VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN ENSEÑANZA DE LA MATEMÁTICA

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TITULO DE MAGISTER EN ENSEÑANZA DE LA MATEMÁTICA

TEMA: TÉCNICAS DE BIG DATA A DATOS CLÍNICOS

AUTORES: Ing. CHÁVEZ VINUEZA, JORGE LUIS

Ing. ÑAUÑAY PANCHO, JUAN MANUEL

DIRECTOR: Mat. MEDINA VÁSQUEZ, PAUL LEONARDO, Ph.D.

SANGOLQUÍ

2019

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación **TÉCNICAS DE BIG DATA A DATOS CLÍNICOS** fue realizado por las señores **Chávez Vinueza, Jorge Luis y Ñauñay Pancho, Juan Manuel** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustenten públicamente.

Sangolquí, 27 de marzo de 2019

Firma:

Ph.D. Paúl Leonardo Medina Vásquez

CC: 1712227295

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

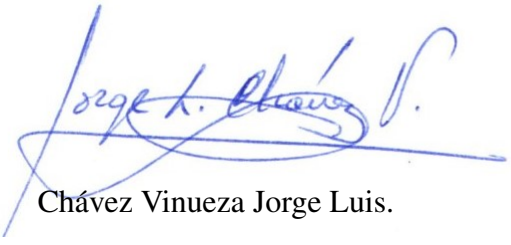
AUTORÍA DE RESPONSABILIDAD

Nosotros, **Chávez Vinueza, Jorge Luis con cédula de ciudadanía 1705912747 y Ñauñay Pancho, Juan Manuel con cédula de ciudadanía 0601631807**, declaramos que el contenido, ideas y criterios del trabajo de titulación: **TÉCNICAS DE BIG DATA A DATOS CLÍNICOS** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación es veraz.

Sangolquí, 27 de marzo de 2019

Firmas:



Chávez Vinueza Jorge Luis.

C.C. 1705912747

Ñauñay Pancho, Juan Manuel.

C.C. 0601631807

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

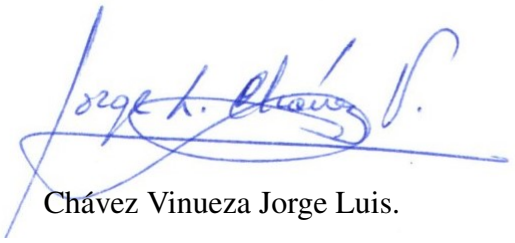
CENTRO DE POSGRADOS

AUTORIZACIÓN

Nosotros, **Chávez Vinueza Jorge Luis y Ñauñay Pancho, Juan Manuel**, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **TÉCNICAS DE BIG DATA A DATOS CLÍNICOS** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 27 de marzo de 2019

Firmas:



Chávez Vinueza Jorge Luis.

C.C. 1705912747

Ñauñay Pancho, Juan Manuel.

C.C. 0601631807

DEDICATORIA

A mi familia, por su infinita paciencia y apoyo.

Jorge

DEDICATORIA

A Dios, a mi esposa Rocío, a mis hijos Diego, Xavier y Andrés quienes con su amor, paciencia y esfuerzo han permitido culminar un sueño más en mi vida.

Manuel

AGRADECIMIENTO

A Paul Leonardo Medina Vásquez, Ph.D. por su tiempo y conocimientos compartidos en la elaboración de este proyecto. Al Ing. Carlos Almeida Ph.D. por sus consejos y observaciones. A los Doctores Gustavo Terán y Hugo Vergara Directivos del Consorcio Médico Biodilab, quienes proporcionaron la Data y brindaron sus criterios de especialistas en el área médica. A la Ing. Lilia Quituisaca por su apoyo incondicional y a todas las personas que contribuyeron a que este proyecto se haga realidad.

Jorge y Manuel

Índice

Índice de Tablas

Índice de Figuras

Índice de Algoritmos

RESUMEN

El presente trabajo de investigación desarrolla modelos matemáticos descriptivos y predictivos a partir de una data que contiene los resultados de exámenes clínicos de pacientes, la cual fue suministrada por un laboratorio de la ciudad de Quito. Este estudio se centró en el análisis de las variables que constituyen el denominado Síndrome Metabólico (Índice de Masa Corporal, Colesterol, Triglicéridos, Glucosa y Presión Arterial); además, se consideran las variables Edad, Sedentarismo y Tabaquismo. Las variables antes citadas se analizan bajo dos enfoques: el primero es la determinación de un Índice de Salud, que se basa en la categorización y posterior evaluación de las variables cuantitativas de los resultados de exámenes clínicos, de acuerdo al nivel de gravedad que es proporcionada por la literatura médica. Este índice permitirá ubicar a los pacientes en las categorías de salud Baja, Normal, Sobre y Crítica. El segundo enfoque es la determinación de factores de riesgo. En este enfoque se transforma las variables numéricas en categóricas dicotómicas de acuerdo al límite de normalidad de cada variable, es decir, dependiendo si la variable constituye o no un riesgo a la salud del paciente, para luego clasificar a la población según el número de factores de riesgo; esta clasificación permite determinar grupos poblacionales con diferentes niveles de riesgo. Adicionalmente, se aplica la técnica del escalado multidimensional, la cual permite visualizar la influencia de los diversos factores de riesgo en la salud del paciente.

PALABRAS CLAVE:

- **BIG DATA**
- **ANÁLISIS MULTIVARIANTE**
- **SÍNDROME METABÓLICO**
- **ESCALADO MULTIDIMENSIONAL**

ABSTRACT

This research develops descriptive and predictive mathematical models about the results of clinical exams of patients from a big data base. This information was provided by a clinical laboratory located in Quito city. This study has focused on the analysis of the variables which constitute the so-called Metabolic Syndrome: body mass index, cholesterol, triglycerides, glucose and blood pressure; in addition, other variables as age, sedentary, and smoking are considered. The variables mentioned above are analyzed through two approaches: the first one is the determination of a health index based on the categorization and subsequent evaluation of the quantitative variables from the clinical examination results according to the level of severity provided by the medical literature. This index will allow locating patients in the categories of 'low normal', 'normal', 'over normal', and 'critical'. The second one is the determination of risk factors. In this approach, the numerical variables are transformed to dichotomous categorical variables according to the normal limit of each one. In consequence, it depends on whether the variable constitutes a risk to the patient's health or not, and then classifies the population according to the number of risk factors. This classification allows determining population groups with different risk levels. Additionally, it is possible to visualize the influence of the various risk factors on the patient's health by applying a technique known as multidimensional scaling.

KEY WORDS:

- **BIG DATA**
- **MULTIVARIATE ANALYSIS**
- **METABOLIC SYNDROME**
- **MULTIDIMENSIONAL SCALING**

Capítulo 1

1. INTRODUCCIÓN

Actualmente las clínicas, hospitales y laboratorios tienen grandes cantidades de datos clínicos, muchos de ellos en formatos físicos o electrónicos; sin embargo, esta información en la mayoría de los casos permanece sin utilizar, pues los formatos son incompatibles o la información no está estandarizada.

Al no poder utilizar la información generada en forma sistemática, se pierde la oportunidad de conocer patrones de comportamiento de grupos humanos con características similares de educación, actividad económica, género, etc., y; por tanto, también la oportunidad predecir y prevenir patologías que se puedan presentar. Un adecuado uso de la información presente en las bases de datos médicas se puede traducir en la implementacón de una medicina preventiva.

La gran cantidad de información producida por los sistemas de salud y la imposibilidad de trabajarla con técnicas convencionales ha logrado posicionar el Big Data en el tratamiento y análisis de esta información. En el Ecuador, específicamente en el área de la salud, no existe información sobre la aplicación de la técnica de Big Data a datos clínicos. En ciertos países europeos la utilización de plataformas tecnológicas ha entrado con fuerza en los últimos años en el ámbito de la salud, tanto en la prestación de servicios sanitarios como en su predicción [Amabili, 2016].

Considerando que las aplicaciones de Big Data en el sector de la salud van en aumento, la presente investigación pretende desarrollar un modelo o modelos matemáticos descriptivos y/o predictivos que faciliten la toma de decisiones médicas. Así, la pregunta que pretendemos responder en la presente investigación es: ¿es posible obtener información relevante, utilizando técnicas de Big Data, en la base de datos clínicos suministrados? Para responderla, analizando los resultados de exámenes clínicos de pacientes, se utilizan las siguientes técnicas:

- Análisis descriptivo de las variables seleccionadas,
- Análisis inferencial y predictivo con:
    - Técnicas convencionales y

- Técnicas contemporaneas de Big Data.

El objetivo general será desarrollar un algoritmo o modelo matemático que brinde información para describir y predecir el comportamiento de determinadas patologías clínicas, con la intensión de poder establecer alertas tempranas que permitan prevenir el desarrollo de las mismas.

Los objetivos específicos serán:

- Analizar la calidad de la data clínica proporcionada,
- Combinar variables para clasificar los grupos de pacientes de acuerdo a ciertas características médicas,
- Describir y predecir a partir de los datos clínicos de una persona, su estado de salud, y
- Desarrollar un algoritmo matemático para transformar los datos en información.

Este trabajo se ha dividido en las siguientes partes: en el Capítulo 1, se presenta el alcance del trabajo, se establecen las interrogantes que conducen esta investigación y los enfoques considerados para responderla; en el Capítulo 2, se habla sobre el Big Data, técnicas de análisis multivariante, los perfiles médicos y las actividades económicas relevantes en la data utilizada; el Capítulo 3, describe la metodología utilizada en el tratamiento de la data; en el Capítulo 4, se hace un análisis y discusión de los resultados obtenidos de acuerdo a los objetivos y preguntas de investigación planteadas y; finalmente, en el Capítulo 5 se presentan las conclusiones y recomendaciones.

Capítulo 2

2. ESTADO DEL ARTE

En este Capítulo se analiza los siguientes temas: Big Data, Análisis Multivariante, Síndrome Metabólico y Actividades Económicas. En Big Data se revisa las caracteristicas y los beneficios de esta técnica, luego se analiza su importancia en el campo de la salud, finalmente se menciona los diferentes programas computacionales que permiten procesar y obtener información a través del Big Data. En el Análisis Multivariante se recopila información de los métodos que se aplican en el presente estudio, estos son, componentes principales, distancias estadísticas, escalado multidimensional y análisis de conglomerados. En el Síndrome Metabólico se analiza y justifica la selección de variables que se utiliza para el presente estudio. Las Actividades Económicas se refieren a la identificación de los sectores de la economía donde se ubican las empresa en las que laboran los pacientes, cuyos datos se encuentran en la data.

2.1. Big Data

En su investigación [Garside and Cox, 2013] comentan que en la actualidad existe gran cantidad de información, cada segundo las computadoras y los sistemas electrónicos crean, procesan, trasmiten y reciben enormes volúmenes de información. Los datos pueden ser almacenados de muchas formas como: sonidos, imágenes, videos, códigos de barras, transacciones comerciales, etc. Independientemente de la fuente los datos normalmente se dividen en dos tipos, estructurados y no estructurados.

Los datos estructurados es información que está altamente organizada dentro de una tabla, donde las computadoras pueden fácilmente manipularlos y organizarlos basándose en diversos criterios. La información contenida en este tipo de datos no es fácilmente apreciada a simple vista. Por ejemplo, un código de barras es irreconocible por el ojo humano pero es altamente estructurado y fácil de leer por las computadoras.

Los datos no estructurados es información que normalmente no tiene un modelo predefinido o no encaja dentro de tablas ordenadas u hojas de cálculo, estos suelen tener un gran volumen de

texto y puede contener datos como fechas y números. Las imágenes, videos y audios se consideran no estructurados, su falta de estructura hace que la compilación sea una tarea que consume tiempo y energía para un sistema informático. Los datos no estructurados son fáciles de entender por los humanos.

Los datos adicionalmente pueden considerarse como cualitativos si presentan información descriptiva, que puede ser a menudo subjetiva, o como datos cuantitativos si presentan información numérica, que a la vez puede ser cuantitativo discreto si se lo puede contar o cuantitativo continuo si se lo puede medir.

