de 3967 respectivamente). Sin embargo se encuentran algunos casos de diabetes en personas con peso adecuado y peso bajo.

El colesterol figura como la novena octava más influyente, con un valor de *chi cuadrado* de 157. La OMS explica que:

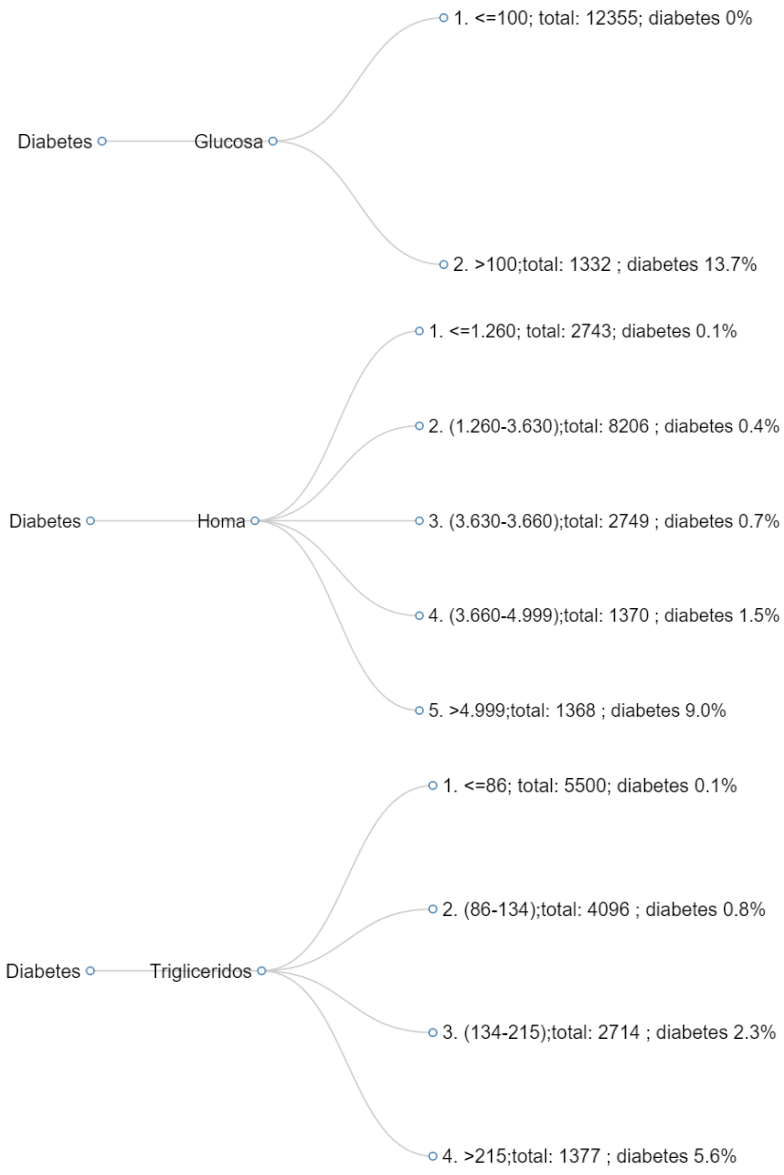
> "El colesterol es una sustancia serosa y parecida a la grasa que se encuentra en todas las células de su cuerpo. Su cuerpo necesita algo de colesterol para producir hormonas, vitamina D y sustancias que le ayuden a digerir los alimentos. Su cuerpo produce todo el colesterol que necesita. El colesterol también se encuentra en alimentos de origen animal, como yemas de huevo, carne y queso.
>
> Si tiene demasiado colesterol en la sangre, puede combinarse con otras sustancias en la sangre para formar placa. La placa se pega a las paredes de sus vasos sanguíneos. Esta acumulación se llama arterioesclerosis. Puede provocar enfermedad de las arterias coronarias, la que puede estrecharlas o incluso bloquearlas. "

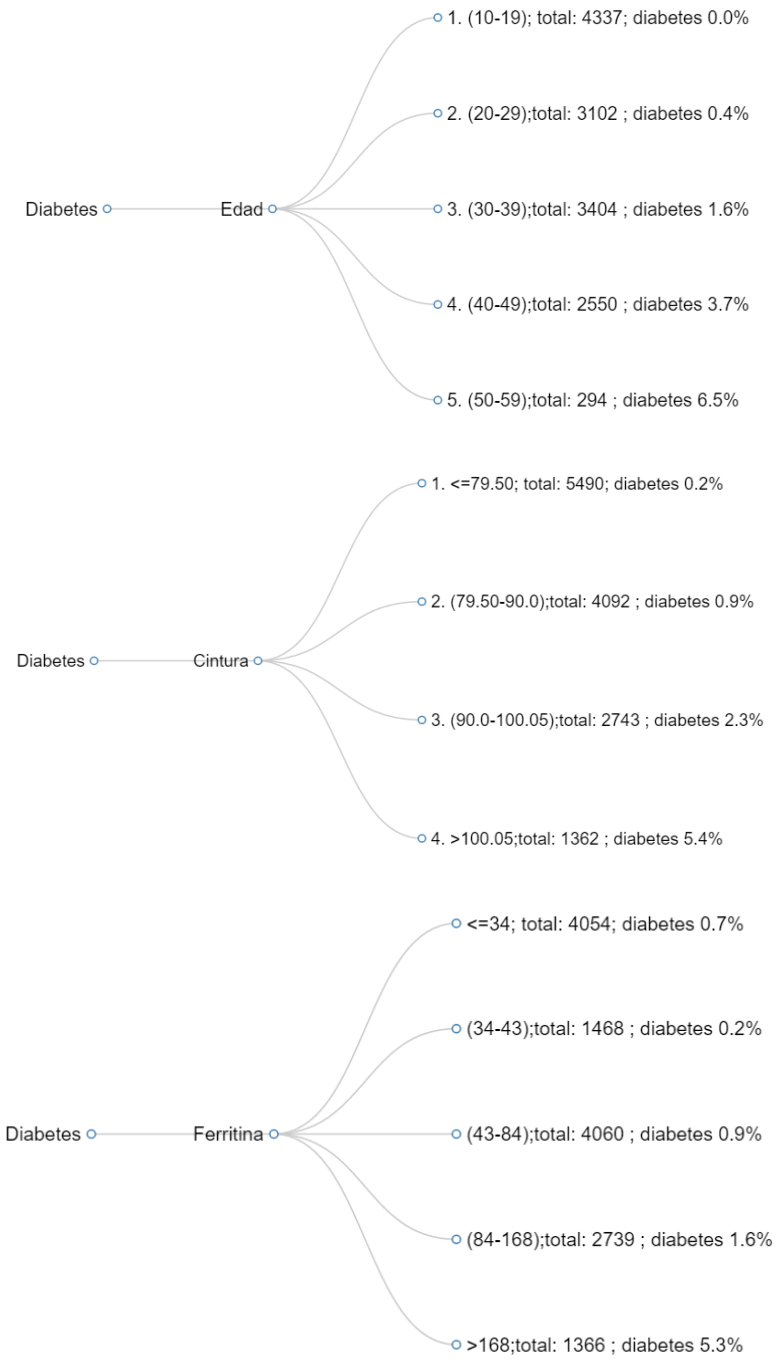
El siguiente árbol de decisión muestra que el 5.6% de los casos de las personas con colesterol superior a 218 mg/dl tienen diabetes (recordemos que los valores normales de colesterol deben ser menores a 200 mg/dL). A su vez, el 2.1% de las personas con colesterol entre 199 y 218 mg/dL tienen diabetes. También hay varios casos de diabéticos cuando los niveles de colesterol se encuentran entre 157 y 199 mg/dL.

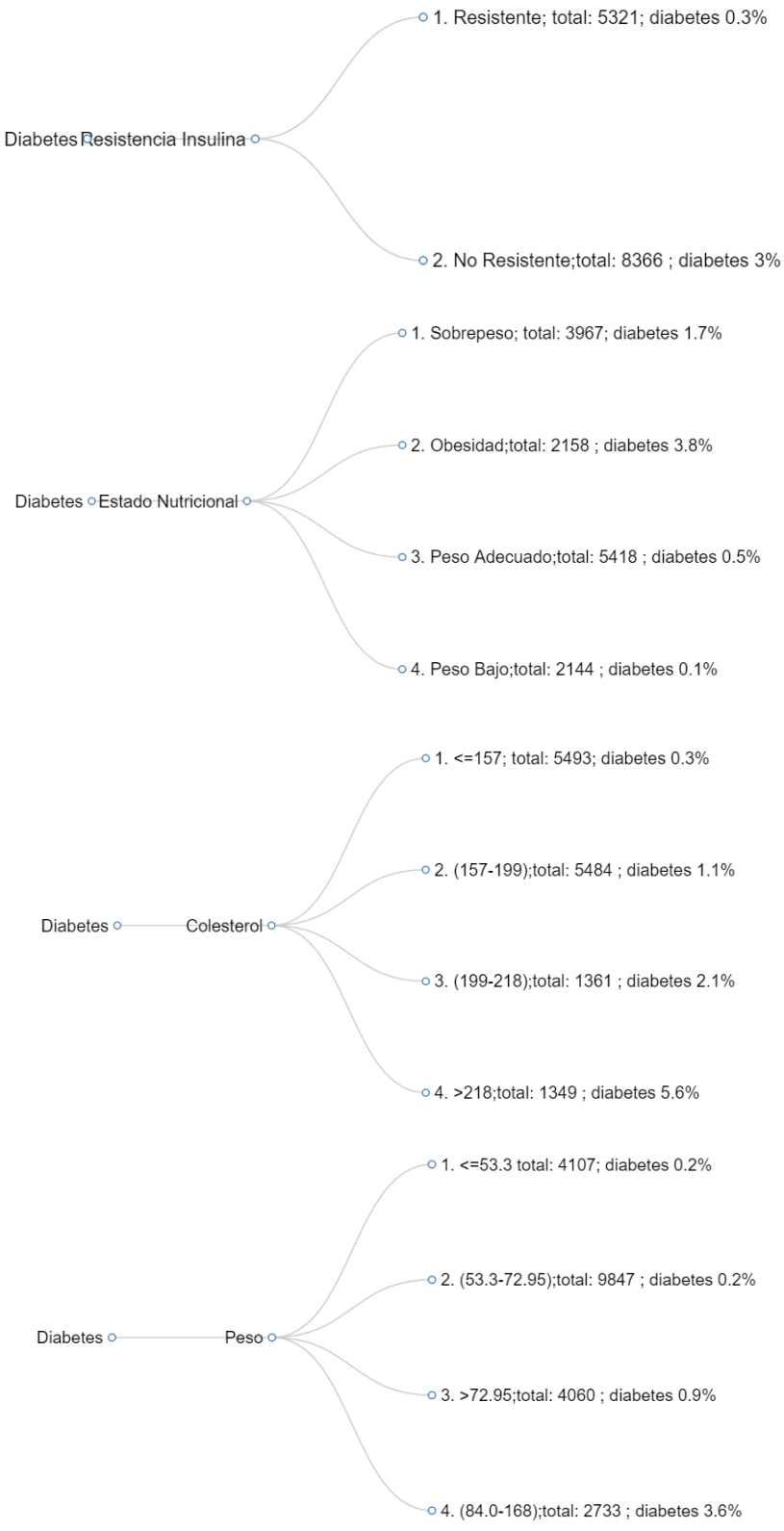
En noveno lugar se encuentra el peso. No hace falta explicar que las personas diabéticas tienden a poseer problemas de sobrepeso es por eso que podemos evidenciar que en los nodos 4 y 5, los cuales albergan a personas con peso superior a los 72.95 kg, se encuentra la mayor cantidad de personas diabéticas, alrededor de 134 personas, especialmente en el nodo 5 el cual tiene a las personas con peso superior a 84 kg. Sin embargo se registran algunos casos de personas diabéticas con peso inferior a los 72.95 kg.

Finalmente, en décimo lugar se encuentra la presión diastólica. Este árbol de decisión define cuatro nodos, el primero agrupa a las personas con valores de presión diastólica menores a 64.67 mmHg el segundo a personas con valores entre 64.67 y 69.67. En ambos casos el porcentaje de personas con diabetes no es muy alto (0.2 % y 0.8 %); mientras que en los dos nodos subsiguientes se encuentran más personas que padecen de diabetes,

destacándose el nodo cuatro con un rango de presión diastólica de 84 mmHg a 168 mmHg albergando a más de 98 personas con diabetes. Evidentemente la mayor parte de las personas con diabetes tienen presión diastólica alta.




Diabetes
Edad
1. (10-19); total: 4337; diabetes 0.0%
2. (20-29);total: 3102 ; diabetes 0.4%
3. (30-39);total: 3404 ; diabetes 1.6%
4. (40-49);total: 2550 ; diabetes 3.7%
5. (50-59);total: 294 ; diabetes 6.5%
Diabetes
Cintura
1. <=79.50; total: 5490; diabetes 0.2%
2. (79.50-90.0);total: 4092 ; diabetes 0.9%
3. (90.0-100.05);total: 2743 ; diabetes 2.3%
4. >100.05;total: 1362 ; diabetes 5.4%
Diabetes
Ferritina
<=34; total: 4054; diabetes 0.7%
(34-43);total: 1468 ; diabetes 0.2%
(43-84);total: 4060 ; diabetes 0.9%
(84-168);total: 2739 ; diabetes 1.6%
>168;total: 1366 ; diabetes 5.3%


Diabetes
Resistencia Insulina
1. Resistente; total: 5321; diabetes 0.3%
2. No Resistente;total: 8366 ; diabetes 3%
Diabetes
Estado Nutricional
1. Sobrepeso; total: 3967; diabetes 1.7%
2. Obesidad;total: 2158 ; diabetes 3.8%
3. Peso Adecuado;total: 5418 ; diabetes 0.5%
4. Peso Bajo;total: 2144 ; diabetes 0.1%
Diabetes
Colesterol
1. <=157; total: 5493; diabetes 0.3%
2. (157-199);total: 5484 ; diabetes 1.1%
3. (199-218);total: 1361 ; diabetes 2.1%
4. >218;total: 1349 ; diabetes 5.6%
Diabetes
Peso
1. <=53.3 total: 4107; diabetes 0.2%
2. (53.3-72.95);total: 9847 ; diabetes 0.2%
3. >72.95;total: 4060 ; diabetes 0.9%
4. (84.0-168);total: 2733 ; diabetes 3.6%

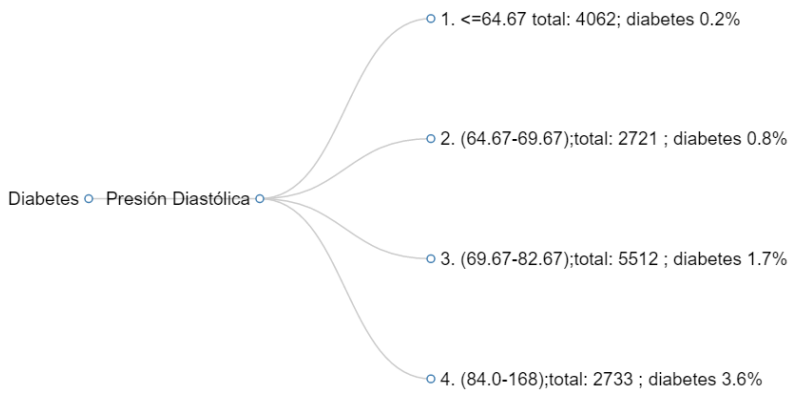


Figura 37. Arboles de decisión de las variables más importantes de la diabetes

Variables más Importantes para la Hipertensión

Al igual que el caso anterior, se aplicó consecutivamente árboles de decisión usando el algoritmo *CHAID* sobre todo el conjunto de datos para identificar las variables que más influyen en la hipertensión; con base en el nivel de ganancia de información que generan utilizando el valor del estadístico *chi cuadrado* como métrica de análisis. Los parámetros usados fueron los mismos que el caso anterior (tabla 5).

Después de encontrar la primera variable, se la excluyó del estudio y se volvió a generar el árbol de decisión y se encontró la segunda variable más influyente, es decir el primer nodo del árbol. Este proceso se continuó realizando hasta encontrar las diez variables más importantes. La siguiente figura muestra el top de las diez variables más influyentes en base al estadístico *chi cuadrado*. Debido a que los valores de presión sistólica y diastólica definen si la persona padece de hipertensión se decidió incluirlas en el análisis y añadir dos variables más al proceso de selección.

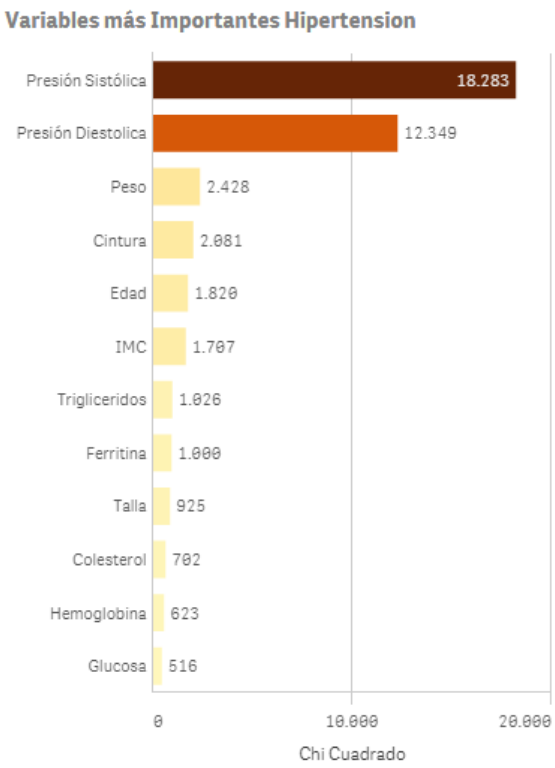


Figura 38. Gráfico de las variables más importantes para la hipertensión

Los dos primeros árboles de decisión indican que la presión sistólica y diastólica son las variables más correlacionadas con la hipertensión, lo cual es evidentemente correcto debido a que la hipertensión se define a partir de valores de presión sistólica superiores a 140 mmHg o valores superiores a 90 mmHg para el caso de la presión diastólica.

En el nodo 9 del primer árbol, mismo que agrupa a las personas con presión sistólica superior a 129.67 mmHg se puede observar que la gran mayoría de los casos de hipertensión se encuentran allí 45.5% de un total de 1386 personas. En el resto de nodos se registran casos de hipertensión pero con una proporción mucho menor.

De manera similar se puede observar que en el segundo árbol el último nodo alberga a la gran mayoría de personas hipertensas teniendo, 44.9% de un total de 1407 casos con valores superiores a 82.67 mmHg. En el resto de nodos se registran casos de hipertensión pero con una proporción mucho menor.

En contraste con el listado de variables influyentes de la diabetes en donde el peso ocupó el décimo lugar, el tercer árbol de decisión muestra a esta variable como la variable más importante para la hipertensión. En donde la cantidad de casos que agrupa cada uno de los nodos es similar, salvo el caso del nodo 9 donde se registran muchos más casos. La proporción de personas con hipertensión empieza a crecer a partir del nodo cinco, mismo que agrupa a las personas con peso entre 57.05 y 60.45 kg. Los nodos 9 y 10 registran la mayor proporción de casos de hipertensión: 8.9% de 2730 para el caso del nodo 9, mismo

que agrupa a personas de entre 68.150 y 80.40 Kg de peso y 206 casos para el nodo 10 que agrupa a las personas con peso mayor a 80.4 Kg.

Como la segunda variable más importante aparece el diámetro de la cintura. En este árbol de decisión encontré la proporción de personas en cada una de las ramas es muy similar, la misma que fluctúa entre 1368 y 1383 a excepción de la rama 3 la cual alberga a 2747 personas. Como era de esperarse mientras más altos son los rangos del diámetro de cintura más casos de hipertensión existen, teniendo, para el nodo 8 el 15.6% de personas con hipertensión de un total de 1362. Para el caso de la diabetes esta variable ocupó el quinto lugar.

En el quinto árbol de decisión denota que la variable edad es la tercera variable en importancia de la hipertensión (para el caso de la diabetes ocupa también el tercer puesto). Los últimos cuatro nodos albergan a la mayor cantidad de personas hipertensas llegando al 17% para el nodo 8, y 10.2% para el nodo 7. Para el nodo 1, que alberga a las personas menores de 13 años también se registra hipertensión pero en mucha menor proporción llegando al 0.3%. Esta es la variable cuya distribución es la más cercana a una distribución normal, ya que los datos con los cuales se trabajó pertenecen a personas de entre 10 a 59 años. Para el caso de la diabetes esta variable ocupó el cuarto lugar.

El IMC ocupa el cuarto lugar en importancia para la hipertensión subiendo 3 escalones con respecto a la diabetes. Los nodos con rangos de IMC más altos son los que más casos de hipertensión albergan teniendo 13.5% para >31.76, en caso del nodo 8; 9% para el nodo 7, 8.7 % para el nodo 6, 5.4% para el nodo 5. Cabe resaltar que estos nodos agrupan a las personas con problemas de sobre peso y obesidad. Sin embargo también se registran casos de hipertensión en personas con peso bajo aunque en una proporción mucho menor como es el caso del primer nodo.

El séptimo árbol de decisión indica que los triglicéridos son la quinta variable en importancia para la hipertensión y, tal y como en los casos anteriores a valores más altos de esta variable mayor es la proporción de personas hipertensas. El nodo 7 tiene la mayor proporción de personas hipertensas 13.4% y cuyos valores son superiores a los 216 [mg/dL]. En contraste con el nodo 1 que tiene 1.6% de hipertensos y con niveles de triglicéridos menores a 64 [mg/dL]. Es importante señalar que a partir del nodo 5 empiezan a aparecer casos de hipertrigliceridemia y es donde la proporción de hipertensos empieza a aumentar más. Esta variable figura como más importante en la diabetes que en el caso de la hipertensión.

En la sexta posición de las variables de influencia se encuentra la ferritina, esta variable figura como más importante en la diabetes que en la hipertensión según el ranking que se realizó. Al igual que para el caso de la diabetes resulta muy interesante encontrar a esta variable en este listado pues podría ser la variable que conecta información entre los macro

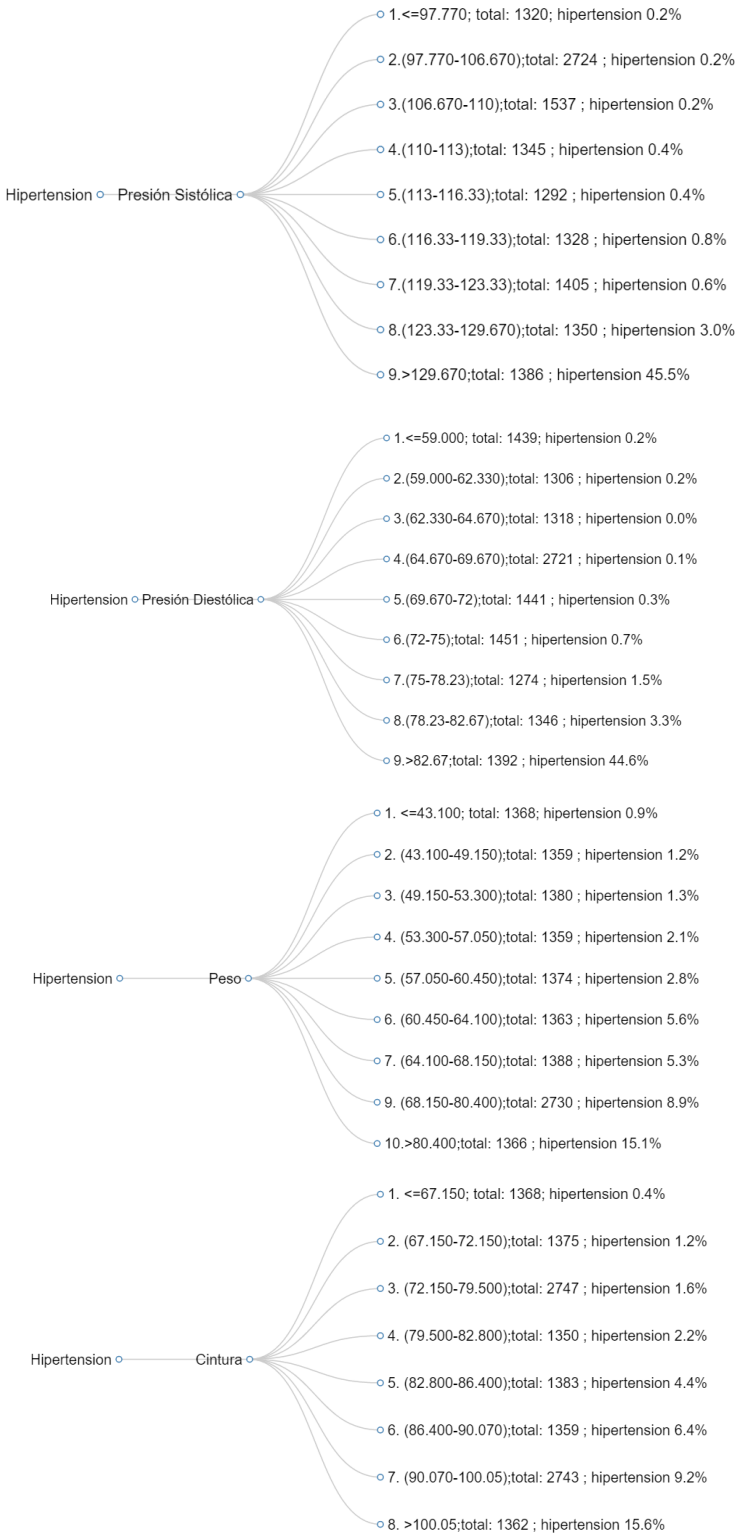
y micronutrientes con la hipertensión. De igual forma que en el resto de variables analizadas a mayores niveles de ferritina mayor es la proporción de hipertensión. También se registran casos de hipertensión en valores bajos, tal y como se puede ver en el primer nodo.

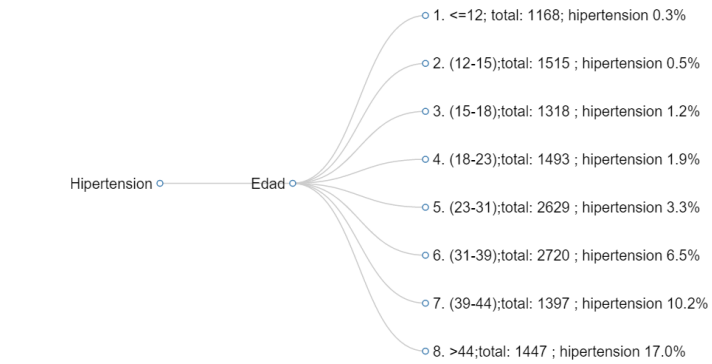
La séptima variable de influencia en este listado es la variable talla, esta es la única variable qué, en compañía de la presión sistólica y la hemoglobina no aparecen en el listado de las variables más influentes en la diabetes. Se va incrementando la proporción de personas hipertensas a medida que los rangos de los nodos son mayores (e.g. nodo 1 <=143.850 y 1.8%; nodo 7 > 168.45 y 8.8%) la cantidad de personas en cada uno de los nodos no se encuentra equitativamente distribuida así que no se puede afirmar que a valores más altos de talla mayores casos de hipertensión. Si puedo afirmar que para las personas del nodo 1 se registran pocos casos de hipertensión, posiblemente porque en su mayoría son niños.

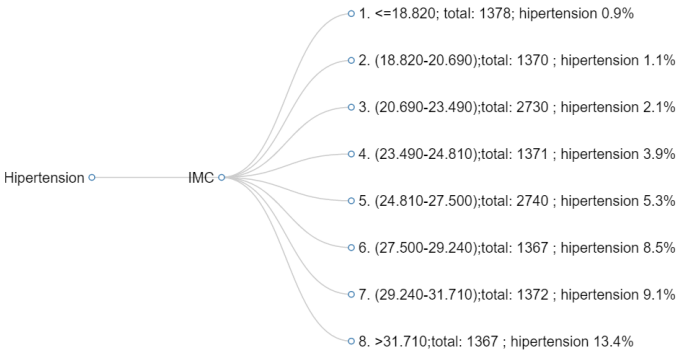
En el octavo lugar se encuentra el colesterol. Para los tres primeros nodos se registran pocos casos de hipertensión. No así para los cuatro nodos subsiguientes en donde el nodo 5 que es el que agrupa a las personas que tienen entre 166 y 199 [mg/dL] es en el que mayor cantidad de casos de hipertensión se registra, en este nodo no hay casos de personas con hipercolesterolemia, pues deberían tener valores superiores a los 200 [mg/dL]. Para los nodos en los cuales hay personas con hipercolesterolemia, también se registran algunos casos de hipertensión pero menos que en el nodo 5.

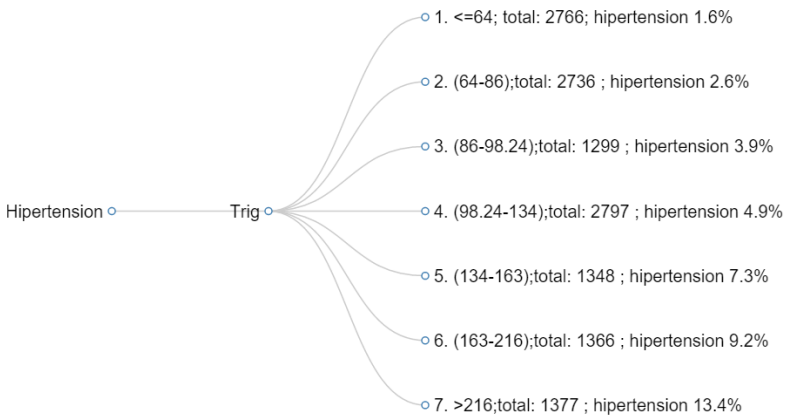
En noveno lugar tenemos a la hemoglobina. El primer nodo alberga a las personas con valores inferiores a los 13.9 y de un total de 5510 el 6.3 % tiene hipertensión, por otro lado los nodos 4 y 5 sugieren que hay una proporción considerable de personas con hipertensión qué, a su vez, tienen también valores de hemoglobina superiores a los 15, casi la mitad de las personas con hipertensión. Esto indica que hay una correlación entre valores altos y bajos de hemoglobina respecto a la hipertensión arterial.

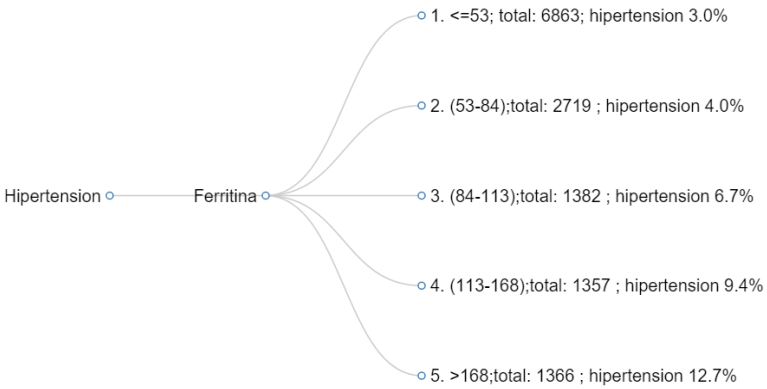
El último árbol de decisión revela algo sumamente interesante, la glucosa figura como la décima variable de influencia para la hipertensión. La variable cuyos valores determinan si una persona padece de diabetes, también se encuentra correlacionada con la hipertensión arterial, conforme va incrementando los valores de los nodos va aumentando la cantidad de personas con hipertensión arterial e.g. el último nodo el cual alberga a las personas con glucosa superiores a 100 de un total de 1332 el 12.3% tienen hipertensión.