Los datos son esencialmente un conjunto de números y caracteres, y serán útiles una vez que se procesan y coloquen en un contexto, esta información transformada en instrucciones permitirá aumentar la efectividad y eficiencia en un proceso de investigación. Por tanto, el procesamiento de datos es el primer paso para crear conocimiento. Ante la generación de gran cantidad de datos en determinadas áreas se han creado estrategias para su comprensión, manipulación y extracción de información conocidas como Big Data.

Big Data es una combinación de tecnologías antiguas y nuevas que ayudan a obtener una perspectiva práctica. Es la capacidad de administrar un gran volumen de datos en diversos formatos, a la velocidad correcta y dentro de un marco de tiempo adecuado para permitir el análisis y la reacción en tiempo real [Hurwitz et al., 2013] y se caracteriza por: volumen, variedad, velocidad, veracidad y valor.

Volumen implica la cantidad de datos generados, en cada institución, a través de diversas fuentes, los datos recolectados son fácilmente del orden de los terabytes [1] o incluso de los petabytes [2] , siendo esta la característica más importante y distintiva del Big Data, ya que determina las técnicas y herramientas ha utilizarse [Jasim et al., 2015].

Variedad implica que los datos se obtendrán en diversos formatos, los que pueden ser estructurados, como los generados mediante tablas de datos o no estructurados como imágenes, audios, videos, e-mails, respuestas de sensores remotos, etc.[Jasim et al., 2015].

Velocidad se refiere a la necesidad de procesar y manipular los datos existentes en tiempos relativamente cortos, para obtener información relevante, útil, aplicable y oportuna. Para procesos

[1] 1 terabyte = 1 024 gigabytes = 2^{10} gigabytes = 2^{20} megabytes = 2^{30}kilobytes = 2^{40}bytes
[2] 1 petabyte = 1 024 terabytes = 2^{50}bytes

sensibles se requiere que los datos se analicen y utilicen a medida de que se generen, lo que maximiza el valor de la información [Jain, 2016].

Veracidad es el grado de confianza de la información obtenida de la data para la toma de decisiones, lo que implica que tanto la captura de los datos como el tratamiento de los mismos deben generar conclusiones verdaderas y útiles. Es necesario identificar la cantidad correcta y tipos de datos que se pueden analizar para lograr resultados impactantes[Cano, 2014].

Valor es la característica que se define por el valor agregado que los datos y la información generada pueden aportar al proceso o actividad que realiza la organización. Tener una cantidad infinita de datos puede resultar inútil si la información no se puede monetizar; las tecnologías aplicadas a grandes volúmenes de datos deben generar soluciones que permiten transformar la forma en que se manejan los negocios, lo que se traduce que tanto la data como los resultados que se obtengan de ella representan un valor económico para la organización [Hurwitz et al., 2013].

En términos generales, Big Data se refiere al manejo de un gran volumen de datos, estructurados como no estructurados. Sin embargo, más que la cantidad, su importancia radica en la información que se obtiene mediante la utilización de nuevas técnicas y algoritmos, y lo que se pueda hacer con esta información [Urueña et al., 2012].

2.1.1. El Big Data en el campo de la salud

En el campo de la salud la aplicación de Big Data podría, entre otras cosas: prevenir enfermedades antes de que se diagnostiquen; tener un seguimiento de nuestras enfermedades en tiempo real sin necesidad de acudir a la consulta del médico; tener acceso a un mapa de salud para que sea analizado por especialistas, etc. [Zaforas, 2016].

El Big Data se consolida a partir de la información antigua y nueva incorporando tanto las provenientes de las redes sociales como aquellas derivadas del Internet. Optimizar esta información permite tener un mayor conocimiento del paciente gracias a la información existente de historias médicas, registros electrónicos de salud, registros personales de salud y análisis e imágenes clínicas. También, genera nueva información para detectar efectos secundarios de los fármacos, tratamientos más adecuados y con mayor rapidez, así como avanzar en la medicina personalizada y preventiva [Urueña et al., 2012].

Uno de los conceptos que cobra importancia en la relación entre salud y el Big Data es lo que se denomina “datos reales mundiales” y que se refiere a como éste se nutre de información obtenida en condiciones reales; por ejemplo, en la Agencia de Medicamentos de Estados Unidos, ha detectado mediante la aplicación de algoritmos a grandes bases de datos, a veces no estructurados y, procedentes del mundo real nuevas interacciones, efectos adversos de medicamentos y otros problemas de seguridad que han llevado al retiro del fármacos o la modificación de sus indicaciones [Urueña et al., 2012].

Los profesionales de la salud cada vez entienden mejor que Big Data supone un cambio de paradigma en la práctica de la medicina. Las empresas farmacéuticas quieren utilizarlo para diseñar medicamentos cada vez más efectivos y con menor costo de investigación. Por su parte, las administraciones quieren comprobar la eficacia de los nuevos medicamentos gracias a la información proveniente de la obtenida en el mundo real. Pero también este nuevo concepto complementa a que los ensayos clínicos realizados por los laboratorios aseguren la eficacia de un nuevo fármaco. Por ejemplo, los profesionales sanitarios pueden utilizar la analítica del Big Data en tiempo real para saber donde se está extendiendo un virus de la gripe y a que rapidez, con lo cual pueden dar una respuesta oportuna, garantizando el stock de vacunas [Joyanes and Poyatos, 2013] .

En una entrevista que realizó la revista EFE salud, al investigador y neurólogo Ignacio Hernández Medrano, manifiesta entre otras cosas que: “Somos capaces de anticipar con algoritmos si un paciente va a volver por urgencias o no, también detectar cuál es la pauta óptima para un paciente que tiene anemia o insuficiencia renal”. Afirma este neurólogo que se trata de una evolución, más que una revolución, porque la medicina siempre progresa, nunca ha dejado de progresar y tecnológicamente un día apareció el fonendo, que fue muy bueno para saber qué pasaba dentro del cuerpo, luego llegó la resonancia magnética o el escáner que nos permitió mirar también dentro del cuerpo y ahora es el tiempo de los algoritmos matemáticos aplicados a la medicina [Gonzáles, 2017].

Los algoritmos, manifiesta Hernández, son capaces de afinar a nivel individual que es lo mejor para cada paciente lo que permiten hacer una medicina personalizada, aplicada de forma individual o a un grupo pequeño de personas, al tiempo que facilitan la medicina predictiva, “porque lo bueno de los algoritmos es que cuando tienes un buen histórico de datos, podemos predecir, como en meteorología”.

En este contexto la aplicación de Big Data es cada vez más evidente en el mundo de la salud y la medicina. Se utilizará para predecir, prevenir y personalizar enfermedades, es probable que se

amplíe a otros campos de la salud, en particular se puede citar: la investigación y secuenciación del genoma; investigación clínica; epidemiología; el uso de los wearables; medicina personalizada; farmacología, etc. [Urueña et al., 2012].

Un ejemplo importante es la **genómica**, campo de la biología molecular que estudia el material genético de los organismos, incluye la secuenciación, mapeo y análisis de los códigos ADN [3] y ARN [4] para comprender cómo funcionan los genes y qué impacto tienen en las enfermedades. La cantidad de datos que se generan en este proceso son enormes. El genoma de una persona tiene entre 20 000 y 25 000 genes y son necesarios cientos de petabytes para almacenar esta información. El proyecto que consiguió secuenciar y analizar el primer genoma humano tardó 10 años con un costo cercano a los 3 000 millones de dólares (Estados Unidos. Department of Energy. Human Genome Project. 1990), este costo se ha reducido y ahora se puede secuenciar y analizar un genoma completo en pocas horas por menos de 1 000 dólares [Urueña et al., 2012].

Entre las principales aplicaciones del estudio del genoma tenemos: el uso de modelos predictivos para identificar pacientes de alto riesgo (diabetes de tipo 1); clasificación de enfermedades en subtipos para seleccionar el tratamiento clínico dirigido (cáncer); y proveer de información para la elaboración de fármacos [Zaforas, 2016].

En **investigación clínica** el Big Data permite a los profesionales sanitarios ofrecer diagnósticos más rápidos y precisos, con un mejor respaldo desde la perspectiva científica [Urueña et al., 2012].

En **epidemiología** el Big Data ofrece información para el análisis y control de epidemias, pues permite, predecir la propagación de un virus, geolocalizarlos y realizar un mejor seguimiento a fin de definir las áreas y los centros de tratamiento contra la enfermedad o restringir los movimientos poblacionales si fuera el caso [Urueña et al., 2012].

Los **wearles** son dispositivos digitales que se usa para tomar medidas de datos biométricos de los pacientes durante todo el día. Esta información es analizada en tiempo real por los médicos y especialistas, realizando un seguimiento como nunca visto hasta ahora. Por ejemplo se pueden monitorear parámetros como la saturación de oxígeno, el pulso cardiaco, las calorías quemadas o realizar un seguimiento de la calidad del sueño [Zaforas, 2016].

[3] El ADN es el ácido desoxirribonucleico que contiene toda la información genética hereditaria en los seres vivos, consiste en dos cadenas que se enrollan entre ellas y forman una estructura de doble hélice. Cada cadena contiene azúcares y grupos fosfatos [Curtis et al., 2008]

[4] El ARN es el ácido ribonucleico que ayuda al ADN en las funciones de transmisión de genes y de síntesis de proteínas vitales para el desarrollo de las características y funciones grabadas en el ADN [Curtis et al., 2008]

Esta forma de análisis ha permitido una nueva era para la ciencia de los datos. Gracias al gran volumen de información se puede aplicar técnicas de inteligencia artificial, como el diseño de nuevos algoritmos de aprendizaje automático (machine learning algorithms) para su predicción y prevención de enfermedades. Esta nueva revolución podría afectar incluso el rol de los médicos que tendrán que apoyarse en expertos en análisis de datos. O quién sabe la forma de diagnosticar y prevenir enfermedades evolucionará radicalmente [Zaforas, 2016].

2.1.2. Programas que trabajan con Big Data

Existen varios programas que permiten procesar y obtener información a través del Big Data; por ejemplo, programas comerciales como MATLAB y SPSS, y programas de software libre como OCTAVE, R, Python, Julia y Tableau.

- **Matlab**

Es un lenguaje de computación técnico de alto nivel y un entorno interactivo para el desarrollo de algoritmos, visualización de datos, análisis de datos y cálculo numérico. Matlab cuenta con una amplia gama de aplicaciones que incluyen procesamiento de señales e imágenes, comunicaciones, diseño de sistemas de control, sistemas de prueba y medición, modelado y análisis estadísticos y biología computacional. Para el análisis estadístico, Matlab posee el Toolbox Statistics que proporciona un conjunto completo de herramientas para evaluar e interpretar Big Data [Pateiro, 2016].

- **SPSS**

Proporciona poderosas herramientas para el tratamiento estadístico de Big Data, ésta plataforma ha desarrollado soportes en todas las fases del proceso de minería de datos, que incluye el desarrollo, implementación y actualización del modelo. SPSS genera activos analíticos, el término genérico activo analítico, se utiliza para describir una colección de operaciones que resuelve un problema comercial. Los científicos de datos a menudo usan los términos modelo o modelo predictivo cuando describen activos desarrollados en minerías de datos. Un activo analítico de SPSS puede incluir pasos de preparación de datos [Pateiro, 2016].

- **R**

Es un entorno de software libre para computación y gráficos estadísticos. El lenguaje R es ampliamente utilizado entre estadísticos y mineros de datos para el desarrollo de software estadístico y análisis de datos; su popularidad ha aumentado en los últimos años. Se compila y se ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOs. R posibilita cargar librerías o paquetes con finalidades específicas de cálculo [Pateiro, 2016].

- **Python**

Es un lenguaje de alto nivel. Por su naturaleza interactiva, es una buena opción para el desarrollo algorítmico y el análisis de datos exploratorios. Se utiliza cada vez más no solo en entornos académicos sino también en la industria [Drake, 2009].

- **Julia**

Es un lenguaje de alto nivel. Al igual que Python, por su naturaleza interactiva, es una buena opción para el desarrollo algorítmico y el análisis de datos exploratorios. Se usa no solo en entornos académicos sino también en la industria [Goette, 2014].

- **Tableau**

Tableau es una herramienta gratuita de visualizaciones interactivas de datos. Tiene su origen en una investigación del Departamento de Informática de la Universidad de Stanford. La empresa fue creada en el 2003, pero recién en 2010 se creó la versión gratuita Tableau Public. Hoy Tableau Public es la principal herramienta que nos permite crear visualización de datos de manera rápida y sencilla a partir de plantillas de Excel. Los datos son públicos, se los puede descargar desde las mismas visualizaciones y estas a su vez se las puede socializar. Trabaja con una gran variedad de gráficos: barras, barras apiladas, tortas, tablas, mapas con polígonos, líneas, puntos, etc. [Trigo and López, 2012].

- **¿Por qué el uso de Matlab para el presente análisis?**

En el presente trabajo se utilizará Matlab por ser un lenguaje potente y amistoso, que permite realizar nuestras propias rutinas de programación en las cuales se pueden incluir poderosos comandos para el análisis estadístico multivariante.

Entre las principales herramientas que posee Matlab tenemos la Statistics and Machine Learning Toolbox, la cual proporciona funciones y aplicaciones para describir, analizar y modelar datos. Puede usar estadísticas descriptivas y gráficos para el análisis de datos exploratorios, ajustar las distribuciones de probabilidad a los datos y realizar pruebas de hipótesis. Los algoritmos de regresión y clasificación le permiten sacar inferencias de los datos y construir modelos predictivos. Para el análisis de datos multidimensionales, Statistics and Machine Learning Toolbox proporciona selección de características, regresión por pasos, análisis de componentes principales (ACP), regularización y otros métodos de reducción de dimensionalidad que le permiten identificar variables o características que afecten su modelo. Este paquete proporciona algoritmos de aprendizaje automático supervisados y no supervisados, incluyendo máquinas de vectores de soporte (MVS), árboles de decisión potenciados y empaquetados, vecino más cercano a k, k-medias, k-medoides, agrupamiento jerárquico, y modelos de mezcla gaussianos. Muchas de las estadísticas y algoritmos de aprendizaje automático se pueden usar para cálculos en conjunto de datos que son demasiado grandes para almacenarse en la memoria [MathWorks, 2017].

2.2. Análisis multivariante

Para describir las características de una persona o grupo de personas necesitamos analizar diferentes variables al mismo tiempo, así por ejemplo para describir las características físicas de una persona podemos utilizar variables como la estatura, el peso, el diámetro de cintura, etc.; para analizar el perfil lipídico [5] utilizaremos variables como, el Colesterol, Triglicéridos, Glucosa e Índice de Masa Corporal.

Las técnicas que nos brinda el análisis multivariante nos permite, entre otras cosas, determinar cuál o cuáles de las variables analizadas son las de mayor influencia, y reducir la dimensionalidad que caracteriza a los datos de acuerdo a sus similitudes (cercania) o no.

[5] También conocido como “panel de lípidos", mide las concentraciones de distintos tipos de grasas en la sangre [Múnera and Escobar, 2007]

Los métodos que se aplicarán en el análisis multivariante de datos son: estadísticos descriptivos, componentes principales, distancias estadísticas, escalado multidimensional y análisis de conglomerados.

2.2.1. Estadísticos descriptivos

El análisis estadístico descriptivo permite comprender la estructura de los datos y extraer la información que estos contienen, lo que ayuda a caracterizar a la población de estudio y formar grupos específicos. Sobre los grupos establecidos se determina parámetros estadísticos como media, mediana y varianza.

2.2.2. Análisis de componentes principales

El análisis de componentes principales consiste en disminuir el número de variables presentes en los datos a cambio de una pequeña pérdida de la información presente. Las nuevas variables se representan como una combinación lineal de las variables originales, si las variables originales tienen alta dependencia un pequeño número de nuevas variables explican la mayor parte de la variabilidad original. Aunque se necesitan el mismo número de componentes principales que las p variables originales, para reproducir toda la variabilidad del sistema, generalmente la mayor parte de esta variabilidad es explicada por un número pequeño de k componentes principales; en estos casos las k componentes principales reemplazan a las p variables originales. Con frecuencia este análisis revela relaciones que no se sospechaba inicialmente, lo que permite interpretaciones de los datos que no podrían ser derivadas a partir de las variables originales [Castaño, 2012].

2.2.3. Distancias estadísticas

La Distancia Estadística entre objetos o individuos está presente en casi todas las técnicas del análisis multivariado, ésta es una interpretación geométrica de objetos como puntos de un espacio métrico adecuado. Esta interpretación es válida no solo para variables cuantitativas, sino también, cuando las variables observadas son de tipo más general, siempre que tenga sentido obtener una medida de proximidad entre los objetos o individuos [Cuadras, 2014]. De manera particular tenemos la siguiente formalización

Definición 1. *Sea E un conjunto finito con n elementos diferentes, $\mathbb{R}$ el conjunto de los números reales, una disimilaridad es una aplicación $\delta : E \times E \longrightarrow \mathbb{R}$, que cumple:*

- $\delta_{ij} = \delta_{ji}$ *para todo* i, j *elementos de* E *y,*
- $\delta_{ii} = 0$ *para todo* i *elemento de* E*,*

Si la aplicación δ, adicionalmente, cumple la desigualdad triangular.

$\delta_{ij} \leq \delta_{ik} + \delta_{kj}$ *para todo* i, j, k *elementos de* E*, se habla de una distancia [Cuadras, 2014].*

Las definiciones que se muestran a continuación son tomadas de [Grané, 2008].

Distancias para variables cuantitativas que dependen de la escala de datos

- Distancia de Minkowski (δ_{mq})

$$\delta_{mq}(i,j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right)^{\frac{1}{q}},$$

donde q es un entero positivo, p el número de variables y x_{ij} es el elemento i, j de la matriz de datos. Aquí se presupone que las variables son incorrelacionadas y de varianza uno. En base a esta distancia se presenta los siguientes casos particulares:

- Distancia de Manhattan o City Block $(\delta_{CB} = \delta_{m1})$

$$\delta_{CB}(i,j) = \delta_{m1}(i,j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|.$$

- Distancia Euclídea $(\delta_E = \delta_{m2})$

$$\delta_E(i,j) = \delta_{m2}(i,j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^2 \right)^{\frac{1}{2}}.$$

- Distancia de Tchebychev o del máximo $(\delta_T = \delta_{m\infty})$

$$\delta_T(i,j) = \delta_{m\infty}(i,j) = max\{|x_{i1} - x_{j1}|, ..., |x_{ip} - x_{jp}|\}.$$

Las distancias antes mencionadas, son muy sensibles a las unidades en las que están medidas. Lo indicado se puede visualizar con un ejemplo.

Tabla 1
Datos para un ejemplo de cálculo de distancias.

Persona	Estatura (m)	Edad (años)
1	1.85	18
2	1.78	18
3	1.65	20

Con los datos de la Tabla 1 se calculará para ejemplificar la distancia Euclídea entre los individuos 1 y 2, y entre los individuos 1 y 3, considerando las variables Estatura y Edad. A fin de evidenciar la influencia que tiene el cambio de escala al realizar los cálculos, la variable Estatura será tanto en metros como en centímetros, obteniendose:

- distancias obtenidas utilizando la variable estatura en metros

  $\delta_{1-2} = \sqrt{(1.78 - 1.85)^2 + (18 - 18)^2} = 0.07$ y,

  $\delta_{1-3} = \sqrt{(1.65 - 1.85)^2 + (20 - 18)^2} = 2.01;$

- distancias obtenidas utilizando la variable estatura en centímetros

  $\delta_{1-2} = \sqrt{(178 - 185)^2 + (18 - 18)^2} = 7.00$ y,

  $\delta_{1-3} = \sqrt{(165 - 185)^2 + (20 - 18)^2} = 20.10.$

Los valores de distancia calculados en metros dan la sensación que el individuo 1 es más parecido o más cercano al individuo 2 que al 3; sin embargo, al cambiar la escala los valores de distancia entre estos individuos aumenta, dando la sensación que la distancia es mayor, cambio que no es en la misma proporción en los dos casos. Esto muestra la sensibilidad de estas distancias a la escala de las variables. Por lo tanto, es importante tener presente las unidades de las variables que intervienen en el cálculo de este tipo de distancias. Una alternativa para evitar la influencia de las unidades, de las variables, en los cáculos de las distancias es estandarizar los datos de todas las variables a media cero y desviación estandar uno.

Distancias para variables cuantitativas invariantes frente a cambios de escala

- Distancia de Canberra (δ_{Can}), es una modificación de la distancia Manhattan,

$$\delta_{Can}(i,j) = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}.$$

- Distancia de Karl Pearson δ_k, es una modificación de la distancia Euclídea, depende de la varianza s_k^2 de la variable X_k,

$$\delta_k(i,j) = \left(\sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|^2}{s_k^2} \right)^{\frac{1}{2}}.$$

- Distancia de Mahalanobis δ_M, es adecuada como medida de discrepancia entre los datos ya que considera las correlaciones entre las variables y es invariante frente a transformaciones lineales no singulares de las variables,

$$\delta_M(i,j) = \left((X_i - X_j) S^{-1} (X_i - X_j)' \right)^{\frac{1}{2}},$$

donde S es la matriz de varianzas y covarianzas, de la matriz de datos, con dimensión $p \times p$; y, X_i es la fila i de la misma matriz, con dimensión $1 \times p$.

Consideremos los datos en la Tabla 1, con estos, se calcula para ejemplificar la distancia de Karl Pearson entre los individuos 1 y 2, y entre los individuos 1 y 3, considerando las variables Estatura y Edad; con la variable Estatura tanto en metros como en centímetros, obteniendose:

- distancias obtenidas utilizando la variable estatura en metros:
  varianza de la variable estatura = 0.006866,
  varianza de la variable edad = 0.888888,
  distancias entre los individuos:

$$\delta_{1-2} = \sqrt{\frac{(1.78 - 1.85)^2}{0.006866} + \frac{(18 - 18)^2}{0.888888}} = 0.8447 \quad \text{y,}$$
$$\delta_{1-3} = \sqrt{\frac{(1.65 - 1.85)^2}{0.006866} + \frac{(20 - 18)^2}{0.888888}} = 3.2133;$$

- distancias obtenidas utilizando la variable estatura en centímetros:
  varianza de la variable estatura =68.66,
  varianza de la variable edad = 0.8888,
  distancias entre los individuos:

$$\delta_{1-2} = \sqrt{\frac{(178-185)^2}{68.66} + \frac{(18-18)^2}{0.8888}} = 0.8447 \quad \text{y,}$$
$$\delta_{1-3} = \sqrt{\frac{(165-185)^2}{68.66} + \frac{(20-18)^2}{0.8888}} = 3.2133.$$

Los valores de distancia calculados ratifican que el individuo 1 es más parecido o más cercano al individuo 2 que al 3, y al cambiar la escala de la variable estatura, los valores de distancia entre estos individuos no cambia, en los dos casos.

Distancias para variables dicotómicas

Según Cuadras [Cuadras, 2014], cuando se trata de variables dicotómicas es necesario definir, en primer lugar, el concepto de similaridad, pues a partir del mismo se pueden definir distancias entre variables cualitativas.