Hipertension
Presión Sistólica
1.<=97.770; total: 1320; hipertension 0.2%
2.(97.770-106.670);total: 2724 ; hipertension 0.2%
3.(106.670-110);total: 1537 ; hipertension 0.2%
4.(110-113);total: 1345 ; hipertension 0.4%
5.(113-116.33);total: 1292 ; hipertension 0.4%
6.(116.33-119.33);total: 1328 ; hipertension 0.8%
7.(119.33-123.33);total: 1405 ; hipertension 0.6%
8.(123.33-129.670);total: 1350 ; hipertension 3.0%
9.>129.670;total: 1386 ; hipertension 45.5%
Hipertension
Presión Diestólica
1.<=59.000; total: 1439; hipertension 0.2%
2.(59.000-62.330);total: 1306 ; hipertension 0.2%
3.(62.330-64.670);total: 1318 ; hipertension 0.0%
4.(64.670-69.670);total: 2721 ; hipertension 0.1%
5.(69.670-72);total: 1441 ; hipertension 0.3%
6.(72-75);total: 1451 ; hipertension 0.7%
7.(75-78.23);total: 1274 ; hipertension 1.5%
8.(78.23-82.67);total: 1346 ; hipertension 3.3%
9.>82.67;total: 1392 ; hipertension 44.6%
Hipertension
Peso
1. <=43.100; total: 1368; hipertension 0.9%
2. (43.100-49.150);total: 1359 ; hipertension 1.2%
3. (49.150-53.300);total: 1380 ; hipertension 1.3%
4. (53.300-57.050);total: 1359 ; hipertension 2.1%
5. (57.050-60.450);total: 1374 ; hipertension 2.8%
6. (60.450-64.100);total: 1363 ; hipertension 5.6%
7. (64.100-68.150);total: 1388 ; hipertension 5.3%
9. (68.150-80.400);total: 2730 ; hipertension 8.9%
10.>80.400;total: 1366 ; hipertension 15.1%
Hipertension
Cintura
1. <=67.150; total: 1368; hipertension 0.4%
2. (67.150-72.150);total: 1375 ; hipertension 1.2%
3. (72.150-79.500);total: 2747 ; hipertension 1.6%
4. (79.500-82.800);total: 1350 ; hipertension 2.2%
5. (82.800-86.400);total: 1383 ; hipertension 4.4%
6. (86.400-90.070);total: 1359 ; hipertension 6.4%
7. (90.070-100.05);total: 2743 ; hipertension 9.2%
8. >100.05;total: 1362 ; hipertension 15.6%


Hipertension
Edad
1. <=12; total: 1168; hipertension 0.3%
2. (12-15);total: 1515 ; hipertension 0.5%
3. (15-18);total: 1318 ; hipertension 1.2%
4. (18-23);total: 1493 ; hipertension 1.9%
5. (23-31);total: 2629 ; hipertension 3.3%
6. (31-39);total: 2720 ; hipertension 6.5%
7. (39-44);total: 1397 ; hipertension 10.2%
8. >44;total: 1447 ; hipertension 17.0%


Hipertension
IMC
1. <=18.820; total: 1378; hipertension 0.9%
2. (18.820-20.690);total: 1370 ; hipertension 1.1%
3. (20.690-23.490);total: 2730 ; hipertension 2.1%
4. (23.490-24.810);total: 1371 ; hipertension 3.9%
5. (24.810-27.500);total: 2740 ; hipertension 5.3%
6. (27.500-29.240);total: 1367 ; hipertension 8.5%
7. (29.240-31.710);total: 1372 ; hipertension 9.1%
8. >31.710;total: 1367 ; hipertension 13.4%


Hipertension
Trig
1. <=64; total: 2766; hipertension 1.6%
2. (64-86);total: 2736 ; hipertension 2.6%
3. (86-98.24);total: 1299 ; hipertension 3.9%
4. (98.24-134);total: 2797 ; hipertension 4.9%
5. (134-163);total: 1348 ; hipertension 7.3%
6. (163-216);total: 1366 ; hipertension 9.2%
7. >216;total: 1377 ; hipertension 13.4%


Hipertension
Ferritina
1. <=53; total: 6863; hipertension 3.0%
2. (53-84);total: 2719 ; hipertension 4.0%
3. (84-113);total: 1382 ; hipertension 6.7%
4. (113-168);total: 1357 ; hipertension 9.4%
5. >168;total: 1366 ; hipertension 12.7%

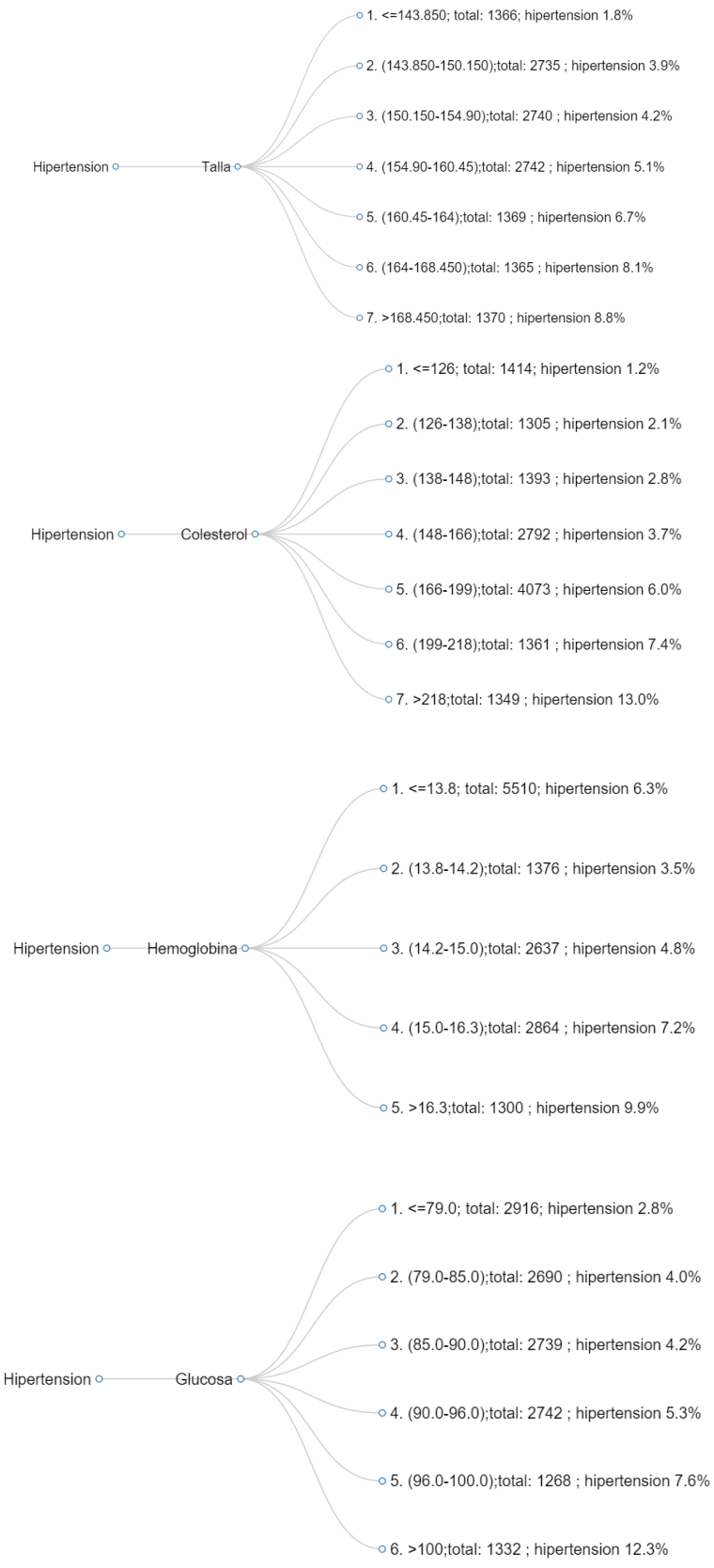


Figura 39. Arboles de decisión de las variables más importantes de la hipertensión

5.3. Análisis Poblacional

5.3.1. Aspectos Antropométricos

Análisis Univariado – Aspectos Geográficos

Gran parte de la población ecuatoriana, es decir, el 40% tiene un peso adecuado. Las mujeres presentan más problemas de sobrepeso y obesidad que los hombres, 52% contra 36% respectivamente. Las provincias que sobresalen por tener un IMC elevado son Galápagos, Esmeraldas, Sucumbíos, Carchi y Cotopaxi. Mientras que las provincias que en promedio presentan un estado nutricional adecuado son Manabí, Tungurahua, Bolívar, Pichincha, Orellana y Napo, Los Ríos. En este grupo solamente dos provincias de la costa.

En cuanto al peso promedio, las cinco provincias que se destacan por tener los valores más altos son Esmeraldas, Galápagos, Guayas, Sto. Domingo de los Tsáchilas y Sucumbíos. Mientras que las provincias que sobresalen por la altura de sus habitantes son Guayas, Esmeraldas, Pichincha, Manabí y Sto. Domingo. (Figura 40).

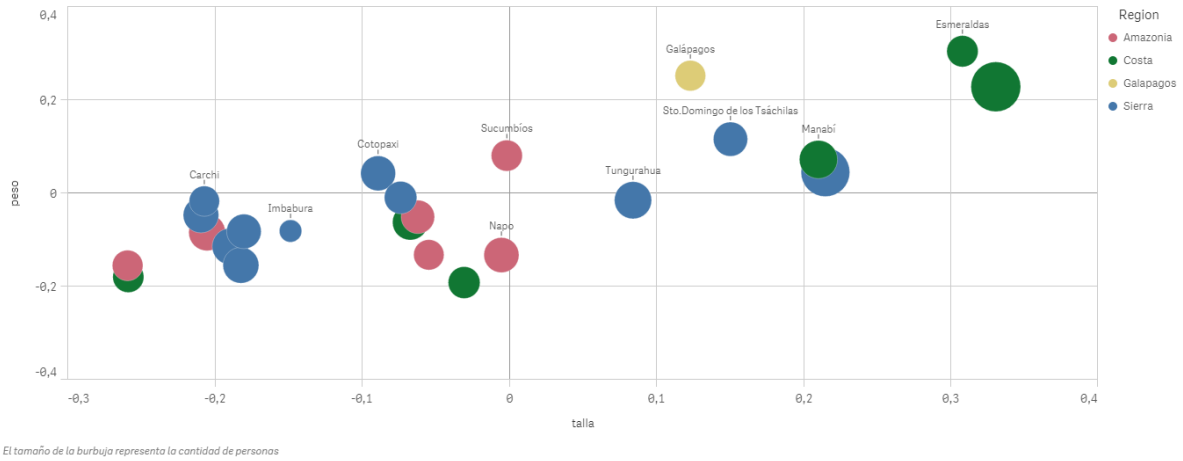


Figura 40. Talla, peso y población por provincia y región

Análisis Multivariado – Económicos y Demográficos

Aplicando un análisis de correspondencias se encontró que las personas de entre 30 a 49 años tienden al sobrepeso y a la obesidad, a tener niveles no normales de cintura, a tener hipertensión y a tomar medicación para la presión. En cambio, las personas de 10 a 19 años se caracterizan por su tendencia a ser hipotensos y a no tener problemas de sobrepeso.

Adicionalmente, se encontró que los indígenas, en su gran mayoría se encuentran en el quintil económico 1. Esto no es nada nuevo ya que históricamente, éste es el grupo étnico que más exclusión social ha experimentado en el país. Algo similar pasa con los montubios y afro ecuatorianos ya que en su mayoría ocupan los quintiles 1 y 2, los cuales son lo los quintiles económicos más bajos. El resto de la población se encuentra distribuida entre los quintiles económicos de forma más homogénea.

Los quintiles 4 y 5, los cuales representan a los grupos económicos de mayor riqueza son muy similares entre sí. La población de entre 50 a 59 años es el grupo etario minoritario.

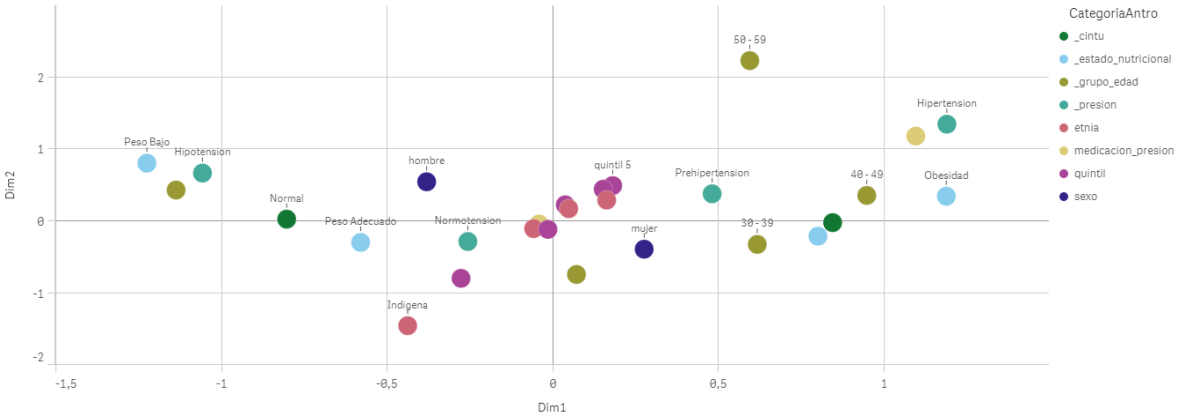


Figura 41. Análisis de Correspondencia Aspectos Antropométricos, Económicos y Demográficos

Análisis Multivariado – Componentes Principales y Conglomerados

Con la finalidad de estudiar las características de las 24 provincias del país en relación a los aspectos antropométricos, se decidió realizar un análisis de correlaciones. Los resultados de este estudio indican que existe una alta correlación entre el *IMC*, el diámetro de la cintura y el peso de la persona. También se encuentran correlacionados pero en menor proporción la presión sistólica y diastólica, y con menor magnitud el peso y la talla de la persona. (Figura 42).

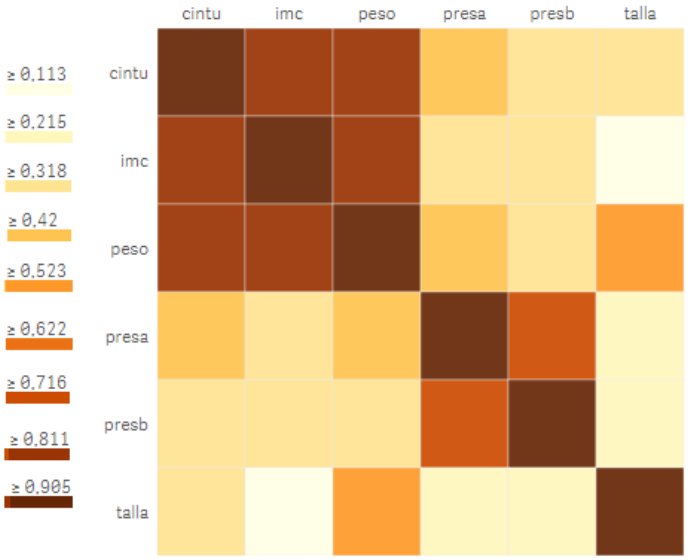


Figura 42. Análisis de Correlaciones Aspectos Antropométricos

Estos resultados me dieron el aval necesario para poder aplicar *componentes principales* como técnica de reducción de dimensiones, puesto que al existir correlaciones altas entre las variables tiene sentido aplicar este tipo de técnicas. Las variables que se usaron para este

análisis fueron el *IMC*, el diámetro de la cintura, la presión sistólica, y la presión diastólica. Se obtuvo el 94% de la variabilidad del conjunto de datos con las primeras dos componentes principales.

La primera componente es la que acumula la mayor variabilidad, cerca del 64% y tiene todos sus coeficientes positivos, se interpreta como una componente de tamaño. Es decir, que mientras más alto sea el valor de esta componente, mayores serán los valores de *IMC*, presión sistólica, presión diastólica, y el diámetro de la cintura.

Por otro lado, la segunda componente acumula el 30% de la variabilidad del conjunto de datos, tiene coeficientes positivos y negativos, quiere decir que es una componente de forma o de contraste, Por lo tanto, mientras más alta sea su componente mayores serán los contrastes entre las variables con coeficientes positivos y negativos. Es decir, si una provincia tiene un valor de CP2 alto, va a tener valores de presión altos en relación a sus valores de cintura e IMC [33].

Variable	CP1	CP2
cintu	0,55	0,36
imc	0,39	0,68
presa	0,53	-0,42
presb	0,51	-0,47

Tabla 6. Componentes Principales. Aspectos Antropométricos

Posteriormente, se aplicó diferentes métodos de conglomerados tales como *k - means*, *fuzzy* y *PAM* y se utilizó el coeficiente silhouette para identificar la tendencia al cluster de los datos, obteniendo los mejores resultados con *k – medias* con valores de *k* de 2 y 3.

K igual a 2

Para *k* igual a 2, con valor de silhouette de 0.44, se llegó a los siguientes resultados:

La componente principal 1 definió los conglomerados de la siguiente manera (Gráfico 6 y 7). Las provincias que tienen los valores bajos de esta componente formaron parte de un mismo grupo, mientras que el resto de las provincias formaron parte del otro. El grupo 1 se caracterizó, principalmente porque las provincias que lo compone tienen valores bajos en todos los indicadores antropométricos en relación al resto de provincias.

En la figura 45 se puede ver claramente la distribución de las variables analizadas para los dos grupos generados, tal y como se explicó en párrafos anteriores el conglomerado 2 tiene los aspectos antropométricos más altos mientras que en el conglomerado 1 se encuentran provincias mediciones más bajas en estas variables. No se encontró valores atípicos.