Definición 2. *Una similaridad es una aplicación* $S : E \times E \longrightarrow \mathbb{R}$*, que cumple:*

- $S_{ij} = S_{ji}$ *para todo* i, j *elementos de* E *y,*
- $0 \leqslant S_{ij} \leqslant S_{ii} = 1$ *para todo* i, j *elementos de* E*.*

La distancia entre los individuos i, j, para las variables dicotómicas, se obtiene a partir de los índices de similaridad mediante la relación:

$$\delta_{ij} = S_{ii} + S_{jj} - 2S_{ij}.$$

Los coeficientes de similaridad entre los individuos i, j, se calculan de acuerdo a las frecuencias a, b, c y d, donde:

- a es el número de variables con respuesta 1 en ambos individuos,
- b es el número de variables con respuesta 0 en el individuo i y 1 en el individuo j,

- c es el número de variables con respuesta 1 en el individuo i y 0 en el individuo j y,
- d es el número de variables con respuesta 0 en ambos individuos.

El número total de variables es p, tal que

$$p = a + b + c + d.$$

De acuerdo a [Baroni and Buser, 1976], existen muchos coeficientes de similaridad, entre los que se mensiona:

- Jaccard

  $$S_{ij} = \frac{a}{a+b+c},$$

  el cual considera solo la presencia de la variable entre los individuos i, j. No pondera ninguna frecuencia.

- Dice

  $$S_{ij} = \frac{2a}{2a+b+c},$$

  considera solo la precencia de la variable entre los individuos i, j. Se pondera la presencia de la variable en los dos individuos.

- Ochiai

  $$S_{ij} = \frac{a}{\sqrt{(a+b)(a+c)}},$$

  considera solo la precencia de la variable entre los individuos i, j. El cociente representa la media geométrica entre la precencia de la variable en el individuo i y en el individuo j.

- Russell-Rao

  $$S_{ij} = \frac{a}{a+b+c+d} = \frac{a}{p},$$

  considera la presencia común de la variable sobre el total de las variables (ya se incluye la ausencia común de la variable).

- Sokal-Michener

  $$S_{ij} = \frac{a+d}{a+b+c+d} = \frac{a+d}{p},$$

  considera las presencia y ausencias comunes de la variable sobre el total de las variables.

- Rogers-Tanimoto

$$S_{ij} = \frac{a+d}{a+d+2(b+c)},$$

considera las presencia y ausencias comunes de la variable sobre el total de las variables, ponderando las frecuencias b y c.

Un inconveniente en la interpretación de estos coeficientes es su dependencia de p, sobre todo, para valores pequeños de p, pues la probabilidad de una estimación adecuada es baja, aumentandose a medida que el valor de p se incrementa.

Si se considera que la data contiene la totalidad de las variables de comparación entre los individuos, será recomendable utilizar un coeficiente que ponderé las ausencias comunes como los de Sokal-Michener o de Rogers-Tanimoto; si por el contrario, se estima que aún quedan variables no consideradas en la data, se debera usar índices que solo consideren las frecuencias a, b y c, como los coeficientes de Jaccard, Dice o Ochiai [Saiz, 1980].

En nuestro estudio, para la determinación de la matriz de distancias, dado que se considera todas las variables del perfil médico definido, se utilizará el índice de similaridad de Sokal-Michener, ya que permite obtener una matriz de distancias Euclídeas.

2.2.4. Escalado multidimensional

Esta técnica permite, a partir de una matriz D, cuadrada, de dimensión $n \times n$, que representa las distancias o las disimilaridades entre los n elementos de un conjunto dado, obtener una matriz que contiene un conjunto de variables ortogonales $Y_1, ..., Y_p$ donde $p < n$, de manera que las distancias Euclídeas entre elementos, en estas nuevas variables, sean iguales o lo más próximas posibles a las distancias de la matriz original D. Es decir, a partir de la matriz D, se pretende obtener una matriz X, de dimensiones $n \times p$, que pueda interpretarse como la matriz de p variables en los n individuos, y donde la distancia Euclídea entre los elementos reproduzca, aproximadamente, la matriz de distancias D inicial. Cuando el valor de $p > 2$, las variables pueden ordenarse de acuerdo a su importancia, y se pueden realizar representaciones gráficas en dos o tres dimensiones para entender la estructura existente en el grupo de estudio [Peña, 2002].

El escalado multidimensional permite describir e interpretar los datos, la representación de los elementos por unas pocas variables faculta entender que elementos tienen propiedades similares,

si aparecen grupos de elementos con características similares o si existen elementos atípicos. Si se puede interpretar las variables obtenidas se tendrá un conocimiento más relevante del grupo en estudio [Peña, 2002].

El problema consiste en construir una matriz X, de dimensión $n \times p$, en la cual las variables tengan media cero, a partir de una matriz de distancias al cuadrado D de dimensión $n \times n$, con elementos δ_{ij}^2. En el proceso se construye una matriz simétrica Q, de dimensión $n \times n$, y a partir de esta, se genera la matriz X.

- **Obtensión de la matriz Q a partir de la matriz de distancias D**

Dado que se considera que en la matriz X, todas sus variables tienen media cero, el producto de matrices $X' \times \mathbf{1} = \mathbf{0}$, donde $\mathbf{1}$ es una matriz de unos de dimensión $n \times 1$ y $\mathbf{0}$ es una matriz nula de dimensión $n \times 1$. También se considera $Q \times \mathbf{1} = \mathbf{0}$, es decir, la suma de todos los elementos de una fila o columna de la matriz de similitudes Q debe ser cero.

Considerando la relación entre distancia y similitud se tiene que

$$\delta_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}; \tag{1}$$

en particular, la suma de la fila i en la matriz D genera

$$\sum_{i=1}^{n} \delta_{ij}^2 = \sum_{i=1}^{n} q_{ii} + nq_{jj} = t + nq_{jj}, \tag{2}$$

donde por la condición impuesta $Q \times \mathbf{1} = \mathbf{0}$, se tiene que

$$t = \sum_{i=1}^{n} q_{ii} = traza(Q);$$

luego la suma de los elementos de la fila i es cero

$$\sum_{i=1}^{n} q_{ij} = 0,$$

dado que q_{jj} no depende del indice i, se tiene que

$$\sum_{i=1}^{n} q_{jj} = q_{jj} \sum_{i=1}^{n} 1 = nq_{jj};$$

despejando q_{jj} de (2) se tiene que

$$q_{jj} = \frac{1}{n}\sum_{i=1}^{n}\delta_{ij}^2 - \frac{t}{n}. \tag{3}$$

De manera similar, la suma de la columna j de la matriz D genera

$$\sum_{j=1}^{n}\delta_{ij}^2 = t + nq_{ii}.$$

Esto permite obtener una expresión para q_{ii}, tal que

$$q_{ii} = \frac{1}{n}\sum_{j=1}^{n}\delta_{ij}^2 - \frac{t}{n}. \tag{4}$$

Por similar análisis, sumando todos los elementos de la matriz D se tiene que

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}^2 = 2nt.$$

Ahora por (1), (3) y (4) se tiene que

$$\delta_{ij}^2 = \frac{1}{n}\sum_{j=1}^{n}\delta_{ij}^2 - \frac{t}{n} + \frac{1}{n}\sum_{i=1}^{n}\delta_{ij}^2 - \frac{t}{n} - 2q_{ij}, \tag{5}$$

llamando a la media de la fila i como δ_{i*}^2, se tiene que

$$\delta_{i*}^2 = \frac{1}{n}\sum_{j=1}^{n}\delta_{ij}^2, \tag{6}$$

a la media de la columna j como δ_{*j}^2, se tiene que

$$\delta_{*j}^2 = \frac{1}{n}\sum_{i=1}^{n}\delta_{ij}^2, \tag{7}$$

y a la media total como δ_{**}^2, se tiene que

$$\delta_{**}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}^2 = \frac{2t}{n}; \tag{8}$$

por (5), (6), (7) y (8) se obtiene que

$$\delta_{ij}^2 = \delta_{i*}^2 + \delta_{*j}^2 - \delta_{**}^2 - 2q_{ij}. \tag{9}$$

A partir de (9) se concluye que

$$q_{ij} = -\frac{1}{2}\left(\delta_{ij}^2 - \delta_{i*}^2 - \delta_{*j}^2 + \delta_{**}^2\right).$$

Esta expresión nos permite construir la matriz de similitud Q a partir de la matriz de distancias D [Peña, 2002].

- **Obtensión de la matriz X a partir de la matriz Q**

Suponiendo que Q es definida positiva de rango p, se puede representar mediante la expresión

$$Q = V \times \Lambda \times V',$$

donde:

- V es una matriz de dimensión $n \times p$ y contiene los vectores propios correspondientes a valores propios no nulos de Q,
- Λ es una matriz diagonal de dimensión $p \times p$ que contiene los valores propios de Q y,
- V' es la transpuesta de la matriz V y tiene dimensión $p \times n$;

reescribiendo a la matriz Q como

$$Q = (V \times \Lambda^{\frac{1}{2}}) \times (\Lambda^{\frac{1}{2}} \times V'),$$

y llamando

$$X = V \times \Lambda^{\frac{1}{2}},$$

se obtiene una matriz $n \times p$ con p variables no correlacionadas que reproducen la métrica inicial y es la solución buscada [Peña, 2002].

2.2.5. Análisis de conglomerados

Conocido también como análisis de clúster, es una técnica estadística multivariante que busca agrupar elementos tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos. Este análisis es una técnica descriptiva, a teórica y no inferencial; no tiene bases estadísticas, se utiliza fundamentalmente como una técnica exploratoria, descriptiva pero no explicativa. Las soluciones no son únicas, dependen de las variables consideradas y del método de análisis de clúster empleado. La adición o exclusión de variables relevantes puede tener un impacto substancial sobre la solución resultante [Fuente, 2011].

El análisis de conglomerados estudia tres tipos de problemas: partición de datos, construcción de jerarquías y, clasificación de variables.

- **Partición de datos**

Se divide el conjunto de observaciones en un número prefijado de k grupos, de manera que cada observación pertenezca a uno y solo a uno de los grupos. Todo elemento queda clasificado y cada grupo será internamente homogéneo.

Inicialmente se selecciona k puntos, como centros de los grupos, luego se calcula las distancias Euclídeas de cada elemento al centro de los grupos y se asigna cada elemento al grupo más próximo. Definidos los grupos se calculan las coordenadas de la media del grupo, convirtiéndose este punto en el nuevo centro del grupo. Se define un criterio de optimalidad y finalmente se verifica si reasignando uno a uno cada elemento de un grupo a otro mejora el criterio; si no es posible mejorar el criterio de optimalidad, termina el proceso [Peña, 2002].

- **Construcción de jerarquías**

Inicialmente cada observación es un grupo y la matriz de distancias entre los elementos representa la matriz de distancias entre los grupos. A continuación se determinan los grupos más cercanos, los que se unen y pasan a formar un nuevo grupo de un nivel superior, luego se determina la distancia del nuevo grupo al resto de grupos con algún criterio de distancia entre grupos (distancia entre centroides, distancia entre los elementos más cercanos, distancia entre los elementos más lejanos, etc.) y se repite el proceso hasta que todas las observaciones formen un solo grupo. El resultado

es una estructura jerárquica de conglomerados, que puede representarse gráficamente en un árbol, llamado dendrograma [Peña, 2002].

- **Conglomerados por variables**

Es un procedimiento exploratorio que puede sugerir procedimientos de reducción de la dimensión. La idea es construir una matriz de distancias o similitudes entre las variables y aplicar a esta matriz un algoritmo jerárquico de clasificación. Para que la distancia no dependa de las unidades de las variables estas deben estar estandarizadas, es decir, tener media cero y varianza uno [Peña, 2002].

En el presente estudio la conformación de conglomerados o grupos se efectuará en función de variables cualitativas como: Sexo, Actividad Económica, Educación etc.; lo que permitirá obtener grupos con características homogéneas, sobre los cuales se aplica las estrategias de estudio considerando un perfil médico determinado.

2.3. Síndrome Metabólico

El Síndrome Metabólico, SM, se refiere a la presencia de tres o más de las siguientes condiciones: exceso de grasa abdominal; nivel elevado de glucemia; hipertensión; triglicéridos elevados,y colesterol total elevado [Múnera and Escobar, 2007].

El SM se presenta en el 5 % de las personas de peso normal, en el 22 % de las personas que tienen sobre peso y en el 60 % de las obesas. El riesgo de presentar diabetes se multiplica por 5, el padecer de una enfermedad cardiovascular se multiplica por 2. También se puede citar el aumento de riesgo de hígado graso, enfermedad renal crónica, aumento de ácido úrico, etc. [Calle, 2011].

En el año 2005 la IDF [6] propuso que para que una persona tuviera SM era requisito tener un diámetro de la cintura superior a 94 cm, para los hombres, y 80 cm para el caso de mujeres,[7] la glucosa igual o mayor a 100 mg/dl; triglicéridos por encima de 150 mg/dl, colesterol total por encima de 200 mg/dl y la presión arterial sistólica por encima de 140 mmHg [Calle, 2011].

Entre las principales patologías que ocasiona el SM se encuentra la diabetes y la aterosclerosis.

[6] IDF International Diabetes Federation (Federación Internacional de Diabetes).

[7] una persona que supera estas medidas presenta obesidad y asocia un aumento de grasa en las vísceras, [Calle, 2011].

La diabetes es una enfermedad en la que la cantidad de glucosa (azúcar) de la sangre está elevada. La glucosa se obtiene de los alimentos que se consume. La insulina es una hormona que ayuda a que la glucosa entre a las células para suministrar energía. Existen dos tipos de diabetes: la tipo 1 en la cual el cuerpo no produce insulina; en la tipo 2, el cuerpo no produce o no usa la insulina de manera adecuada. Sin suficiente insulina, la glucosa permanece en la sangre y si ésta sobrepasa de niveles permitidos puede ocasionar problemas en los ojos, los riñones, los nervios, enfermedades cardiacas, derrames cerebrales, amputaciones de miembros, etc. [MedlinePlus, 2019b].

La aterosclerosis es un factor de riesgo el cual se caracteriza por la formación de placas en forma de parches (ateromas) en la capa interna de las arterias. Las placas contienes lípidos, células inflamatorias, células musculares lisas y tejido conectivo. Los factores de riesgo para el desarrollo de esta enfermedad son la dislipidemia,[8] la diabetes, el tabaquismo, los antecedentes familiares, el estilo de vida sedentario, la obesidad y la hipertensión arterial. Los síntomas aparecen por el crecimiento o rotura de las placas que obstruyen el flujo sanguíneo. El diagnóstico se confirma con una evaluación clínica, ecografía, angiografía u otros estudios de diagnóstico de imagen [Lam, 2016]. Los principales problemas que esta puede ocacionar son el bloqueo de las arterias coronarias produciendo angina de pecho y ataque cardíaco; el bloqueo de las arterias carótidas (ataque cerebral) y el bloqueo de las arterias periféricas.

El Dr. Valentín Fuster [9] en la conferencia dictada en CEDE / ICLD (International Center for Leadership Development) manifiesta lo siguiente:

> *...que el infarto de miocardio y el infarto cerebral son la causa número uno de la mortalidad en el mundo. Existen siete factores de riesgo que dañan las arterias que van al corazón y al cerebro; de estos siete factores tenemos dos mecánicos, la obesidad y la presión arterial; dos químicos, el colesterol y la diabetes; dos considerados como hábitos, si es fumador y si hace ejercicio; y el séptimo factor corresponde a la edad, por encima de los 55 años. El 95 % de las personas con infarto de miocardio o cerebral tiene dos de estos siete factores. De manera particular en España, el 75 % de la población por encima de los 50 años tiene dos factores de riesgo.*

El deterioro de la salud de la población debido al incremento de los malos hábitos nutricionales,

[8] La dislipidemia es la elevación anormal de la concentración de grasas en la sangre (triglicéridos, colesterol HDL y LDL[Lam, 2016]

[9] El Dr. Valentín Fuster es un médico investigador, Director General del Centro Nacional de Investigaciones Cardiovasculares Carlos III (España) y Director del Instituto Cardiovascular del Hospital Monte Sinaí de Nueva York [Fuster, 2013]

la obesidad y la hipertensión, hacen que busquemos soluciones a este problema. En esta era de alta tecnología, el cuidado de la salud requiere de la interacción de varias disciplinas médicas y especialidades en donde el laboratorio clínico es un aporte importante para prevenir, monitorear y curar una enfermedad. Los exámenes básicos de laboratorio permiten detectar la función de los órganos. A este grupo de pruebas se los describe como perfiles, según el órgano que se seleccione para monitorear; por ejemplo: perfil renal, perfil hepático, perfil lipídico, perfil tiroídeo, entre otros [Fuster, 2013].

Por ejemplo el perfil lipídico es uno de los exámenes de laboratorio más requeridos, pues mide las concentraciones de distintos tipos de grasas en la sangre, se solicita para investigar el riesgo de desarrollar una enfermedad cardiovascular y ateroesclerosis producto de un trastorno en el metabolismo de lípidos, los valores que mide este perfil son: colesterol total, colesterol HDL [10], colesterol LDL [11] y triglicéridos [12].

En este estudio nos centraremos en analizar el Síndrome Metabólico el cual considera las variables: Índice de Masa Corporal (IMC), Colesterol, Triglicéridos, Glucosa, Presión Sistólica y Presión Diastólica. Sin embargo, debido a la importancia de la edad y los hábitos Sedentarismo (S) y Tabaquismo (T) en nuestro estudio estas variables también serán consideradas, por lo cuál a este conjunto de variables las llamaremos **Síndrome Metabólico Plus (SM+)** .

2.4. Actividad económica

La actividad económica donde se desenvuelve el paciente, definirá aspectos como su tipo de alimentación, nivel de actividad física diaria, actividad intelectual, nivel de estrés, etc., los mismos que influirán notablemente en su forma de vida y estado de salud.

Dado que se está trabajando con una población económicamente activa, resulta de interés investigar cual es el comportamiento “de salud” en cada rama de actividad, pues en el caso de encontrar patrones particulares; esto, permitiría orientar recomendaciones de salud a cada una de ellas.

La data original proporcionada por el laboratorio clínico contenia unicamente información del nombre de la empresa donde labora el paciente; con esta información, para determinar la actividad

[10] Colesterol HDL o lipoproteína de alta densidad también llamado colesterol “bueno”.

[11] Colesterol LDL o lipoproteína de baja densidad también llamado colesterol “malo”.

[12] Los Triglicéridos almacenan energía hasta que el organismo la necesita. Si el cuerpo acumula demasiados triglicéridos, también puede obstruir los vasos sanguíneos.

económica a la que pertenece el paciente fue necesario ligar el nombre de la empresa con su registro en la Superintendencia de Compañías y en el Servicio de Rentas Internas, SRI, de la actividad principal que realiza esta empresa; obteniendose su categorización de acuerdo a la Clasificación Nacional de Actividades Económicas CIIU-4.0, que es una adaptación realizada por el Instituto Nacional de Estadística y Censos, INEC de la Clasificación Industrial Internacional Uniforme (CIIU) de las Naciones Unidas. Esta labor fue ralizada por los directores de este proyecto de investigación.

Las principales actividades económicas, en porcentajes mayores al 10 % de la población total, que se encuentra en la data son:

- Explotación de Minas,
- Actividades Financieras,
- Construcción,
- Actividades Profesionales y,
- Otras actividades

La Figura 1, presenta la distribución de la población total por actividades económicas.

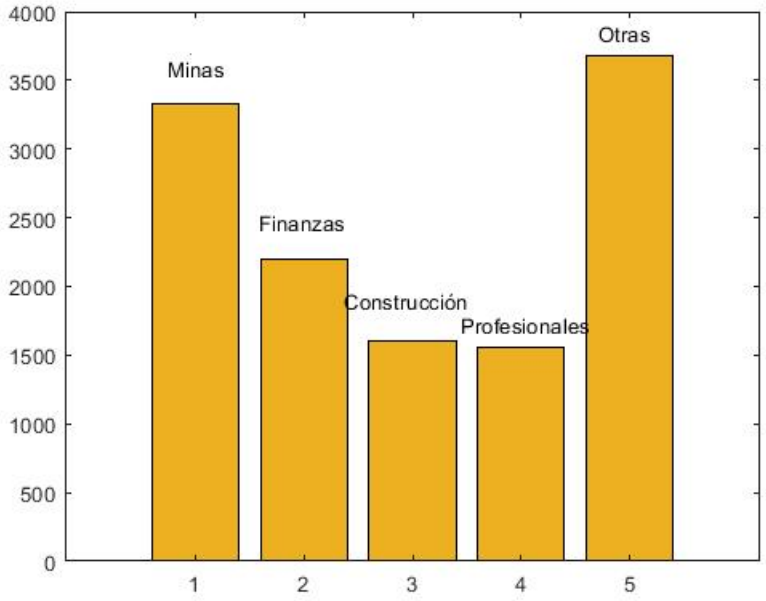


Figura 1. Distribución de la población considerando las actividades económicas más relevantes presentes en la data.

Capítulo 3

3. METODOLOGÍA

La data con la que se va a trabajar ha sido proporcionada por un laboratorio clínico de la ciudad de Quito, en la misma se encuentra a individuos que mantienen una relación de dependencia con diversas empresas, es decir, la población de estudio es un segmento económicamente activo, con una relación laboral dependiente.

El laboratorio a ido almacenando la información de los pacientes, con los respectivos resultados de exámenes clínicos, cabe indicar que toda la información confidencial a sido anonimizada, precautelando la privacidad de los pacientes.

3.1. Preparación y estrategia para el análisis de la data

Para el uso de la data es necesario realizar un análisis de la misma para eliminar posibles inconsistencias que pueda existir en sus datos como: datos faltantes, atípicos, redundantes, etc.

La data proporcionada por el laboratorio se lo ordena en una matriz, las filas corresponden a las observaciones o pacientes que se realizaron exámenes clínicos en un momento determinado y las columnas contienen las siguientes variables: datos personales, hábitos, características físicas como peso, estatura o signos vitales; y resultados de exámenes clínicos. Una vez organizada la información se obtuvo 23 336 observaciones y 104 variables de tipo cualitativo y cuantitativo. La Tabla 2 muestra un extracto del Anexo 1 que contiene la lista de variables, el porcentaje de datos existentes en cada variable y su tipo; así por ejemplo, la variable correspondiente a Colesterol, contiene datos en el 79.94 % del total de las observaciones y es de tipo cuantitativa.

Dado que existen variables con un bajo porcentaje de observaciones, solo se consideran las que tienen 15 por ciento o más, reduciéndose la data a 23 336 observaciones y 46 variables. Cabe indicar que la no existencia de datos en determinadas variables no necesariamente se debe a un mal registro, sino a que los individuos correspondientes no se realizaron esos exámenes, pues no fueron solicitados o no les correspondía.

Tabla 2

Extracto de las variables existentes en la data.

Número	Variable	Porcentaje	Tipo
1	Identificación	100.00	N
2	Fecha admisión	100.00	N
⋮			
6	Sedentarismo	100.00	L
7	Tabaquismo	100.00	L
⋮			
39	Glucosa	90.97	N
40	Triglicéridos	79.71	N
41	**Colesterol**	**79.94**	**N**
⋮			
48	Bilirrubina directa	5.09	N
49	Bilirrubina indirecta	5.09	N
50	Bilirrubina total	1.29	N
⋮			
103	Anticuerpos anti herpes IGG	0.01	N
104	Anticuerpos anti herpes IGM	0.01	N

N: Variable cuantitativa; L: Variable cualitativa

Para el caso de individuos que presentan más de un control médico, y por tanto tienen más de una observación en la matriz de datos, se considera únicamente la observación corespondiente a la última cita médica, para así tener una data donde las observaciones sean estadísticamente independientes, reduciéndose la data a 18 657 observaciones y 46 variables.