Figura 43. Mapa Ecuador Aspectos Antropométricos. K=2

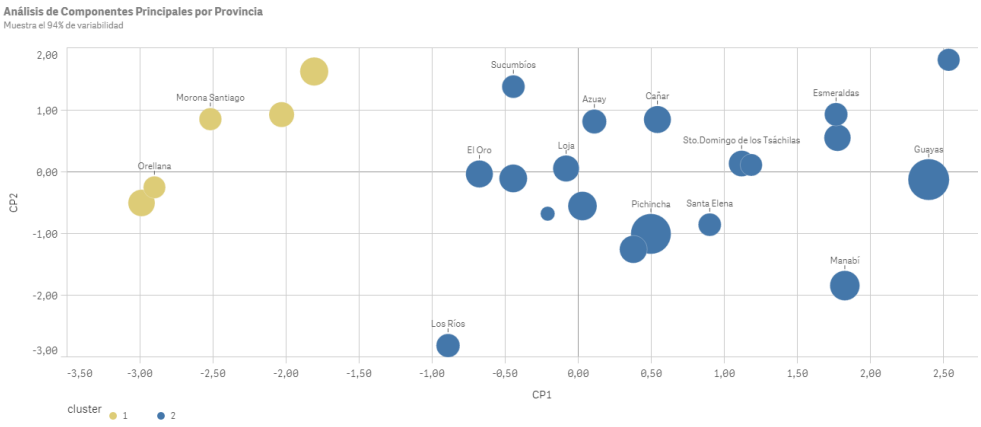


Figura 44. Componentes Principales. Aspectos Antropométricos. K=2

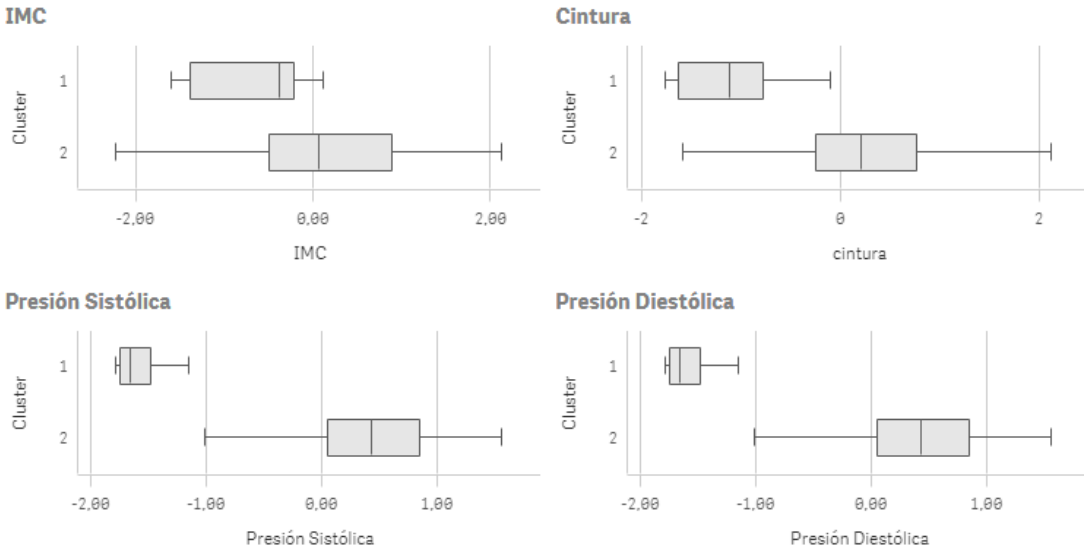


Figura 45. Gráficos de Caja de las provincias de Ecuador. Aspectos Antropométricos. K=2

El conglomerado 1 es el grupo menos numeroso y lo componen todas las provincias de la Amazonía, exceptuando Sucumbíos, es decir, Napo, Orellana, Morona Santiago, Pastaza y Zamora Chinchipe (Figura 43). Lo cual llevó a generar la hipótesis de que existen factores ambientales, culturales, gastronómicos e históricos que hacen que esta provincia sea distinta al resto de provincias de la región. (Anexo 1).

La provincia de Sucumbíos se constituyó como tal a finales de los años 90, es la cuarta provincia más poblada de la región amazónica, y a su vez, desde la extracción del primer barril de petróleo ha sido y sigue siendo el mayor centro productor de este hidrocarburo en el país. El 15 de febrero de 1967, se inicia la explotación del pozo "Lago Agrio 1", el 29 de marzo de 1967 brotaron 2.610 barriles diarios de petróleo [48]. Estos factores, sumados a la migración generada por la extracción de petróleo parecen ser determinantes al momento de separar a esta provincia del resto de las provincias de la amazonia en relación a las variables antropométricas analizadas.

La provincia de Napo, a pesar de tener el valor más bajo de la componente principal 1 (entendiéndose por esto como la provincia con los índices antropométricos más bajos), limita geográficamente con provincias como Sucumbíos, Pichincha y Cotopaxi cuyos índices antropométricos son muy superiores. Esto denota una clara diferenciación regional entre las provincias de la Amazonía y sus homologas de la sierra. Orellana tiene características muy similares a las de Napo. (Figuras 46 y 47).

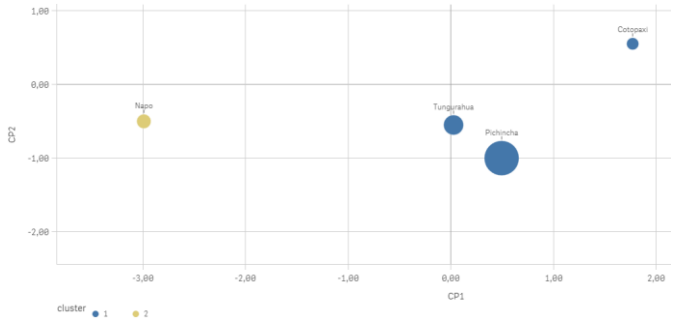


Figura 46. Componentes Principales. Aspectos Antropométricos. K=2. Napo

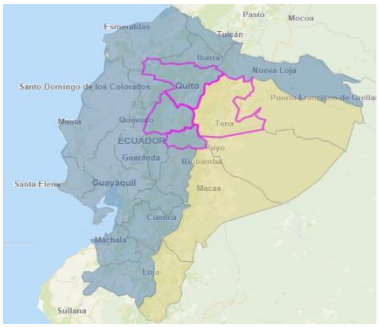


Figura 47. Mapa Ecuador. Napo

En el grupo 1 se encuentran las 19 provincias restantes, en donde se encontraron algunas similitudes entre pares de provincias como Carchi y Santo Domingo, Pichincha y Bolívar, y El Oro y Chimborazo. Los Ríos es la provincia cuyo componente principal 1 es el más bajo de todo el grupo (entendiéndose por esto como la provincia con los índices antropométricos más bajos) y, a su vez, es la única provincia de la Costa que no tiene salida al mar.

Por otro lado, Galápagos es la provincia con los valores más altos en ambas componentes. Quiere decir que, además de tener valores altos de IMC, cintura, presión sistólica y diastólica. Los valores de IMC y cintura son considerablemente altos para los valores de presión. Esta provincia está ubicada a aproximadamente 1000 kilómetros de distancia de la costa continental del Ecuador [49].

Las Islas Galápagos fueron declaradas Patrimonio de la Humanidad por la UNESCO en 1978. Este archipiélago tiene como mayor fuente de ingresos el turismo y recibe alrededor de 200,000 turistas al año. En la figura 44 se puede apreciar que tiene características muy similares a Esmeraldas y Guayas, las cuales son provincias de la costa con alta densidad poblacional (analizando los valores de las componentes principales).

Finalmente, cabe resaltar que en cuanto a la proporción de personas con diabetes e hipertensión se refiere se registran muchos más casos en el grupo 1 que en el grupo 2. Obteniendo un 5.8% contra 2.8% para la hipertensión y 1.7% contra 1.1% para la diabetes. Esto es lógico ya que el grupo 1 se caracteriza por tener los índices antropométricos altos y son precisamente estas variables son las más relevantes para las ENCT; cintura, IMC (o estado nutricional) y peso para la diabetes y presión, peso, IMC y talla para el caso de la hipertensión.

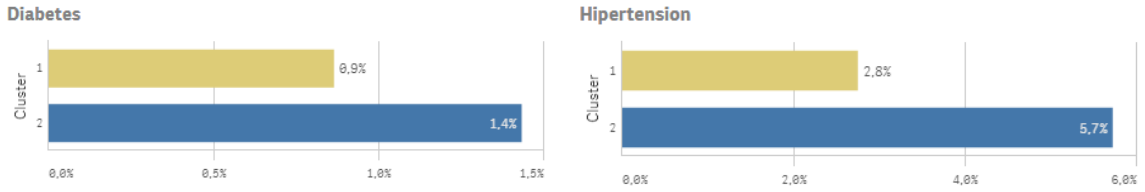


Figura 48. Proporción de Diabetes e Hipertensión. Aspectos Antropométricos. K=2

K igual a 3

Para *k* igual a 3, con valor de silhouette de 0.36, se obtuvieron a los siguientes resultados:

De manera similar que en el caso anteriormente explicado la componente principal 1 es la que finalmente determina la separación entre el conglomerado 1 y los conglomerados 2 y 3 (Figura 51). Mientras que la componente principal 2 es la que determina la separación entre los conglomerados 2 y 3. Si comparo esta agrupación con la obtenida cuando se usó k = 2 estos dos últimos conglomerados representan el conglomerado que tiene los aspectos antropométricos más altos.

En otras palabras el conglomerado 1 que se encuentra conformado por todas las provincias del oriente a excepción de Sucumbíos obtuvo los aspectos antropométricos más bajos. El conglomerado 2 está conformado por provincias de todas las regiones del país las cuales son Galápagos, Esmeraldas, Guayas, Carchi, Santo Domingo, Cotopaxi, Carchi, Azuay y Sucumbíos. Este grupo es el que tiene más alta la componente principal 2, mostrando contrastes considerablemente altos entre el índice de masa corporal y cintura versus los valores de presión sistólica y diastólica. Este grupo tiene las tres provincias que se encuentra más al norte del país.

Finalmente puedo decir que el conglomerado 3 tiene valores intermedios en sus variables antropométricas y los contrastes entre las variables IMC y cintura versus la presión sistólica

y diastólica son bajos en comparación con el conglomerado 2. Está compuesta por 4 provincias de la costa Manabí, Santa Elena, Los Ríos y El Oro y 5 provincias de la sierra Loja, Chimborazo, Bolívar, Tungurahua, Pichincha e Imbabura. Resulta evidente que el Ecuador es un país que se encuentra poderosamente caracterizado por aspectos regionales, es decir, las provincias pertenecientes a una misma región son muy similares entre sí, principalmente las provincias de la Amazonía.

Esto se encuentra reflejado en la clasificación de las provincias con k medias igual a 3. Sin embargo, se puede afirmar que hay casos en los cuales las provincias son más parecidas a provincias contiguas que a provincias pertenecientes a su misma región e.g. Sucumbíos está fuera del grupo de las provincias de la Amazonía. Guayas y Esmeraldas no están en el grupo en el que la mayoría de sus integrantes son provincias de la costa. Mientras que las provincias de las sierra están divididas entre el conglomerado 2 y conglomerado 3 (Figura 50).

Adicionalmente se encontró que Galápagos, Zamora Chinchipe y Sucumbíos son las provincias que obtuvieron los valores más altos de la componente principal 2. Es decir, que sus valores de IMC y cintura son altos en relación a sus niveles de presión sistólica y diastólica. Por el contrario, Los Ríos y Manabí, ambas provincias de las Costa tuvieron los valores más bajos en esta componente, indicando que sus valores de IMC y cintura son bajos en relación a sus niveles de presión.

Los gráficos de caja expuestos en la figura 51 muestran, de manera más detallada las características fundamentales de cada grupo. Por un lado, el grupo 1 tiene los aspectos antropométricos más bajos de los tres grupos, sin tener ningún valor atípico dentro de la distribución de sus variables. Adicionalmente, los grupos 2 y 3 tienen valores más altos en sus variables especialmente para los valores de presión. Sin embargo para el IMC y cintura el grupo 2 tiene valores más altos. Ambos grupos presentan valores atípicos en estas dos variables.

Galápagos sobresale por encima del resto de provincias por poseer el IMC y el diámetro de cintura promedio más alto de todas. Sucumbíos, como se explicó en párrafos anteriores es la única provincia de la Amazonía que no se encuentra en el conglomerado 1 siendo la provincia con menos diámetro de cintura promedio del grupo 2. Para el grupo 3, Los Ríos (la única provincia de la costa que no tiene salida al mar) figura como la que tiene menos IMC y diámetro de cintura promedio de su grupo.

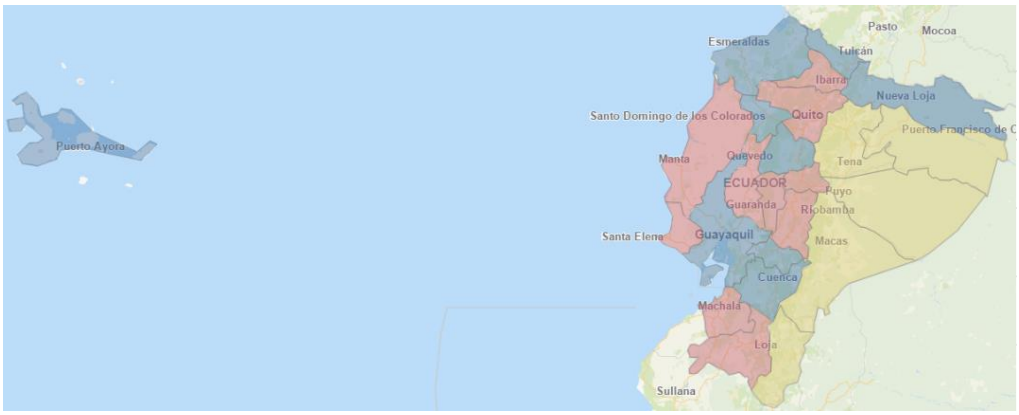


Figura 49. Mapa Ecuador. Aspectos Antropométricos K=3

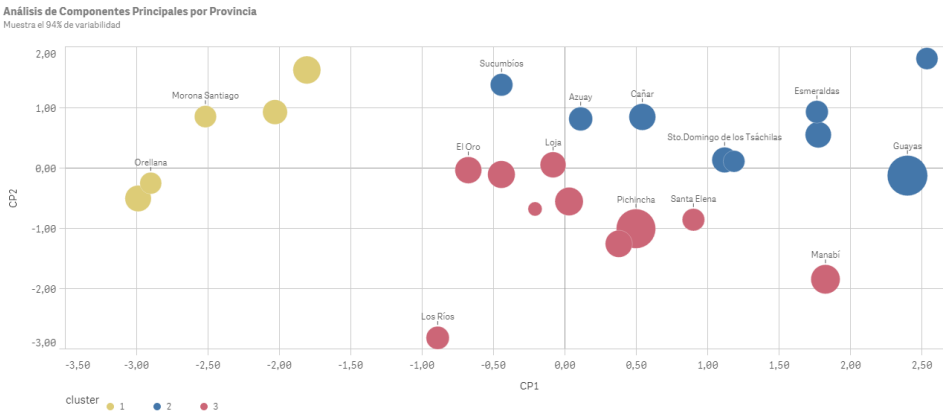


Figura 50. Componentes Principales. Aspectos Antropométricos. K=3

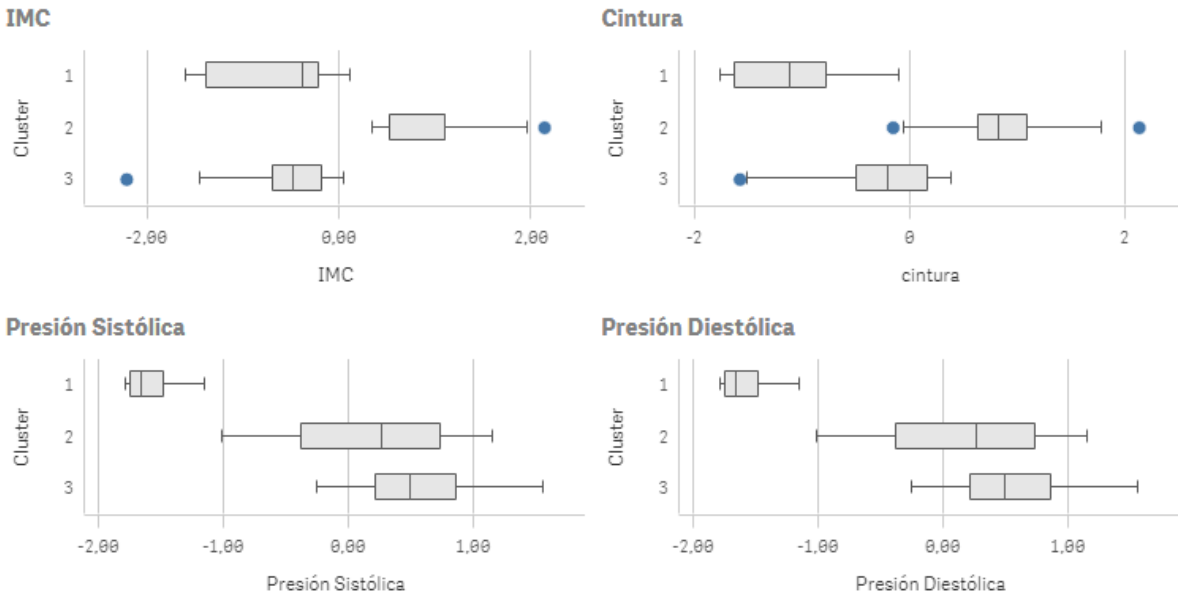


Figura 51. Gráficos de Cajas. Aspectos Antropométricos. K=3

De manera similar que en el caso de k=2 la mayor proporción de personas con diabetes e hipertensión se encontraron en los conglomerados 2 y 3 es decir, las grupos conformados mayoritariamente por provincias de la costa y de la sierra. Obteniendo el 5.7 y 5.8 por ciento para la hipertensión y 1.5 y 2.4 por ciento para la diabetes respectivamente. Las provincias de la costa son las que más casos de diabetes e hipertensión registran. La tabla 7 resume las características geográficas fundamentales de cada uno de los conglomerados con los dos valores de *k* que se usaron.