A esta nueva data se añaden 4 variables: dos que indican la actividad económica de la empresa donde labora el individuo (CIIU 40; CIIU 40 desagregado) de acuerdo a las normas CIIU,[13] la tercera es una variable de identificación de las observaciones, en este estudio para mantener el anonimato de los pacientes, esta variable se llama ID y; finalmente la cuarta variable es la Edad que corresponde a la edad del paciente en años, al momento de realizarce el chequeo, la misma que se construye a partir de las variables Fecha de Nacimiento y Fecha de Admisión. De esta manera se tiene una data de 18 657 observaciones y 50 variables.

Las variables Colesterol LDL calculado, con datos en el 70.28 % de las observaciones y Colesterol LDL cuantificado, con datos en el 1.98 % se fusionan en una sola variable llamada Colesterol LDL, ya que si la variable Colesterol LDL cuantificado posee información en una observación la

[13]Es un instrumento que sirve para clasificar a las unidades de producción, dentro de un sector de la economía, según la actividad económica principal que desarrolle.

variable Colesterol LDL calculado no la posee y viceversa.

Analizando las variables, una a una, se procede a detectar y eliminar datos inconsistentes. El proceso consiste en ordenar todos los valores de la variable en forma ascendente lo que facilita ubicar los valores extremos e inconsistentes que se eliminan. Así, por ejemplo, en la variable Presión Sistólica la observación número 6 852 tiene un valor de 1 y la observación 3 287 tiene un valor de 1 125, valores que fueron eliminados por inconsistentes.

Tabla 3
Variables seleccionadas para el estudio.

Variable	Unidades	Símbolo	Tipo
Edad	años	E	cuantitativa
Indice de Masa Corporal	kg/m^2	IMC	cuantitativa
Colesterol	mg/dl	C	cuantitativa
Triglicéridos	mg/dl	Tri	cuantitativa
Glucosa	mg/dl	G	cuantitativa
Presión Sistólica	mmHg	Ps	cuantitativa
Presión Diastólica	mmHg	Pd	cuantitativa
Sedentarismo (Hábito S)		S	cualitativa
Tabaquismo (Hábito T)		T	cualitativa
Sexo		Sx	cualitativa
Educación		Edu	cualitativa
Actividad Económica (CIIU 40)		Ae	cualitativa
Identificación		ID	cualitativa

Para las 13 variables indicadas en la Tabla 3, existen 12 363 observaciones que poseen información en todas estas; por tanto, la data para el estudio de este perfil, queda con 12 363 observaciones y 13 variables.

El análisis de los datos tiene como finalidad encontrar procedimientos para resumir la información que esta contenida en ellos, es decir, construir modelos matemáticos para representar la realidad que esta oculta en la data, para luego inferir comportamientos para determinados segmentos de la población. El estudio de los datos, además permite entender las posibilidades y limitaciones que tiene la investigación propuesta, así como para desarrollar el pensamiento crítico que permita diferenciar las conclusiones que puedan obtenerse de los datos, de otras que carezcan de sustento [Peña, 2008].

En el proceso de análisis se clasifica la población de acuerdo al número de categorías definidas sobre las variables cuantitativas:

- Si se construyen dos o más categorías según diversos niveles de gravedad que presentan estas variables, entonces el enfoque será la creación del "Índice de Salud".
- Si se consideran dos categorías, es decir, si la variable representa o no un riesgo para el paciente, se tendrá el enfoque por "factores de riesgo".

Además sobre dos grupos representativos de la población se aplicará la técnica del **escalado multidimensional**, lo que permitirá obtener diversos criterios sobre el comportamiento de las variables en la población, su grado de influencia en el quebrantamiento de la salud de los pacientes e inferir recomendaciones para mantener un adecuado estado de salud en la población.

En resumen, el análisis de los datos estará constituido por:

- Análisis descriptivo de las variables cuantitativas, con:
  - Categorización de las variables cuantitativas y,
  - Análisis estadístico descriptivo de las variables cuantitativas, y
- Análisis inferencial predictivo, con:
  - Construcción del índice de salud,
  - Determinación de los factores de riesgo para todos los pacientes de la data y,
  - Aplicación de la técnica de Escalado Multidimensional a un grupo de la población.

3.2. Análisis descriptivo de las variables cuantitativas

El análisis descriptivo se aplica como primer paso para comprender la estructura de los datos y extraer la información que contienen, antes de pasar a los métodos más complejos [Peña, 2002].

3.2.1. Categorización de las variables cuantitativas

Las 7 variables cuantitativas, indicadas en la Tabla 3, se categorizan de acuerdo a rangos pre definidos en la literatura médica; a estas categorías se les asigna valores donde:

- 1 representa una condición bajo valores de normalidad,
- 2 la condición de normalidad,
- 3 condición sobre normalidad y,
- 4, 5 o 6 para condiciones críticas.

Para las variables consideradas se tiene las siguientes tablas:

Tabla 4
Categorías para la variable Edad.

Categoría	Símbolo	Rango	Valor
menores a 30 años	m30	<30	1
de 30 a 40 años	30_40	$[30, 40]$	2
de 40 a 50 años	40_50	$[40, 50]$	3
de 50 a 60 años	50_60	$[50, 60]$	4
mayores a 60 años	M60	≥ 60	5

Tabla 5
Categorías para la variable Índice de Masa Corporal.

Categoría	Símbolo	Rango	Valor
bajo peso	IMC_b	<18.5	1
normal	IMC_n	$[18.5, 25]$	2
sobre peso	IMC_sp	$[25, 30]$	3
obeso 1	IMC_ob1	$[30, 35]$	4
obeso 2	IMC_ob2	$[35, 40]$	5
obeso extremo	IMC_ob3	≥ 40	6

Fuente: [OMS, 2018]

Tabla 6
Categorías para la variable Colesterol.

Categoría	Símbolo	Rango	Valor
normal	C_n	<200	2
alto	C_a	≥ 200	3

Fuente: [Múnera and Escobar, 2007]

Tabla 7
Categorías para la variable Triglicéridos.

Categoría	Símbolo	Rango	Valor
normal	Tri_n	<150	2
alto	Tri_a	≥ 150	3

Fuente: [Calle, 2011]

Tabla 8
Categorías para la variable Glucosa.

Categoría	Símbolo	Rango	Valor
baja	G_b	<70	1
normal	G_n	$[70, 100]$	2
alta	G_a	≥ 100	3

Fuente: [Calle, 2011]

Tabla 9
Categorías para la variable Presión Sistólica.

Categoria	Símbolo	Rango	Valor
baja	Ps_b	<100	1
normal	Ps_n	$[100, 140]$	2
alta	Ps_a	≥ 140	3

Fuente: [Rotaeche del Campo, 2002]

Tabla 10
Categorías para la variable Presión Diastólica.

Categoía	Símbolo	Rango	Valor
baja	Pd_b	<60	1
normal	Pd_n	$[60, 90]$	2
normal alta	Pd_na	$[90, 95]$	3
alta	Pd_a	≥ 95	4

Fuente: [Rotaeche del Campo, 2002]

3.2.2. Análisis estadístico descriptivo de las variables cuantitativas

Para el análisis descriptivo de las 7 variables cuantitativas indicadas en la Tabla 3, se elabora una tabla que contiene medidas de localización como la media y mediana; de variabilidad como la varianza, además de la cantidad y el porcentaje de elementos presentes en las diferentes categorías de cada variable. También se tiene un histograma que ayuda a visualizar este análisis. A partir de la varianza, se calcula la desviación estándar la cual permite relacionar las categorías de cada variable. Finalmente se calcula los coeficientes de asimetría[14] y de curtosis[15] que complementan este análisis.

[14]El coeficiente de asimetría de una variable mide el grado de asimetría de la distribución de sus datos con respecto a su media, si es negativo la distribución se alarga para valores menores a la media, si es positivo la distribución se alarga para valores superiores a su media. Un valor muy alto indica la presencia de datos atípicos [Peña, 2008].

[15]El coeficiente de curtosis nos indica el grado de concentración de los valores en torno a su media, Este coeficiente es siempre mayor o igual que 1, valores cercanos a 1 representa una distribución muy heterogenea, valores alrededor de 2 representa una distribución achatada, el valor de tres corresponde a una distribución normal, un valor mayores a 3 indican que la distribución es escarpada; si es muy alto mayor que 6, indica la presencia de valores atípicos [Peña, 2008].

Estadísticos descriptivos de Edad

La Tabla 11 presenta la media, mediana y varianza de la variable Edad; de la población total y por categorías de edad. Los datos presentados se encuentra en **años**.

Tabla 11
Estadísticos de la variable Edad, por categorías.

Categoría	Elementos	Media	Mediana	Varianza	Porcentaje
m30	3 849	25.62	26	7.16	31.10
30-40	4 828	34.12	34	7.90	39.10
40-50	2 461	43.85	44	8.06	19.80
50-60	1 021	53.54	53	7.33	8.30
M60	204	62.82	62	8.68	1.70
Total	12 363	35.49	34	91.87	100.00

El histograma de esta variable se observa en la Figura 2:

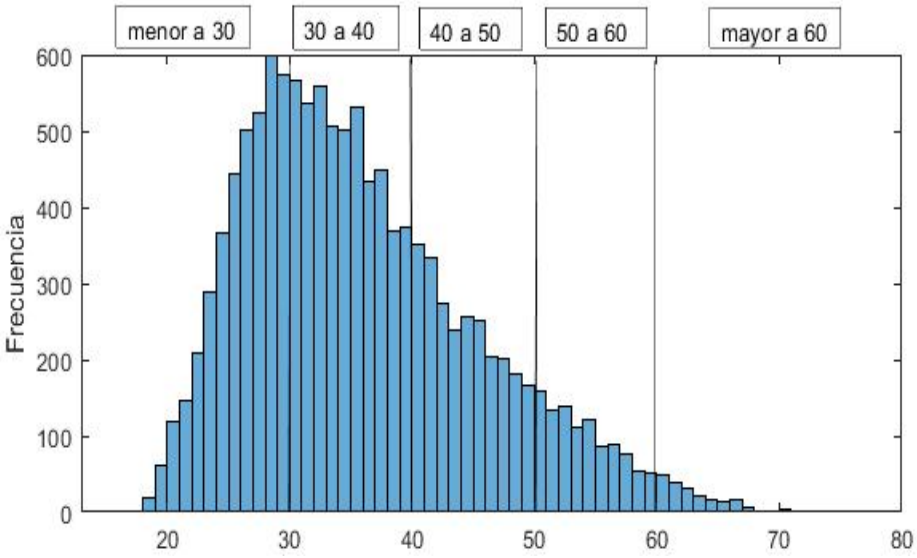


Figura 2. Histograma de la variable Edad.

- La media de la variable edad μ es de 35.49 años, su varianza σ^2 es 91.87 y su desviación estándar σ es 9.58 años.
- Los grupos predominantes son: menores de 30 años con el 31.10 %, de 30 a 40 con el 39.10 %, y de 40 a 50 con el 19.90 %; en conjunto representan el 90.00 % de la población.
- El coeficiente de asimetría es de 0.736, lo que implica que existe una leve asimétrica a la derecha. La cola de la derecha es más larga que la cola de la izquierda, esto ratifica la

presencia de pacientes en las categorías de 50 a 60 años y en mayores a 60 años, es decir, población económicamente activa o jubilados.

- El coeficiente de curtosis es de 3.101, lo cual indica que la población tiene una distribución normal; el intervalo $[\mu - \sigma\ ,\ \mu + \sigma] = [26\ ,\ 45]$ años contendrá aproximadamente el 68 % de los datos.

Estadísticos descriptivos de IMC

La Tabla 12 presenta la media, mediana y varianza de la variable IMC de la población total y por categorías de IMC. Los datos presentados se encuentran en $\mathbf{kg/m^2}$.

Tabla 12
Estadísticos de la variable IMC, por categorías.