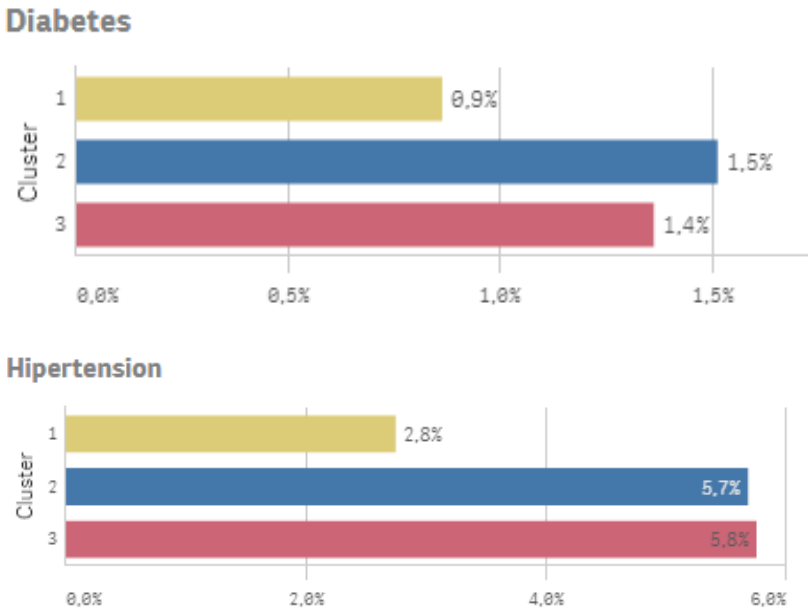


Figura 52. Proporción de Diabetes e Hipertensión. Aspectos Antropométricos. K=3

Aspectos Antropométricos k=2	
Cluster 1	Amazonía a excepción de Sucumbios
Cluster 2	Resto de provincias

Aspectos Antropométricos k=3	
Cluster 1	Amazonía a excepción de Sucumbios
Cluster 2	Tres provincias al extremo norte y al centro del pais y Galápagos
Cluster 3	Provincias del centro y sur del pais del pais

Tabla 7. Características Conglomerados Aspectos Antropométricos

5.3.2. Aspectos Bioquímicos

Análisis Univariado – Aspecto Geográficos

El 20% de la población ecuatoriana tiene hipercolesterolemia. El 16% muestra ldlc aumentado, mientras que el 36,5% tiene hdlc disminuido. Adicionalmente el 10% posee glucosa alterada en ayunas, mientras que el 1.33% tiene diabetes. El 20% de las personas

tienen hipertrigliceridemia. El 19% hipoinsulinemia mientras que 6.0% hiperinsulinemia. Además el 27% presenta casos de pre hipertensión, el 5% hipertensión y el 2.23% hipotensión.

Por otra parte, las provincias que sobresalen por tener niveles altos de colesterol son Morona Santiago, Imbabura y Zamora Chinchipe. Las provincias que poseen valores altos de ldlc son Morona Santiago, Zamora Chinchipe y Carchi. Las que tienen niveles bajos hdlc son Orellana, Bolívar y Loja.

Las tres provincias con valores de glucosa más alta son Napo, Los Ríos y Azuay. Las provincias que se destacan por tener niveles altos de triglicéridos son Bolívar, Imbabura y Orellana. Por el contrario, las provincias que tienen valores altos de insulina son Imbabura, Pastaza y Manabí. En cuento a la presión, Cotopaxi, Chimborazo y Morona Santiago son las provincias con la presión sistólica más alta. Mientras las provincias con la presión diastólica más alta son Cotopaxi, Chimborazo y Los Ríos.

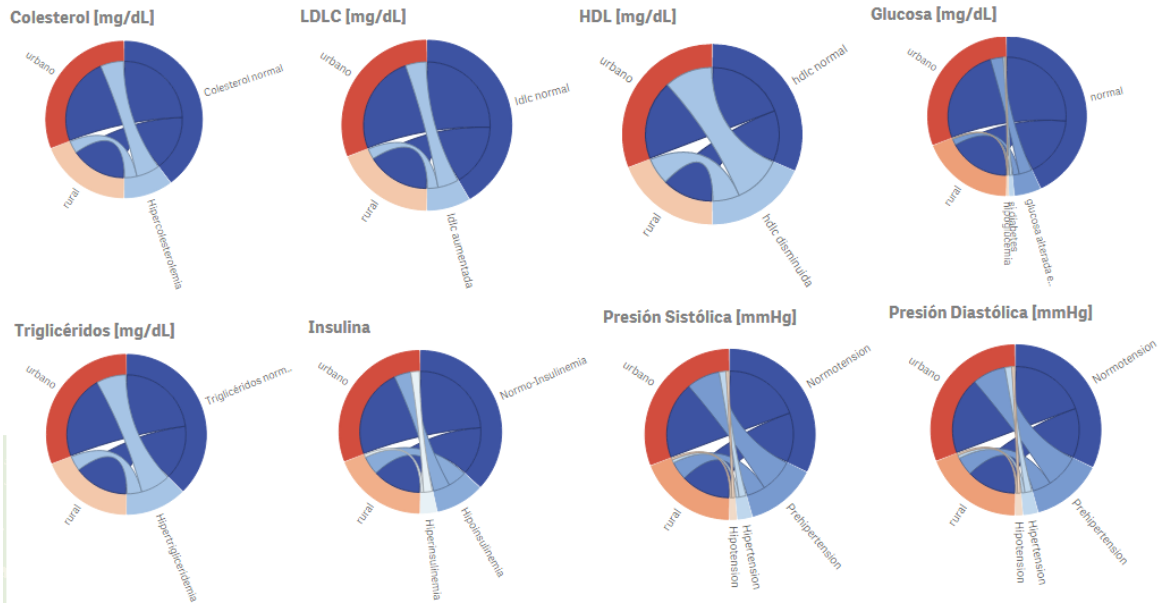


Figura 53. Proporción de Valores Normales de Variables Bioquímicas

Análisis Multivariado – Valores Normales

Aplicando un análisis de correspondencias se encontró que las personas de género masculino tienden a tener valores normales de hemoglobina, mientras que las mujeres a tener valores no normales de hemoglobina. Adicionalmente, existe una asociación alta entre grupos de indicadores bioquímicos tales como colesterol normal y triglicéridos normales, hdlc normal y ldlc normal, hipercolesterolemia y ldlc aumentada.

La mayor parte de la población tiende a tener normoinsulinemia, glucosa normal, a no tener diabetes y a no tener deficiencias de hierro. No existe una asociación aparente entre la

glucosa alterada en ayunas y otros indicadores bioquímicos. Hay muy pocos casos de diabetes (Figura 54).

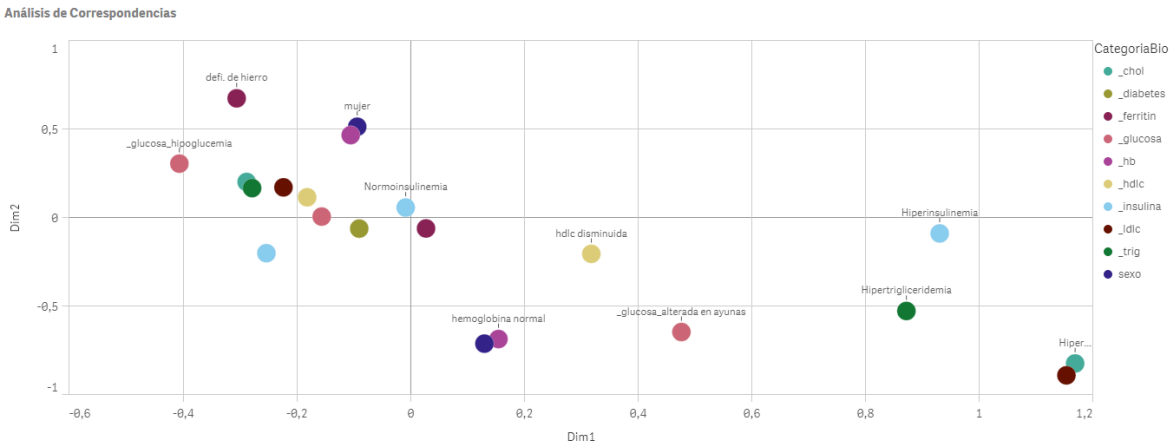


Figura 54. Análisis de Correspondencias. Valores Normales. Aspectos Bioquímicos

Análisis Multivariado – Componentes Principales y Conglomerados

Con la finalidad de estudiar las características de las 24 provincias del país en relación a las variables que están relacionadas con problemas cardiovasculares, se decidió realizar un análisis de correlaciones. Los resultados de este estudio indican que existe una alta correlación entre el colesterol y todas las variables, especialmente con ldlc, presión sistólica, insulina, hdlc. También hay correlaciones altas entre la presión sistólica y las variables presión diastólica, ldlc y glucosa. También se encuentran correlacionados pero en menor proporción los niveles de insulina y ldlc. (Figura 55).

Estos resultados me dieron el aval necesario para poder aplicar *componentes principales* como técnica de reducción de dimensiones. Las variables que se usaron para este análisis fueron colesterol, glucosa, hdlc, insulina, ldlc, presión sistólica, presión diastólica y triglicéridos. Se obtuvo el 86% de la variabilidad del conjunto de datos con las primeras tres componentes principales.

La primera componente es la que acumula la mayor variabilidad, alrededor del 62% y tiene todos sus coeficientes positivos, se interpreta como una componente de tamaño. Es decir, que mientras más alta sea esta componente, mayores serán los valores de las variables implicadas, en especial, los valores de colesterol, presión sistólica y ldlc (Tabla 7).

Por otro lado, la segunda componente acumula el 14% de la variabilidad del conjunto de datos, tiene coeficientes positivos y negativos, quiere decir que es una componente de forma o de contraste, Por lo tanto, mientras más alta sea su componente mayores serán los contrastes entre el grupo de variables conformado por colesterol, ldlc, hdlc y el grupo de variables compuesto por presión sistólica, presión diastólica, glucosa, insulina y triglicéridos [44]. Resaltando los contrastes entre triglicéridos y hdlc.

De la misma manera, la tercera componente acumula el 9% de la variabilidad de los datos, tiene coeficientes positivos y negativos, quiere decir que es una componente de forma o de contraste, Por lo tanto, mientras más alta sea su componente mayores serán los contrastes entre las variables insulina y glucosa versus el resto de variables implicadas. Resaltando los contrastes entre insulina y glucosa contra la presión diastólica (Tabla 7).

Aspectos Bioquimicos			
Variable	CP1	CP2	CP3
chol	0,42	-0,11	0,15
presa	0,40	0,12	0,08
ldlc	0,40	-0,16	0,04
presb	0,36	0,07	0,35
insulina	0,35	0,03	-0,46
glucosa	0,32	0,12	-0,67
hdlc	0,29	-0,64	0,22
trig	0,23	0,70	0,32

Tabla 8. Componentes Principales. Aspectos Bioquímicos

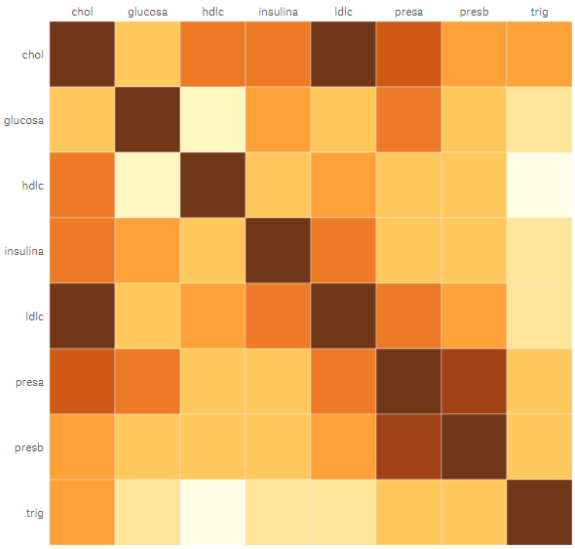


Figura 55. Análisis de Correlaciones. Aspecto Bioquímicos

Posteriormente, se aplicó diferentes métodos de conglomerados tales como *k - means*, *fuzzy clustering* y *PAM* y se utilizó coeficiente silhouette para identificar la tendencia al cluster de los datos; obteniendo los mejores resultados con *k – medias* con valores de *k* de 2 y 3. Para *k* igual a 2, con valor de silhouette de 0.42, se llegó a los siguientes resultados:

La componente principal 1 definió los conglomerados de la siguiente manera (Figuras 56 y 57). Las provincias que tienen los valores bajos de esta componente formaron parte de un mismo grupo, mientras que el resto de las provincias formaron parte del otro. El grupo 2 se caracterizó, principalmente porque las provincias que lo componen tienen valores bajos en todas las variables bioquímicas en relación al resto de provincias (Figura 57).

Este grupo lo componen todas las provincias de la Amazonía, en donde Sucumbíos es la provincia cuyo componente principal 1 es el más alto, es decir, la que tienen los valores más altos de las variables bioquímicas (especialmente de colesterol, presión sistólica y ldlc) del grupo. Mientras que Orellana y Morona Santiago son muy parecidas entre sí y son las que presentan los valores más bajos de colesterol e insulina. Pastaza, Napo y Zamora Chinchipe son provincias muy parecidas entre sí también en cuanto a las variables bioquímicas se refiere.

El grupo 1 es el más numeroso y lo componen todas las provincias de la costa, sierra y la región insular del país. En el gráfico 33 se puede apreciar que las provincias Pichincha y Guayas son muy parecidas entre sí, pese a que geográficamente se encuentran muy distantes. Esto se puede deber a que en estas provincias se encuentran las ciudades más densamente pobladas y las más importantes a nivel económico del país y posiblemente por el estilo y ritmo de vida son parecidas en cuanto a las variables bioquímicas se refiere.

Por otro lado, en la figura 56 y figura 57 también se puede apreciar que Azuay, El Oro y Loja son provincias muy parecidas entre sí y son geográficamente aledañas. También hay una relación de similitud entre Tungurahua y Cañar. Por otro lado, Manabí, Cotopaxi y Santa Elena son las que tienen los valores más altos de la primera componente principal tanto dentro del grupo como en todo el conjunto de datos.

Analizando el figura 57 el tamaño de las burbujas representa el valor de la componente principal 3, en donde sobresalen las provincias de El Oro, Galápagos, Santa Elena y Pastaza como las provincias con menos valor en esta componente lo cual implica tienen valores de presión diastólica, triglicéridos y hdlc son altos mientras que para la insulina y glucosa presentan valores bajos.

Los gráficos de caja en la figura 58 muestran claramente la distribución de las variables agrupadas por conglomerado. Como se explicó en párrafos anteriores el grupo 1 se caracteriza por tener los valores más altos en cada una de las variables, Manabí sobresale por tener los valores más altos de triglicéridos y presión diastólica. Cotopaxi figura como la provincia con valores más altos de ldlc. Carchi y Santa Elena con altos niveles de hdlc y Santa Elena con los valores más altos de insulina. Por otro lado Orellana y Morona Santiago figuran como las provincias con los valores más bajos de Colesterol e Insulina.

Finalmente, cabe resaltar que en cuanto a la proporción de personas con diabetes e hipertensión se refiere se registraron prácticamente los mismos patrones que en el análisis de los aspectos antropométricos. Y también que si comparamos la clasificación obtenida, versus la clasificación que otorgó k medias para los aspectos antropométricos k igual a 2 se encontró qué, si no se considerara a la provincia de Sucumbíos, la clasificación sería esencialmente la misma (Figura 56).