Categoría	Elementos	Media	Mediana	Varianza	Porcentaje
IMC_b	81	17.71	17.96	0.57	0.70
IMC_n	4 604	22.83	23.13	2.43	37.20
IMC_sp	5 611	27.26	27.14	1.91	45.40
IMC_ob1	1 696	31.77	31.51	1.83	13.70
IMC_ob2	311	36.85	36.51	1.89	2.50
IMC_ob3	60	43.16	42.23	8.60	0.50
Total	12 363	26.48	26.16	15.74	100.00

El histograma de esta variable se observa en la Figura 3:

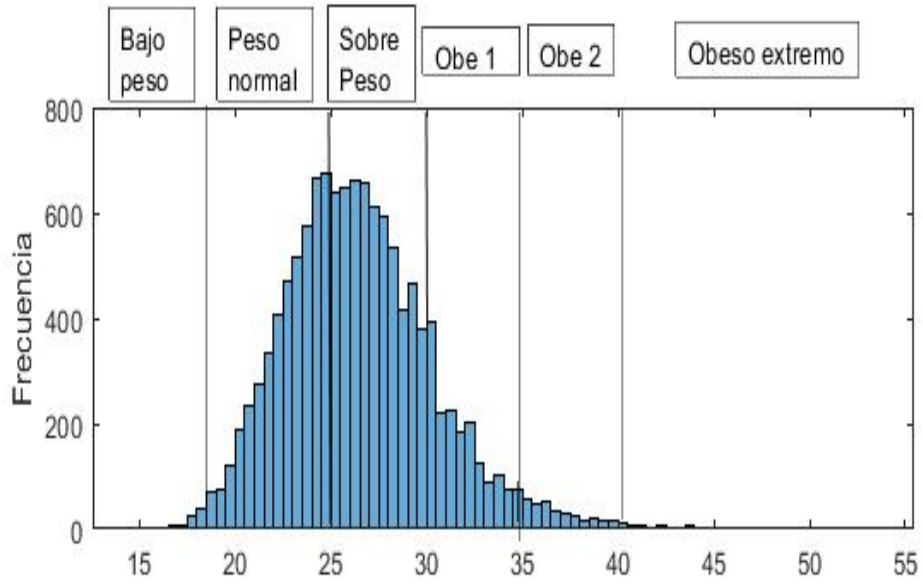


Figura 3. Histograma de la variable Índice de Masa Corporal.

- La media μ de la variable IMC es de 26.48 kg/m^2, valor que se ubica en el grupo de sobrepeso. Este primer resultado indica que más de la mitad de la población tiene sobrepeso o algún nivel de obesidad. La varianza σ^2 es 15.74 kg^2/m^4 y la desviación estándar σ es 3.97 kg/m^2.

- Los grupos predominantes son: Normal con el 37.20 %, Sobrepeso con el 45.40 %, y Obeso 1 con el 13.70 %.

- El porcentaje acumulado de las personas que rebasaron los límites de normalidad es del 62.10 % (aproximadamente 2 de cada 3 personas).

- El valor de la desviación estándar indica que un paciente del grupo normal con media 22.83 kg/m^2 facilmente puede pasar a un valor $\mu+\sigma = 26.79$ kg/m^2, que corresponde a sobrepeso; o pasar de un estado de sobrepeso con media 27.26 kg/m^2 a un valor $\mu + \sigma = 31.22$ kg/m^2 que corresponde a obeso1; si no mantiene un adecuado control de su alimentación.

  Al subir el IMC, aumenta el riesgo de desarrollar enfermedades como la diabetes, patologías cardiovasculares, o algunos tipos de cánceres, lo que puede producir la muerte prematura o discapacidad en la edad adulta.

- El coeficiente de asimetría es 0.763, esto implica que existe una leve asimetría derecha, es decir, existen más datos a la derecha de la media en el sector de sobrepeso y los distintos niveles de obesidad.

- El coeficiente de curtosis es de 4.595. Esto implica un mayor agrupamiento de los datos alrededor del valor central, es decir, en las categorías normal y sobrepeso, lo que se confirma con los datos ya que estas categorías en conjunto contienen el 82.6 % de los datos.

Estadísticos descriptivos de Colesterol

La Tabla 13 presenta la media, mediana y varianza de la variable Colesterol de la población total y por categorías de Colesterol. Los datos presentados estan dados en **mg/dl**

Tabla 13
Estadísticos de la variable Colesterol, por categorías.

Categoría	Elementos	Media	Mediana	Varianza	Porcentaje
C_n	7 535	167.08	170	483 11	60.90
C_a	4 828	230.35	224	796.67	39.10
Total	12 363	191.78	189	1 558.00	100.00

El histograma de esta variable se observa en la Figura 4:

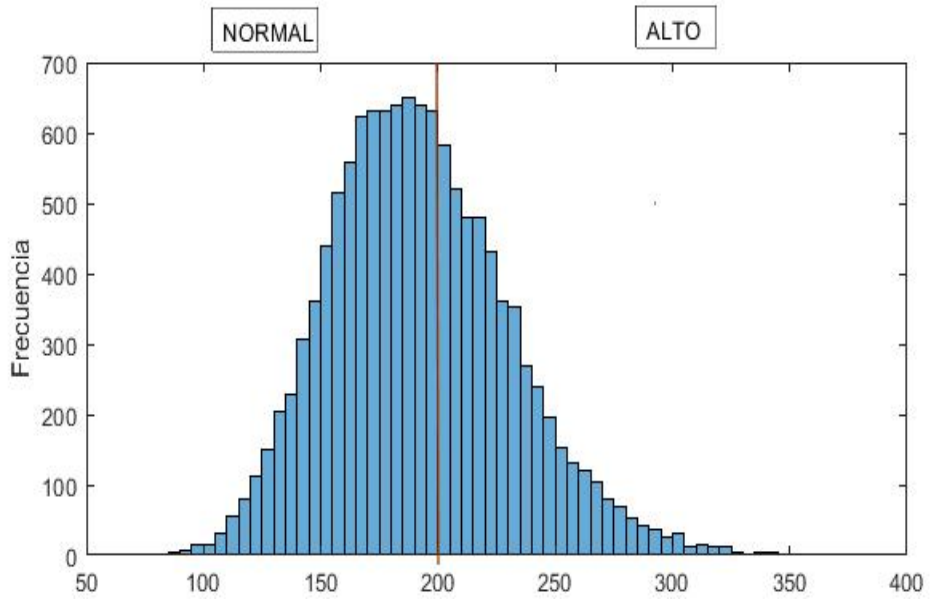


Figura 4. Histograma de la variable Colesterol.

- La media μ de la variable Colesterol es de 191.78 mg/dl, valor que está en el rango normal, sin embargo muy, cerca de su límite superior. La varianza σ^2 es de 1 558 mg^2/dl^2 y la desviación estándar σ es de 39.47 mg/dl.

- El porcentaje de personas que tienen Colesterol Alto es del 39.10 % (aproximadamente 2 de cada 5 personas).

- El valor de la desviación estándar indica que un paciente del grupo normal con media 167.08 mg/dl fácilmente puede pasar a un valor $\mu + \sigma = 206.55$ mg/dl, que corresponde a la categoría Colesterol Alto; o pasar de un estado de Colesterol Alto con media 230.35 mg/dl a un valor $\mu - \sigma = 190.88$ mg/dl que corresponde a Colesterol Normal; lo anterior significa que un paciente que mantiene esta variable en el rango de normalidad, fácilmente puede pasar a una condición de Colesterol Alto, pero también se ve que un paciente que ya se encuentra en el grupo Colesterol Alto, con un adecuado control puede llegar a la condición de normalidad.

  Si se tiene demasiado colesterol en la sangre, se produce una acumulación de placas en las arterias, que pueden estrecharlas o incluso bloquearlas originando enfermedades coronarias [MedlinePlus, 2019a].

- El coeficiente de asimetría es 0.9963, esto implica que existe asimetría derecha, es decir, existen más datos a la derecha de la media que corresponde al sector de Colesterol Alto.

- El coeficiente de curtosis es de 12.57. Esto implica un mayor agrupamiento de los datos alrededor del valor central y la presencia de datos atípicos.

Estadísticos descriptivos de Triglicéridos

La Tabla 14 presenta la media, mediana y varianza de la variable Triglicéridos de la población total y por categorías de Triglicéridos. Los datos presentados estan dados en **mg/dl**.

Tabla 14
Estadísticos de la variable Triglicéridos, por categorías.

Categoría	Elementos	Media	Mediana	Varianza	Porcentaje
Tri_n	7 856	94.67	93	848.72	63.54
Tri_a	4 507	244.28	208	16 509.00	36.46
Total	12 363	149.21	122	11 743.00	100.00

El histograma de esta variable se observa en la Figura 5:

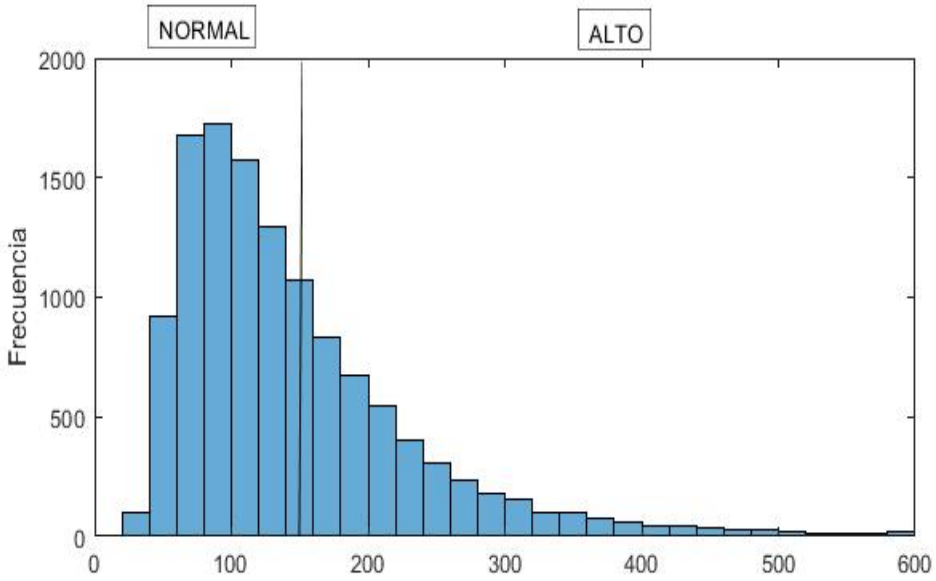


Figura 5. Histograma de la variable Triglicéridos.

- La media μ de la variable Triglicéridos es de 149.21 mg/dl valor que está muy cercano al límite de normalidad (150 mg/dl). La varianza σ^2 es de 11 743 mg^2/dl^2 y la desviación estándar σ es 108.36 mg/dl

- El porcentaje de personas que tienen problemas con los Triglicéridos, el grupo Alto es del 36.46 % (aproximadamente 1 de cada 3 personas).

- El valor de la desviación estándar indica que un paciente del Grupo Normal con media 94.67 mg/dl, fácilmente puede pasar a un valor $\mu + \sigma = 203.03$ mg/dl, que corresponde a la categoría Triglicéridos Alto; o un paciente del grupo Triglicéridos Alto con media 244.28 mg/dl, llegar a un valor $\mu - \sigma = 135.92$ mg/dl, que corresponde a Triglicéridos Normal; esto indica que un paciente que se encuentra en la categoría Triglicéridos Normal fácilmente puede pasar a la condición Triglicéridos Alto; por otro lado, un paciente que se encuentre en el grupo Triglicéridos Alto, con un adecuado control y cuidado podrá llegar a la condición de normalidad.

  Los triglicéridos es un tipo de grasa que se adquiere de los alimentos, el cuerpo consume lo que necesita y el resto se acumula como calorías adicionales, un exceso produce enfermedades del corazón [MedlinePlus, 2019c].

- El coeficiente de asimetría es 4.682, valor que indica una fuerte asimetría derecha, lo que indica que existen más datos a la derecha de la media que corresponde al sector de Triglicéridos Altos.

- El coeficiente de curtosis es de 51.909. Esto implica un alto agrupamiento de los datos alrededor del valor central y la existencia de datos atípicos.

Estadísticos descriptivos de Glucosa

La Tabla 15 presenta la media, mediana y varianza de la variable Glucosa de la población total y por categorías de Glucosa. Los datos presentados estan dados en **mg/dl**.