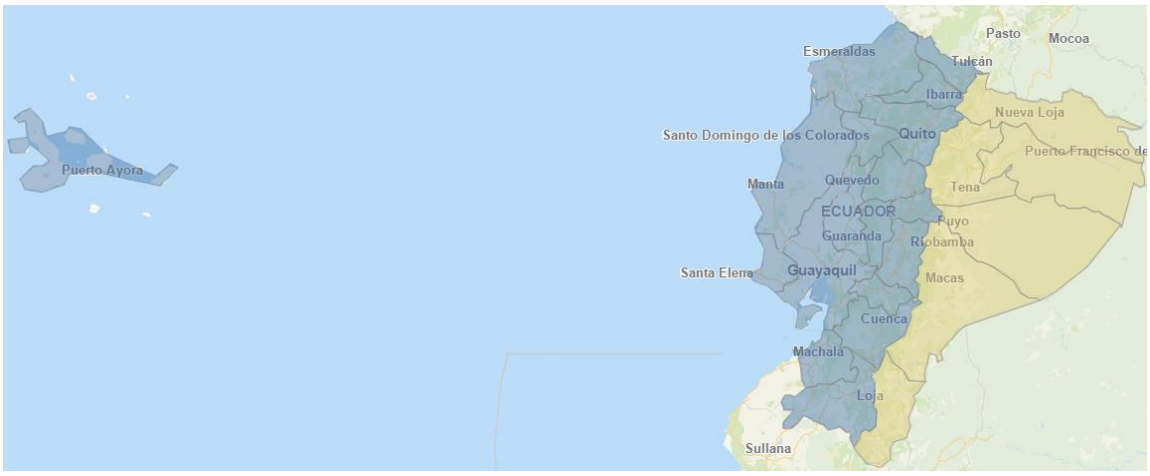


Figura 56. Mapa Ecuador. Aspectos Bioquímicos. K=2

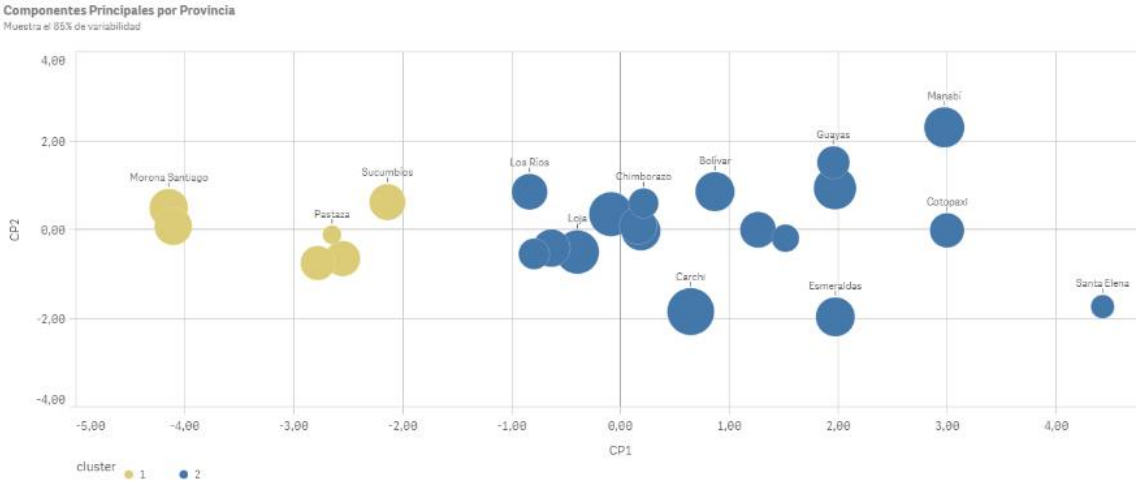


Figura 57. Componentes Principales. Aspectos Bioquímicos. K=2

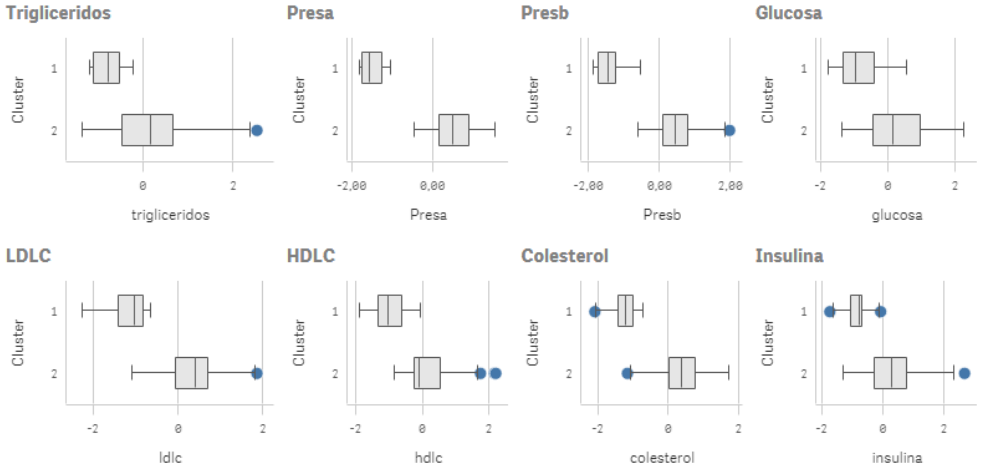


Figura 58. Gráficos de Cajas. Aspectos Bioquímicos. K=3

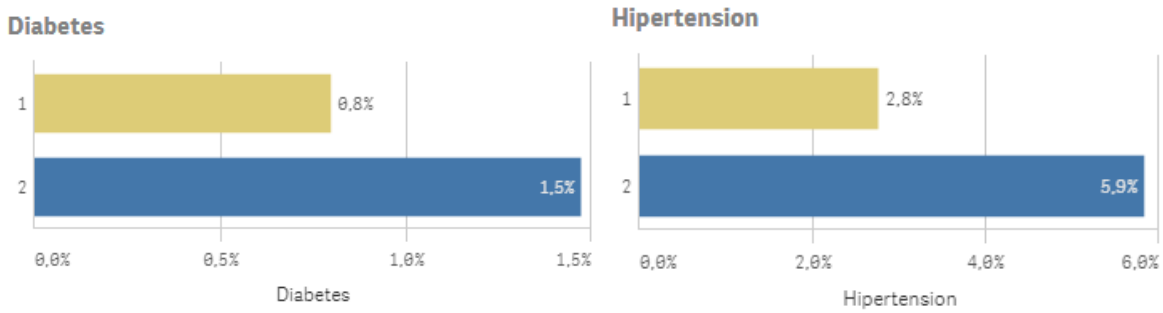


Figura 59. Proporción de Diabetes e Hipertensión. Aspectos Bioquímicos. K=2

Para *k* igual a 3, con valor de silhouette de 0.33, se obtuvieron los siguientes resultados:

Al igual que en el caso anterior la primera componente principal es la que determina la separación de los conglomerados. El conglomerado 2, caracterizado por tener valores bajos en las variables bioquímicas se mantiene inalterable en esta clasificación, es decir, que está conformado por las mismas provincias que se encontraron en el caso anterior. Mientras que el grupo 1 se divido en dos partes esta separación la realiza esencialmente la componente principal 1. Agrupando a las provincias con valores intermedios en esta componente en un grupo y a los que tienen los valores más altos en otro grupo. (Figura 60 y 61).

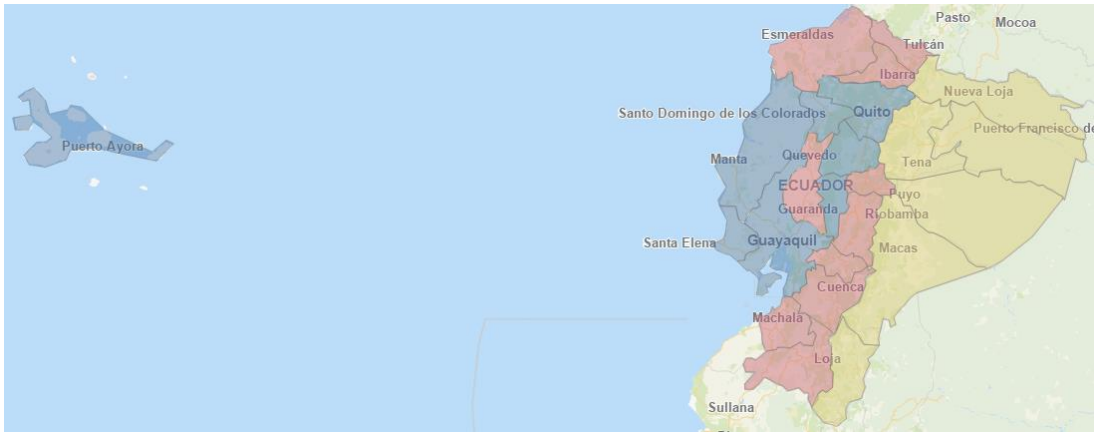


Figura 60. Mapa Ecuador. Aspectos Bioquímicos. K=3

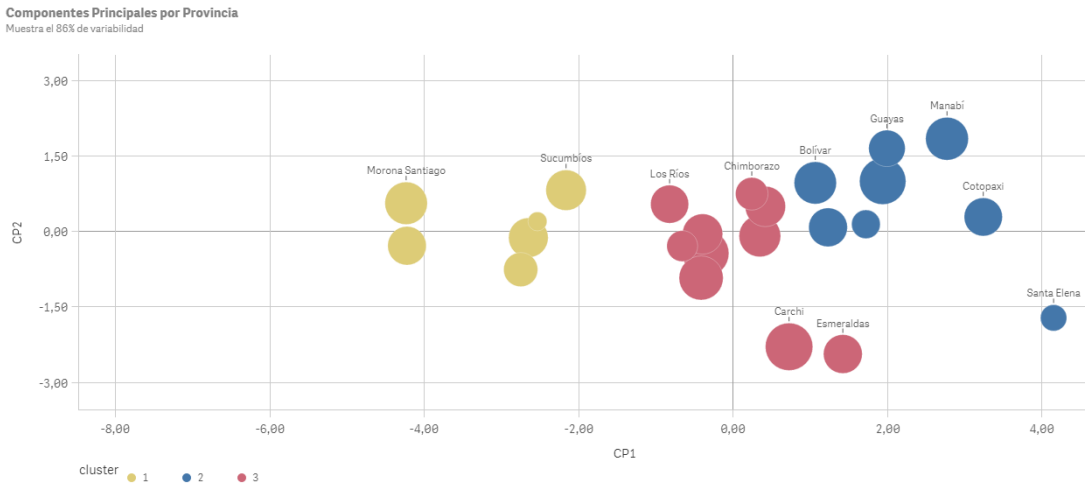


Figura 61. Análisis de Componentes Principales. Aspectos Bioquímicos. K=3

El grupo intermedio está constituido por tres provincias de la costa (Esmeraldas, Los Ríos y El Oro) y por provincias de las sierra Carchi, Imbabura, Tungurahua, Chimborazo, Cañar, Azuay y Loja. El grupo con variables bioquímicas altas está constituido por el resto de provincias de la sierra y de la costa y la región insular las cuales son Santo Domingo, Pichincha, Cotopaxi, Bolívar, Manabí, Santa Elena, Guayas y Galápagos.

En términos generales, se pudo apreciar, cierta correlación entre los valores de las variables bioquímicas (relacionadas con los problemas cardiovasculares) y las regiones del país, para ambos valores de k en el algoritmo k medias ya que en ambos casos todas las provincias de la Amazonía pertenecen a un mismo conglomerado. Esta correlación también se evidencia en los aspectos antropométricos también puesto que el grupo con los menores valores en estas variables permanece inalterable.

Analizando los valores de la segunda componente principal se encontró que Manabí es la provincia que más niveles de triglicéridos tiene en relación a los de hdlc. Mientras que Esmeraldas posee los valores más bajos de triglicéridos en relación al hdlc. Por otro lado investigando la tercera componente principal se pudo evidenciar que Carchi es la provincia con los valores más altos de triglicéridos y presión diastólica en relación a la insulina y glucosa, mientras que Pastaza y Santa Elena tienen los valores más bajos de hdlc y presión diastólica en relación a la insulina y glucosa.

Analizando la proporción de personas con diabetes e hipertensión, se encontró que el grupo caracterizado por tener los valores más altos de las variables bioquímicas (grupo número 2) es, a su vez, el que registra más casos de diabetes e hipertensión. Lo característico de este conglomerado a nivel geográfico es que lo componen las provincias centro occidentales del país a excepción de la provincia de Los Ríos.

Por otra parte, el conglomerado constituido por las provincias de la Amazonía como se explicó en párrafos anteriores registra las proporciones más bajas de diabetes e

hipertensión. Mientras que en el caso del conglomerados con niveles intermedios de las variables bioquímicas la proporción de personas con diabetes e hipertensión es mayor que el conglomerado de las provincias Amazónicas y menor que el conglomerado de las provincias centro occidentales. Además otra característica de este grupo es que la mayoría de sus miembros se ubican al centro sur del país. La tabla 9 resume las características geográficas fundamentales de cada uno de los conglomerados con los dos valores de *k* que se utilizaron.

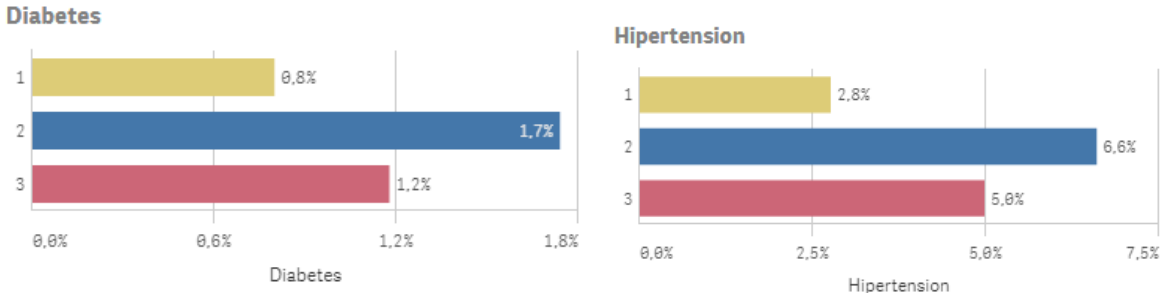


Figura 62. Proporción de Diabetes e Hipertensión. Aspectos Bioquímicos. K=3

Aspectos Bioquímicos k=2	
Cluster 1	Amazonía
Cluster 2	Resto de provincias

Aspectos Bioquímicos k=3	
Cluster 1	Amazonía
Cluster 2	Provincias Centro Occidentales y Galápagos
Cluster 3	Mayoria al Centro Sur del Pais

Tabla 9. Características Conglomerados Aspectos Antropométricos

5.3.3. Consumo de Alimentos

En esta sección se analizó el contenido de proteína, grasa y carbohidratos en la alimentación del país llegando a los siguientes resultados:

Todas las provincias de la costa junto con la provincia de Galápagos, Pichincha y Santo Domingo (estas últimas son colindan con provincias de la costa) son las provincias que consumen más grasas, proteínas y carbohidratos de todo el país, especialmente la provincia del Guayas.

Por otro lado, Sucumbíos es la provincia que se destaca por tener el consumo de proteínas, grasas y carbohidratos más cercano al promedio. Tal y como muestra el siguiente gráfico el resto de provincias de la Amazonía y la gran mayoría de provincias de la sierra consumen menos proporción de estos macro nutrientes. Sobresale la provincia de Morona Santiago como la provincia con el menor consumo de macro nutrientes.

Adicionalmente las figuras 63 y 64 muestran que las provincias con mayor proporción de diabetes e hipertensión son las provincias de la costa, junto con Galápagos y algunas provincias de la sierra incluyendo Pichincha, Imbabura, Carchi y Tungurahua en el caso de la hipertensión y Chimborazo, Santo Domingo y Loja para el caso de la diabetes. Pastaza figura como la única provincia de la Amazonía en el ranking de la diabetes pese a que sus valores medios de proteína, grasa y carbohidratos son relativamente bajos en relación al resto de provincias del top.

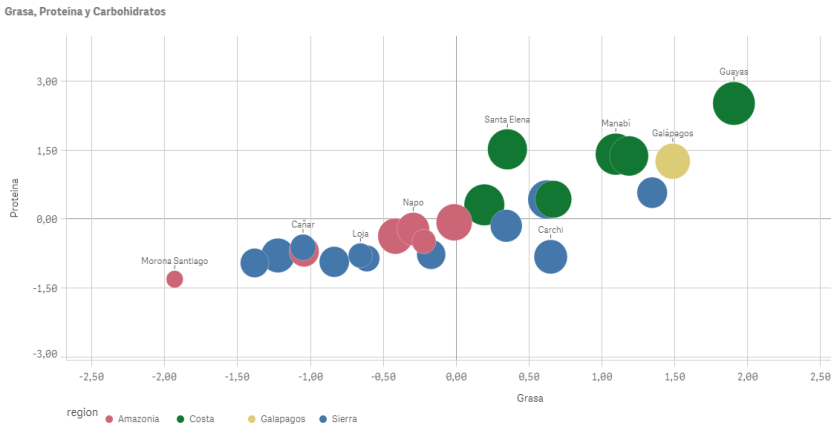


Figura 63. Consumo de Grasa, Proteína y Carbohidratos por provincia y región

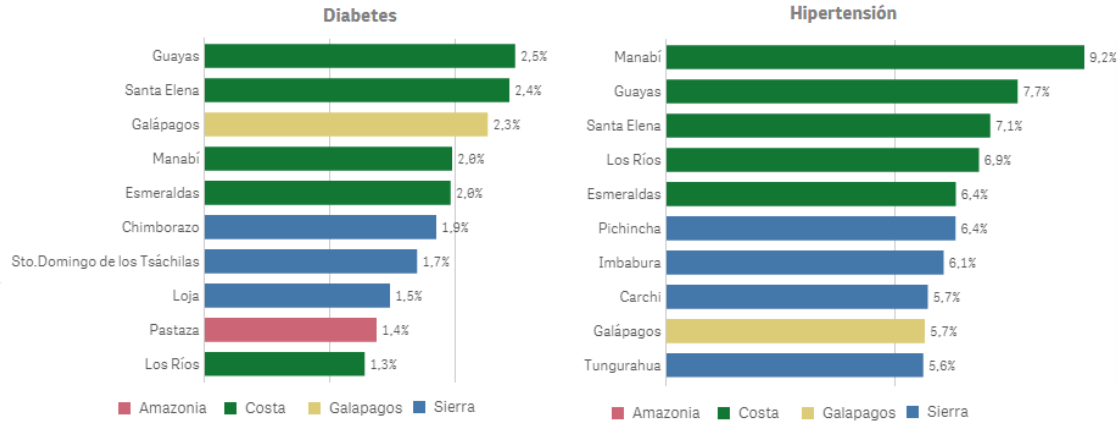


Figura 64. Proporción de Diabetes e Hipertensión por provincia.

Provincias con Particularidades

La tabla 10 resume las características más relevantes de las provincias que sobresalen en cada uno de los aspectos estudiados. La provincia de Sucumbíos, tal y como reveló el análisis de los aspectos antropométricos es muy diferente al resto de provincias de la Amazonía, se destaca por estar en el conglomerado que tiene el IMC y el diámetro de la cintura altos. Sin embargo para el caso de las variables bioquímicas analizadas tales como triglicéridos, glucosa, ldlc, hdlc, colesterol, insulina y presión sistólica y diastólica tiene

niveles bajos. No obstante es la provincia con los valores medios de grasa, proteína y carbohidratos más cercanos a la media.

Por otro lado, Napo y Orellana, ambas provincias de la Amazonía, destacan por ser las provincias con IMC y diámetro de la cintura más bajas de todo el país. De la misma manera pertenecen al grupo que tiene las variables bioquímicas más bajas y, a su vez, sus los valores medios de grasa, proteína y carbohidratos también bajos. Son provincias que al igual que Sucumbíos, no aparece en el listado de provincias con mayores casos de diabetes e hipertensión.

Continuando con el análisis, la provincia de Galápagos es una muy particular ya que sus valores medios de IMC y cintura son los más altos del país, y a su vez, se encuentra en el grupo de provincias con los niveles de triglicéridos, glucosa, ldlc, hdlc, colesterol, insulina y presión sistólica y diastólica más altos, lo mismo pasa con el consumo de alimentos ya que su dieta se basa en alimentos con altos niveles de grasas y proteínas.

Adicionalmente esta provincia se encuentra en el top 3 de las que más casos de diabetes registran y ocupa el noveno lugar en las que tienen más hipertensión. Estas características alimenticias probablemente se deban a que al ritmo de vida sedentario que llevan las personas que habitan en estas islas.

Continuando con el análisis se encuentra la provincia de los Ríos, la única provincia de la costa del Ecuador que no tiene salida al mar. Esta provincia presenta un perfil medio en tanto en los aspectos antropométricos como en las variables bioquímicas y de consumo de alimentos. En cuento a los casos de diabetes ocupa el décimo lugar y el cuarto en el top de los casos de hipertensión.

En contraste con las provincias de Los Ríos, Manabí, otra provincia de la costa muestra valores altos de presión sistólica y diastólica, altos niveles de triglicéridos, glucosa, ldlc, hdlc, colesterol, insulina y presión sistólica y diastólica y también dietas caracterizadas por tener niveles altos de grasas y proteínas. Es una provincia muy similar a Galápagos considerando estos aspectos. A su vez, es la provincia que registra más casos de hipertensión en todo el país y en relación a la diabetes ocupa el cuarto puesto.

Finalmente están las dos provincias más importantes a nivel económico del país. Pichincha y Guayas. En el caso de la provincia de Pichincha, sobresale por tener valores de presión sistólica y diastólica altas, y variables bioquímicas altas tales como triglicéridos, glucosa, ldlc, hdlc, colesterol, insulina. Sin embargo en cuento al consumo de alimentos destaca por consumir alimentos altos en grasa y medios en proteínas. No parece en el ranking de provincias con mayores casos de diabetes pero si en el ranking de provincias con hipertensión, ocupando el sexto lugar.

Para el caso de la provincia del Guayas, se destaca por encabezar el ranking de las provincias con mayores casos de diabetes e hipertensión. Además, en cuento al consumo de alimentos se refiere sus niveles de grasa y proteína están muy por encima del resto de provincias (Figura 63). Sus niveles de cintura e IMC son altos también así como sus variables bioquímicas.

Provincia	Analisis Antropometrico	Analisis Bioquimico	Consumo de Alimentos	Top Diabetes	Top Hipertension
Sucumbíos	IMC y cintura altas	Variables bioquimicas bajas	Valores medios de grasa, proteína y carbohidratos	0	0
Napo y Orellana	IMC y cintura más bajas del país	Variables bioquimicas bajas	Valores de grasa, proteína y carbohidratos bajos	0	0
Galápagos	IMC y cintura las más altas	Variables bioquimicas altas	Valores altos de proteínas y grasas	3	9
Los Rios	IMC y cintura intermedio	Variables bioquimicas intermedios	Valores medios de grasa, proteína y carbohidratos	10	4
Manabí	Presión sistólica y diestólica alta	Variables bioquimicas altas	Valores altos de proteínas y grasas	4	1
Pichincha	Presión sistólica y diestólica alta	Variables bioquimicas altas	Valores altos de grasa e intermedios de proteína	0	6
Guayas	IMC y cintura altas	Variables bioquimicas altas	Valores más altos de grasa, proteína de todas las provincias	1	2

Tabla 10. Provincias con particularidades

6. Discusión

Tal y como se detalla en el punto 2.3.1. Según el artículo, *Applied visual analytics for exploring the National Health and Nutrition Examination Survey* (NHANES) [50] se aplicaron técnicas de visualización para analizar la encuesta de salud y nutrición de la población estadounidense en el 2010. En este sentido, el mencionado trabajo guarda una estrecha semejanza con la herramienta de visualización desarrollada ya que también se utilizaron datos de una encuesta de salud y nutricional a nivel nacional. Sin embargo, en este caso se usaron datos de Ecuador.

Según el artículo, se utilizaron dispersogramas, gráficos de barras y filtros tales como grupo etario, sexo, etnia y el número de clusters. Además se hizo énfasis en analizar la calidad de la alimentación estadounidense en base a los Índices de Alimentación Saludable (HEI) y a una Guía Nutricional (DGA). En este sentido también hay una gran similitud con la aplicación de visualización desarrollada puesto que se usan técnicas tales como *brushing, linking y zoom and filter*. Teniendo la posibilidad de realizar varias selecciones y que todas las visualizaciones se vean afectadas por ellas.

Sin embargo, la presente tesis se enfocó en analizar las enfermedades crónicas no transmisibles así como las variables y factores relacionados a estas. Adicionalmente, la herramienta de visualización desarrollada permite hacer un análisis más profundo del estado de salud y nutrición ya que considera no solo el consumo de alimentos, sino también los aspectos geográficos demográficos, sociales, bioquímicos de la población. Así como el uso de tecnicas de visualización tales como mapas de calor, mapas geográficos, gráficos de cajas *dependency wheel*, entre otras. También se usan técnicas de datamining tales como análisis de clusters, arboles de decisión, etc.

En el trabajo expuesto por Padua [51] se desarrolló una herramienta de visualización para la generación automática de múltiples árboles de decisión y la posterior evaluación de los resultados en forma visual e interactiva. El enfoque principal de esta herramienta es analizar la incidencia de los parámetros de configuración de los árboles de decisión.

Se usó el algoritmo CHAID, histogramas, mapas de calor, tablas dinámicas, gráficos de dispersión, entre otras visualizaciones. En ese sentido, la herramienta de visualización desarrollada en la presente tesis guarda cierta relación con el trabajo expuesto por Padua, puesto que si bien esta herramienta no permite la configuración dinámica de los parámetros de los árboles de decisión si permite analizar los resultados de los mismos de forma dinámica e interactiva.

Uno de los hallazgos más relevantes del presente trabajo es la aparición de la ferritina como una de las variables más significativas y que más información aportan en la diabetes e hipertensión. Al respecto vale la pena destacar sobre la mencionada proteína que no solo se encuentra relacionada a las deficiencias de hierro en el organismo, como es sabido, sino

también a procesos inflamatorios agudos, según se pudo constatar tras una larga investigación con especialistas en nutrición y en ECNT, entre los cuales se encuentra mi codirector de tesis [52].

Acorde a las investigaciones realizadas por de Myiaki [18] las variables más relevantes en cuanto a la enfermedades cardiovasculares son el peso y la edad, los mismas que figuran como altamente relevantes para la diabetes e hipertensión en el presente trabajo. Adicionalmente, acorde a las investigaciones llevadas a cabo por Sigurdardottir la "Intervención en la educación diabética" tiene mucho impacto en cuanto a los niveles de glucosa en la sangre se refiere [31].

Según las investigaciones realizadas por Olynyk JK, Cullen DJ, Aquila S, la mayoría de los pacientes (90%) con hiperferritinemia no tienen sobrecarga de hierro [55] [56] muchas condiciones se asocian con altos niveles de ferritina y estas pueden coexistir, entre ellas se encuentran las enfermedades hepáticas, el exceso de alcohol, afecciones inflamatorias agudas, infecciones y el síndrome metabólico (obesidad, diabetes tipo 2, dislipidemia e hipertensión y trastornos inflamatorios) [55]. Esto se encuentra en concordancia con los resultados obtenidos en el árbol de decisión cuya variable principal es la ferritina (Punto 5.2.1), el cual muestra una clara correlación entre valores altos de ferritina y casos de diabetes e hipertensión.

Según los resultados obtenidos en el análisis de componentes principales y el análisis de conglomerados para las variables antropométricas se pudo apreciar claramente que Sucumbíos es la provincia de la Amazonía con los índices más altos. Acorde al artículo publicado en *EcuadorInmediato* [56] la acción extractiva de las petroleras transnacionales ha afectado severamente las condiciones de vida de los habitantes de esta provincia.

Según este articulo el 81.73% de la población total vive en condiciones de pobreza por necesidades básicas insatisfechas y el 31% de la población de entre 10 a 19 años aportan ingresos al hogar. En otras palabras, la acción de las petroleras en esta región amazónica ha influenciado enormemente a que la población adopte un estilo de vida muy distinto provocando que sus indicadores antropométricos se disparen en relación al de sus homologas de la Amazonía.

7. Conclusiones

En relación al primero de los objetivos específicos se aplicó satisfactoriamente el proceso de *KDD* con un proceso de *ETL* bastante importante puesto que consistió tanto en unificar diversas fuentes de información, tales como datos bioquímicos, antropométricos, demográficos, económicos y alimentarios de la población ecuatoriana (recopilada en el 2012) hasta contar con una estructura de datos que permitiese tanto agrupar todos los aspectos a analizar cómo realizar un minucioso trabajo de exploración de datos que permitió definir conjuntamente con el especialista en medicina los criterios de inclusión y exclusión de casos. Ver tabla 4. En cuento a parte de visualización de datos, se creó satisfactoriamente un sistema de visualizaciones que se basa fundamentalmente en el *Overview, Zoom and Filter and Details on Demand* de Daniel Keim [48]

En cuanto al objetivo específico número dos se aplicó técnicas de visualización tales como *coordenadas paralelas* y *gráficos de cajas* que permitieron identificar junto con el especialista los valores atípicos a excluir del estudio debido a posibles errores de tipeo (luego de haber escogido con el especialista las variables a ser estudiadas, las cuales son las variables que más relación tienen con la diabetes e la hipertensión); así como valores atípicos posibles de generarse debido a la naturaleza de las variables y a características específicas de ciertas regiones del país.

Adicionalmente, los *gráficos de cajas* me permitieron obtener una visión amplia de la distribución y valores atípicos que tienen las variables luego de realizar el proceso de exclusión de casos erróneos. De tal manera que se pudo identificar a la *insulina, índice homa, glucosa y triglicéridos* como las variables que tienen una concentración de datos a la izquierda de la distribución, es decir, que tienen muy pocos datos altos, sin embargo estos son significativamente altos en relación al resto de sus valores. En relación a las variables restantes no se identificó una distribución claramente positiva o negativa. En realidad las variables restantes tienen una distribución parecida a una distribución normal.

Siguiendo con el objetivo específico número dos, dentro del grupo de variables relacionadas con la diabetes, aplicando árboles de decisión con el método CHAID se encontró que las variables más relevantes para la diabetes son glucosa, índice homa, triglicéridos, edad, cintura, ferritina, resistencia a la insulina, estado nutricional, colesterol y peso. En este listado se pudo identificar varias cosas. La primera es, que existe cierta similitud entre grupos de variables de este listado.

Por ejemplo, las primeras tres variables, es decir, las que más ganancia de información generan tienen una marcada distribución asimétrica positiva esto seguramente se debe a que hay pocos casos con valores excesivamente altos, lo cual no quiere decir que una persona con valores normales de estas variables especialmente de glucosa no tenga diabetes, ya que puede ser que la persona esté siendo tratada. La variable edad y el “diámetro de la cintura”

tienen una distribución simétrica en donde se registran más casos de diabetes en personas con edades superiores a los 40 años y con el diámetro de la cintura superior a los 100 cm.

El caso de la ferritina es el más interesante de todos debido a que esta proteína que es la encargada de transportar el hierro no se encuentra vinculada directamente con la diabetes. Sin embargo ocupa el sexto lugar en el ranking, inicialmente esta variable es la encargada de conectar la información entre los macronutrientes (proteínas, grasas y carbohidratos) y micronutrientes (yodo, zinc, vitamina A, vitamina B12, etc) con la diabetes. Sin embargo acorde al punto 6 esto se debe a que los valores altos de ferritina están relacionados con proceso inflamatorios agudos de diabetes tipo 2 e hipertensión. Por último tenemos al *IMC*, el colesterol y el peso, las cuales son variables relacionadas con el estado nutricional.

Continuando con el objetivo número dos, dentro del grupo de variables relacionadas con la hipertensión aplicando arboles de decisión con el método CHAID se encontró que las variables más relevantes para esta ECNT son la presión sistólica, presión diastólica, peso, cintura, edad, IMC, triglicéridos, ferritina, talla y colesterol. Aquí se encontró algunas similitudes y diferencias respecto al listado del ranking de variables relacionadas con la diabetes.

Primeramente, las dos variables más importantes son, por supuesto la presión sistólica y diastólica debido a que son ellas las que determinan si la persona tiene o hipertensión arterial, en este punto hago la misma consideración que el caso de la diabetes, es posible que haya personas con hipertensión pero que al ser tratadas pueden poseer niveles normales en estas variables. Las siguientes cuatro variables peso, cintura, edad e IMC, están relacionadas directamente con el estado nutricional de la personas, para las dos ECNT la edad y el diámetro de la cintura, ocupan el mismo lugar en el ranking, entendiéndose así que estas variables generan relativamente la misma proporción de información tanto para la diabetes como para la hipertensión.

Luego aparecen los triglicéridos, un tipo de grasa que al presentarse en cantidades altas puede generar problemas cardiovasculares, esta variable figura más importante en la diabetes. Al igual que en el caso de la diabetes, aparece la ferritina en este listado debido a la existencia de procesos inflamatorios agudos en la hipertensión. A diferencia del caso de la diabetes, para la hipertensión figura la talla de la persona como otra variable importante en el ranking. Finalmente el colesterol ocupa el mismo nivel de importancia tanto en la diabetes como en la hipertensión.

En relación al tercer y cuarto objetivo específico aplicando análisis de componentes principales y análisis de conglomerados se encontró algunos patrones interesantes los cuales se describen a continuación. Existe una clara diferenciación entre las provincias de la Amazonía y el resto de provincias del país tanto para las variables antropométricas, como para las variables bioquímicas.

A excepción de Sucumbíos, las provincias amazónicas son muy similares entre sí, esta provincia tiene más similitudes en cuanto a los aspectos antropométricos con Esmeraldas y Carchi que son provincias que al igual que Sucumbíos se encuentran al norte del país. A su vez, estas provincias son las que menor proporción de personas con diabetes e hipertensión se encontraron.

El análisis de conglomerados desde la perspectiva de los aspectos antropométricos reveló un dato muy interesante. Las tres provincias que tienen frontera con Colombia, junto con algunas provincias de la sierra central del país (Azuay, Cañar, Cotopaxi y Santo Domingo) y las provincias de Galápagos y Guayas conforman un grupo que se caracteriza por tener altos valores de cintura e índice de masa corporal en relación a sus valores de presión sistólica y diastólica, a su vez, este grupo también se caracteriza porque todas las provincias que lo componen tienen en promedio Sobrepeso. Este dato sería altamente importante al momento de realizar políticas de salud.

Continuando con este análisis, se encontró un grupo de provincias caracterizado por tener los valores de IMC y cintura intermedios en relación al grupo anterior y al grupo de las provincias Amazónicas, a su vez, posee valores de presión similares a las del grupo anterior. Las provincias que están dentro de este conglomerado se encuentran ubicadas mayoritariamente en la Sierra Central y Sur del país junto con tres provincias de la costa Santa Elena, Manabí y Los Ríos, esta última tiene el promedio de IMC más bajo de todo el país.

A pesar de que este grupo de provincias tiene mucho menos IMC y cintura que el grupo que se mencionó anteriormente ambos grupos tienen prácticamente la misma proporción de personas con diabetes e hipertensión. Para terminar de realizar este análisis se puede afirmar que hay tres provincias que llaman grandemente mi atención. En primer lugar Galápagos, que es la provincia con más IMC y cintura y es la tercera provincia con más casos de diabetes.

La provincia de Los Ríos que tiene el IMC más bajo del país, lo cual la hace la provincia de la costa que más difiere con sus homologas de la región. Por último puedo señalar a Sucumbíos como la provincia más distinta en relación a sus homologas de la Amazonía siendo la provincia con menos diámetro de cintura promedio dentro del conglomerado que tienen los valores más altos en esta variable.

El análisis de conglomerados desde la perspectiva de los aspectos bioquímicos reveló también algunos datos interesantes. Aquí si existe una total diferenciación entre las provincias de la Amazonía y el resto del país puesto que Sucumbíos, a diferencia del análisis de los aspectos antropométricos si forma parte del conglomerado de las provincias del Oriente. La característica fundamental de este grupo al igual que en el caso de los

aspectos antropométricos es que, proporcionalmente hablando, tiene los valores más bajos en todas las variables bioquímicas (hdlc inclusive).

Este análisis reveló también que provincias ubicadas en centro occidente del país tales como Pichincha, Santo Domingo de los Tsáchilas, Cotopaxi, Bolívar, Guayas, Santa Elena y Manabí, junto con Galápagos son parte del conglomerado con mayores casos de diabetes e hipertensión. A su vez, tienen los niveles más altos de cada una de las variables bioquímicas, presión sistólica y diastólica y glucosa, siendo Manabí y Santa Elena las que más insulina y hdlc tienen respectivamente.

Adicionalmente, este análisis permitió identificar al tercer conglomerado como el que tiene los niveles de triglicéridos, colesterol, insulina, ldlc, hdlc, presión sistólica, presión diastólica y glucosa intermedios en relación a los tres grupos, el mismo que es esencialmente conformado por provincias del sur y norte del país, conjuntamente con la provincia de Los Ríos. La proporción de personas con diabetes en este grupo son menores que el conglomerado anterior pero mayores que en el caso de las provincias Amazónicas.

Finalmente, en respuesta al objetivo general se complementó el análisis de ENSANUT mediante un sistemas de visualizaciones que consta de gráficos de caja, gráficos de barras, dispersogramas, mapas de calor, mapas geográficos, árboles de decisión, *ruedas de dependencia*, análisis de correlaciones, componentes principales, correspondencias, conglomerados para detectar los patrones mencionados en los párrafos anteriores.

8. Recomendaciones

Fortalecer las acciones de prevención en aquellas provincias con mayores niveles de diabetes, hipertensión y con los factores relacionados con estas, tales como Guayas, Santa Elena, Galápagos, Manabí, Esmeraldas, Los Ríos, Pichincha, Imbabura, Carchi y Galápagos (Figura 58).

Desarrollar políticas con un enfoque geográfico, tener en cuenta que Sucumbíos es la provincia del Oriente con los aspectos antropométricos más altos. Hay que considerar también que las tres provincias que tienen frontera con Colombia, junto con algunas provincias de la sierra central del país (Azuay, Cañar, Cotopaxi y Santo Domingo).

Las provincias de Esmeraldas, Galápagos y Guayas conforman un grupo que se caracteriza por tener altos valores de cintura e índice de masa corporal en relación a sus valores de presión sistólica y diastólica. A su vez, este grupo también se caracteriza porque todas las provincias que lo componen tienen en promedio Sobrepeso.

Adicionalmente, provincias de la Sierra Central y Sur tales como, Pichincha, Imbabura, Bolívar, Chimborazo y Loja junto con provincias de la costa Santa Elena, Manabí, El Oro y Los Ríos, tienen los niveles de IMC y cintura más cercanos al promedio. Se recomienda hacer hincapié en provincias tales como Galápagos, por tener los niveles de IMC y cintura más altos del país. Los Ríos, por ser la provincia con los índices más bajos de IMC. Morona Santiago y Santa Elena por tener los niveles más altos de insulina y hdlc.

Entre futuros aspectos a ser investigados está añadir una variable que indique si la persona tiene o no antecedentes de diabetes e hipertensión. Desarrollar varios modelos predictivos que permitan identificar las variables más importantes para determinar los valores normales de los aspectos antropométricos y bioquímicos.

9. Anexos

Anexo 1: Tabla de provincias y regiones.

Provincia	Región
Morona Santiago	Amazonia
Napo	Amazonia
Orellana	Amazonia
Pastaza	Amazonia
Sucumbíos	Amazonia
Zamora Chinchipe	Amazonia
El Oro	Costa
Esmeraldas	Costa
Guayas	Costa
Los Ríos	Costa
Manabí	Costa
Santa Elena	Costa
Galápagos	Insular
Azuay	Sierra
Bolívar	Sierra
Cañar	Sierra
Carchi	Sierra
Chimborazo	Sierra
Cotopaxi	Sierra
Imbabura	Sierra
Loja	Sierra
Pichincha	Sierra
Sto.Domingo de los Tsáchilas	Sierra
Tungurahua	Sierra

10. Bibliografía

[1] Enrique Ayala Mora (2002). Ecuador. Patria de Todos. Universidad Andina Simón Bolivar.

[2] Tierra del Volcán (2018). Ecuador. 20/01/2019. Sitio Web: http://www.tierradelvolcan.com/espanol/tierra-del-volcan/ecuador/

[3] Ecuador Explorer (2018). Datos y Geografía del Ecuador. 20/01/2019. Sitio Web: http://www.ecuadorexplorer.com/es/html/ubicacion-geografia-y-clima.html

[4] Tobar, Alfredo Luna (1997). Historia política internacional de las Islas Galápagos. Editorial: Abya Yala.

[5] René Vallejo (2018). Quito: capitalidad y centralidades. Centro – H, ISSN: 1390-4361

[6] Edición de América Latina Guía Viajes. Quito Clima: época para viajar a Quito. 18/01/2019. Sitio Web: https://www.guiaviajes.org/quito-clima/

[7] Observatorio Social Guayaquil (2018). En 2018 Guayaquil dejaría de ser la ciudad más poblada del Ecuador. 18/01/2019. Sitio Web: https://observatoriosocial.ec/2018/03/guayaquilpoblacion/

[8] Zona Lógica (2017). El Puerto de Guayaquil: Una joya para la economía del Ecuador. 23/01/2019. Sitio Web: https://www.zonalogistica.com/el-puerto-de-guayaquil-una-joya-para-la-economia-del-ecuador/

[9] Edición de América Latina Guía Viajes. Guayaquil clima: época para viajar a Guayaquil. 23/01/2019.Sitio Web: https://www.guiaviajes.org/guayaquil-clima/#

[10] Epping-Jordan, J. E., Galea, G., Tukuitonga, C., & Beaglehole, R. (2005). Preventing chronic diseases: Taking stepwise action. Lancet, 366(9497), 1667–1671. https://doi.org/10.1016/S0140-6736(05)67342-4

[11] Baragou, S., Djibril, M., Atta, B., Damorou, F., Pio, M., & Balogou, A. (2012). Prevalence of cardiovascular risk factors in an urban area of Togo : a WHO STEPS-wise approach in Lome, Togo. Cardiovascular Journal Of Africa, 23(6), 309–312.

[12] Goryakin, Y., Rocco, L., & Suhrcke, M. (2017). The contribution of urbanization to non-communicable diseases: Evidence from 173 countries from 1980 to 2008. Economics and Human Biology, 26, 151–163. https://doi.org/10.1016/j.ehb.2017.03.004

[13] Lucio, R., Villacrés, N., & Henríquez, R. (2011). Sistema de salud de Ecuador. Salud Pública de México.

[14] Freire WB., Ramírez-Luzuriaga MJ., Belmont P., Mendieta MJ., Silva-Jaramillo MK., Romero N., Sáenz K., Piñeiros P., Gómez LF., M. R. (2014). T. I. E. N. de S. y N. de la población ecuatoriana de cero a 59 años. E.-E. 2012. M. de S. P. N. de E. y C. Q.-E. (2014). Encuesta Nacional de Salud y Nutrición (ENSANUT-2012). Primera edición.

[15] OMS. (2017). OMS | Diabetes. 27/09/2017. Sitio Web: http://www.who.int/mediacentre/factsheets/fs312/es/

[16] OMS. (2017). OMS | Diabetes. 11/10/2017. Sitio Web: http://www.who.int/mediacentre/factsheets/fs317/es/

[17] Yepez, R. F., Fuenmayor, G., Pino, A., & Yepez-Garcia, E. (1996). Enfermedades Cronicas no Trasmisibles relacionadas con la Dieta en el Ecuador. Rev. Cuba. Aliment. Nutr, 10(1), 28–34. Retrieved from http://bvs.sld.cu/revistas/ali/vol10_1_96/ali08196.htm

[18] Tripathy, J. P., Thakur, J. S., Jeet, G., & Jain, S. (2017). Prevalence and determinants of comorbid diabetes and hypertension: Evidence from non communicable disease risk factor STEPS survey, India. Diabetes and Metabolic Syndrome: Clinical Research and Reviews. https://doi.org/10.1016/j.dsx.2017.03.036

[19] Menon, J., Vijayakumar, N., Joseph, J. K., David, P. C., Menon, M. N., Mukundan, S., … Banerjee, A. (2015). Below the poverty line and non-communicable diseases in Kerala: The Epidemiology of Non-communicable Diseases in Rural Areas (ENDIRA) study. International Journal of Cardiology, 187(1), 519–524. https://doi.org/10.1016/j.ijcard.2015.04.009

[20] Tripathy, J. P., Thakur, J. S., Jeet, G., & Jain, S. (2017). Prevalence and determinants of comorbid diabetes and hypertension: Evidence from non communicable disease risk factor STEPS survey, India. Diabetes and Metabolic Syndrome: Clinical Research and Reviews. https://doi.org/10.1016/j.dsx.2017.03.036

[21] Zornetzer, H., Gresh, L., Coloma, J., Harris, E., & Monterrey, W. A. (2015). Vigilancia comunitaria: Improving communitybased infectious disease surveillance in nicaragua using a low-cost mhealth tool for data collection and decision support. American Journal of Tropical Medicine and Hygiene, 93(4 Supplement), 188. Retrieved from http://www.ajtmh.org/content/93/4_Suppl/151.full.pdf+html%5Cnhttp://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed18b&NEWS=N&AN=613369474

[22] García Pérez, C., & Alfonso Aguilar, P. (2013). Vigilancia epidemiológica en salud. Revista Archivo Médico de Camagüey, 17(6), 121–128.

[23] Escartín Lasierra, P. López Ruiz, Vicky. Ruiz-Giménez Aguilar, J. L. (2015). La participación comunitaria en salud. Pacap, 17(2), 1–8.

[24] Card, Mackinlay, & Shneiderman (1999). Readings in Information Visualization Using Vision to Think. Morgan Kaufmann. Primera Edición.

[25] Bertin Jacques (1967). Sémiologie graphique. Paris, Mouton et Gauthier- Villars. Tomo 8

[26] Colin Ware (2004). Information Visualization. Perception for Design. Morgan Kaufmann. Segunda Edición.

[27] Ben Fry (2007). Processing: A Programming Handbook for Visual Designers and Artists. The MIT Press Cambridge, Massachusetts. Londres – Inglaterra.

[28] Agrawal, Imielinski, & Swami, 1993. Mining Association Rules between Sets of Items in Large Databases. SIGMOD International Conference on Management of Data. Washington DC

[29] Wong, 2004. Induction Programs That Keep New Teachers Teaching and Improving. NASSP Bulletin

[30] Keim, Kohlhammer, Ellis, & Mansmann, 2010. Mastering the Information Age Solving Problems with Visual Analytics. Eurographics Association. Goslar, Germany

[31] Data-mining technologies for diabetes: a systematic review. 01-10-2011, PUBMED, Sitio web: https://www.ncbi.nlm.nih.gov/pubmed/22226277

[32] Espinoza Valdez, Aurora; Luna Olivera, Beatríz Carely; Solís Perales, Gualberto. (2014). ReCIBE Revista electronica de Computación, Informática Biomédica y Electrónica. 3(2), 1-12

[33] Débora Chan. Universidad de Buenos Aires (2015). Análisis de Componentes Principales [Material de clase]

[34] Débora Chan. Universidad de Buenos Aires (2015). Análisis de Correspondencias [Material de clase]

[35] Débora Chan. Universidad de Buenos Aires (2015). Análisis Inteligente de Datos [Material de clase]

[36] Carlos N. Bouza1 y Agustín Santiago (2012). Modelización Automática de Fenómenos del Medio Ambiente y la Salud, Tomo 2, pp (64-78)

[37] IBM, IBM Knowledge Center. Disponible: https://www.ibm.com/support/knowledgecenter/en/SS4QC9/com.ibm.solutions.wa_an_overview.2.0.0.doc/chaid_classification_tree.html

[38] Marcelo Soria. Universidad de Buenos Aires (2015). PAM [Material de clase]

[39] Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4: 95-104.

[40] Aníbal Goicochea. (2009). CRISP-DM, Una metodología para proyectos de Minería de Datos. 01/05/2017, Sitio web: https://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/

[41] Tamara Munzner, (2014), Visualization Analysis and Design, Boca Ratón - USA, Taylor & Francis Group. pp (135-136)

[42] Freire B. Wilma (2014), Tomo 1 Encuesta Nacional de Salud y Nutrición ENSANUT – ECU 2012, ISBN-978-9942-07-659-5, Quito – Ecuador.

[43] Salesa Barja, Pilar Arnaiz (2011), Insulinemia e índice HOMA en niños y adolescentes chilenos, Rev Med Chile 2011; 139: pp (1435-1443)

[44] Débora Chan. Universidad de Buenos Aires (2015). Análisis Inteligente de Datos [Material de clase]

[45] Stephen Redmond, (2014), Mastering Qlikview, Birmingham - Mumbai, Packt Publishing Ltd. pp (144)

[46] Freire B. Wilma (2014), Tomo 1 Encuesta Nacional de Salud y Nutrición ENSANUT – ECU 2012, ISBN-978-9942-07-659-5, Quito – Ecuador. pp (207)

[47] Débora Chan. Universidad de Buenos Aires (2015). Análisis de Componentes Principales [Material de clase]

[48] Gobierno Autónomo de la Provincia de Sucumbíos. (2019). Historia de la Provincia de Sucumbios, 17/01/2019. Sitio web:http://www.sucumbios.gob.ec/index.php/2015-10-20-00-03-09/2014-10-11-16-35-05/2014-10-11-16-54-02

[49] ENVIAJES.CL. (2016). Mapa de las Islas Galápagos: Ubicación de las Islas Encantadas, 14/10/2018, Sitio web: https://enviajes.cl/ecuador/islas-galapagos/mapa-de-galapagos/

[50] Purdue University Regional Visualization and Analytics Center, + Dept. of Nutrition, Purdue University, (2012), Applied visual analytics for exploring the National Health and Nutrition Examination Survey, IEEE Computer Society, 45th Hawaii International Conference on System Sciences

[51] Luciana María Padua (2014). Comparación Interactiva de Modelos de Minería de Datos Utilizando Técnicas de Visualización. Buenos Aires - Argentina pp (2)

[52] Worwood M. Ferritin in human tissues and serum. Clin Haematol 1982;11:275-307.

[53] Olynyk JK, Cullen DJ, Aquila S, et al. Population based study of the clinical expression of the haemochromatosis gene. N Engl J Med1999;341:718-24.

[54] European Association for the Study of the Liver (EASL). EASL clinical practice guidelines for HFE haemochromatosis. J Hepatol2010;53:3-22.

[55] Hearnshaw S, Thompson NP, McGill A. The epidemiology of hyperferritinaemia. World J Gastroenterol 2006; 12:5866-9.

[56] Ecuador Inmediato. Pobreza y consecuencias de explotación petrolera son lacerantes en Sucumbíos y Orellana. 03/02/2019. Sitio Web: http://www.ecuadorinmediato.com/index.php?module=Noticias&func=news_user_view&id=19122