Tabla 15
Estadísticos de la variable Glucosa, por categorías.

Categoría	Elementos	Media	Mediana	Varianza	Porcentaje
G_b	18	65.00	65.00	7.64	0.10
G_n	10 600	87.72	88.65	53.79	85.80
G_a	1 745	111.43	103.00	893.95	14.10
Total	12 363	91.03	90.00	241.31	100.00

El histograma de esta variable se observa en la Figura 6:

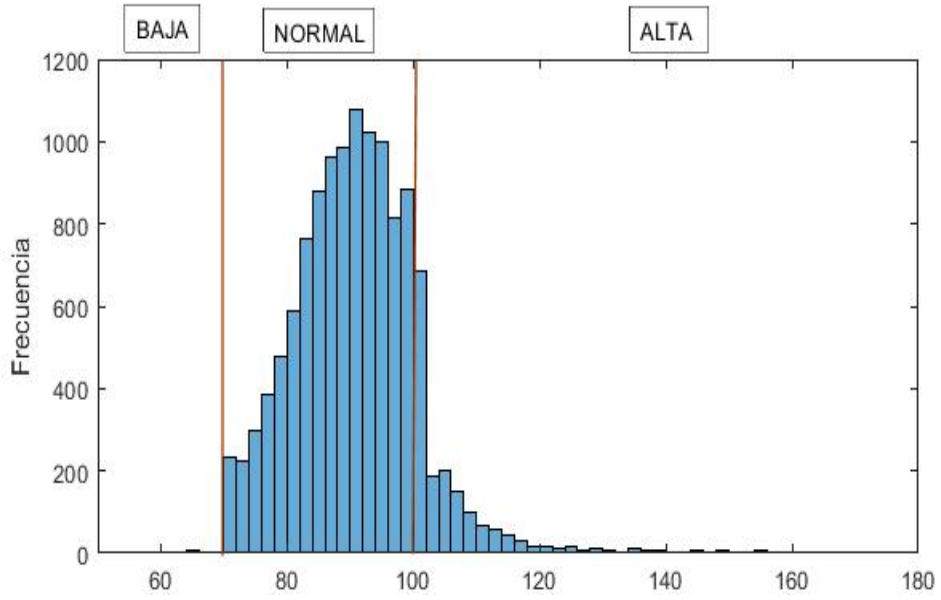


Figura 6. Histograma de la variable Glucosa.

- La media μ de la variable Glucosa es de 91.03 mg/dl, valor que corresponde al grupo de Glucosa Normal. La varianza σ^2 es 241.31 mg^2/dl^2 y la desviación estándar es 15.53 mg/dl.

- El porcentaje de personas que tienen la Glucosa Alta es del 14.10 % (aproximadamente 1 de cada 7 personas), el cual podría considerarse un grupo de riesgo.

- El valor de desviación estándar indica que un paciente del Grupo Normal con media de 87.72 mg/dl, puede pasar a un valor de $\mu + \sigma = 103.25$ mg/dl que corresponde a la categoría de Glucosa Alta; o un paciente del grupo Glucosa Alta con media de 111.43 mg/dl, llegar a un valor de $\mu - \sigma = 95.90$ mg/dl, que corresponde al grupo de Glucosa Normal; lo que indica un paciente que se encuentra en el grupo de Glucosa Normal facilmente puede pasar a la condición de Glucosa Alta, y de igual manera un paciente que se encuentre en el grupo de Glucosa Alta eventualmente podría pasar al grupo de Glucosa Normal. Niveles altos de glucosa (azúcar) produce la diabetes.

- El coeficiente de asimetría es 8.19, lo que indica una gran asimetría derecha, es decir, existen más datos a la derecha de la media en el sector de Glucosa Normal y Glucosa Alta.

- El coeficiente de curtosis es de 136.36. Valor que implica un alto agrupamiento de datos alrededor del valor central correspondiente al sector de Glucosa Normal

Estadísticos descriptivos de Presión Sistólica

La Tabla 16 presenta la media, mediana y varianza de la variable Presión Sistólica de la población total y por categorías de Presión Sistólica. Los datos presentados estan dados en **mmHg**.

Tabla 16
Estadísticos de la variable Presión Sistólica, por categorías.

Categoría	Elementos	Media	Mediana	Varianza	Porcentaje
Ps_b	830	91.82	90	16.07	6.70
Ps_n	11 173	112.73	110	95.96	90.40
Ps_a	360	145.71	141	69.64	2.90
Total	12 363	112.29	110	150.66	100.00

El histograma de esta variable se observa en la Figura 7:

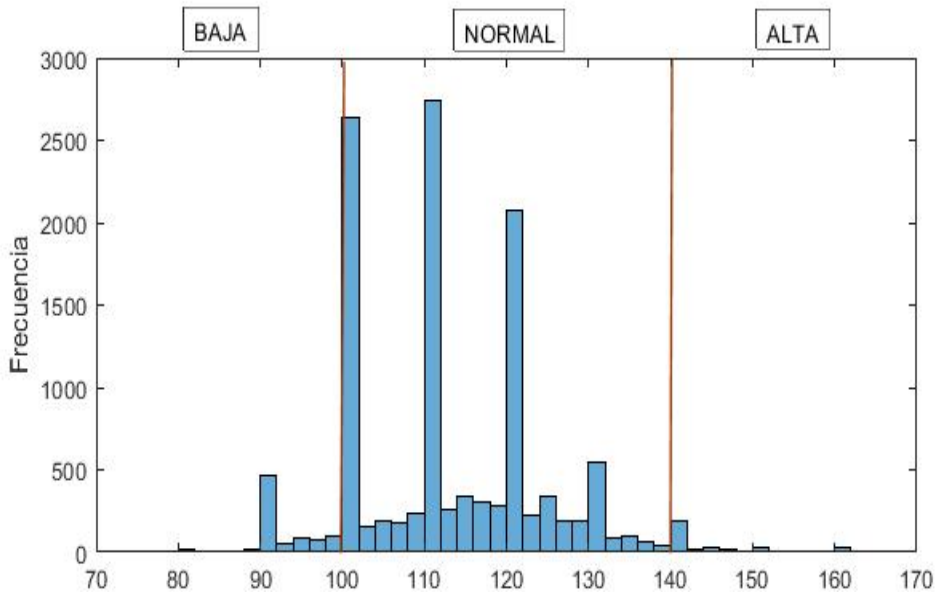


Figura 7. Histograma de la variable Presión Sistólica.

- La media μ de la variable Presión Sistólica es de 112.29 mmHg , valor que corresponde al grupo de Presión Sistólica Normal. La varianza σ^2 es 150.66 $mmHg^2$ y la desviación estándar σ es 12.27 mmHg.
- El grupo predominante es el de Presión Sistólica Normal con 90.40 %, es decir, se puede considerar que la mayoría de la población no tiene problemas con la Presión Sistólica.
- El porcentaje de las personas que tienen Presión Sistólica Alta es apenas del 2.90 % (aproximadamente 1 de cada 30 personas).

- Esta variable es trimodal, en los valores de 100, 110 y 120 mmHg, lo cual indica que un gran porcentaje de la población analizada tiene uno de estos valores

- En esta variable el valor de la desviación estándar indica que un paciente del grupo Presión Sistólica Normal con media de 112.73 mmHg puede pasar a un valor de $\mu + \sigma = 125$ mmHg, valor que esta dentro del mismo grupo, lo cual indica que tiene pocas probabilidades de tener presiones altas; en cambio un paciente del grupo de Presión Sistólica Alta con media de 145.71 mmHg puede tener un valor de $\mu - \sigma = 133.44$ mmHg que correspode al grupo de Presión Sistólica Normal, lo que indica que una persona con Presión Sistólica Alta con un poco de cuidado podría alcanzar la Presión Sistolítica Normal.

- El coeficiente de asimetría es 0.565, lo que representa en esta variable una moderada asimetría hacia la derecha.

- El coeficiente de curtosis es de 3.91. Esto implica un leve apuntalamiento de los datos alrededor del valor central es decir en el grupo de la Presión Sistólica Normal.

Estadísticos descriptivos de Presión Diastólica

La Tabla 17 presenta la media, mediana y varianza de la variable Presión Diastólica de la población total y por categorías de Presión Diastólica. Los datos de presión presentados estan dados en **mmHg**.

Tabla 17
Estadísticos de la variable Presión Diastólica, por categorías.

Categoría	Elementos	Media	Mediana	Varianza	Porcentaje
Pd_b	354	53.87	55	17.56	2.90
Pd_n	11 503	72.06	70	63.33	93.00
Pd_na	370	90.49	90	1.15	3.00
Pd_a	136	100.32	100	25.98	1.10
Total	12 363	72.40	70	88.05	100.00

El histograma de esta variable se observa en la Figura 8:

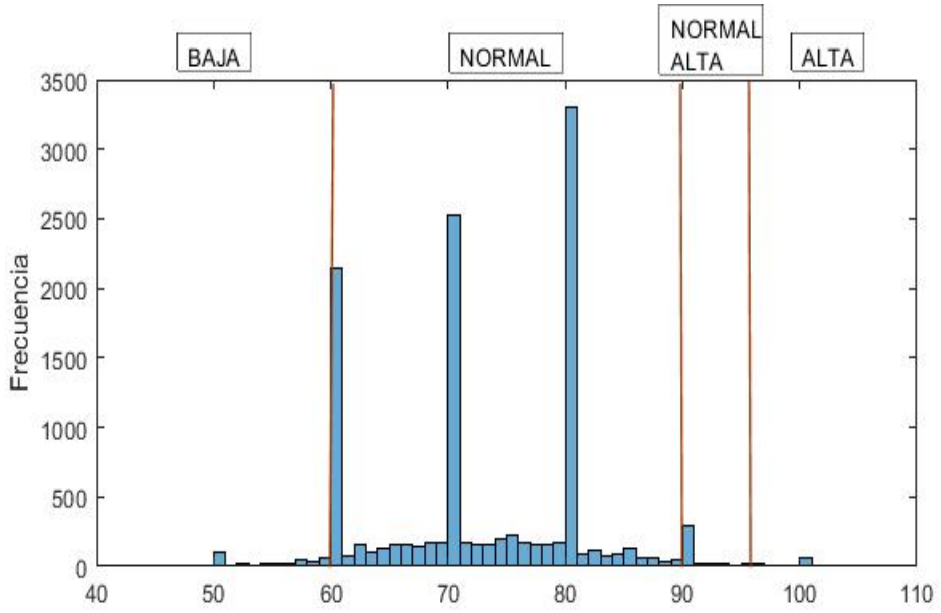


Figura 8. Histograma de la variable Presión Diastólica.

- La media μ de la variable Presión Diastólica es de 72.40 mmHg, valor que corresponde al grupo de Presión Diastólica Normal. La varianza σ^2 es 88.05 $mmHg^2$ y la desviación estándar σ es 9.38 mmHg.

- El grupo predominante es el de Presión Diastólica Normal con el 93.00 %, es decir, la mayoría de la población no tiene problemas con esta clase de Presión.

- El porcentaje acumulado de las personas que tienen Presión Diastólica Alta es apenas del 4.10 % (aproximadamente 1 de cada 25 personas).

- El valor de la desviación estándar indica que un paciente del grupo Presión Diastólica Normal Alta con media de 90.49 mmHg puede pasar a un valor de $\mu + \sigma = 99.875$ mmHg, que esta en el grupo de Presión Diastólica Alta; de igual manera un paciente del grupo Presión Diastólica Alta con media de 100.32 mmHg puede tener un valor de $\mu - \sigma = 90.945$ mmHg que corresponde al grupo de Presión Diastólica Normal Alta; estos valores indican que es relativamente fácil pasar de la Presión Diastólica Alta a la Presión Diastólica Normal Alta y viceversa.

  Los síntomas que pueden indicar presión arterial baja son: debilidad, fatiga, somnolencia, sudor, náuseas, fiebre o escalofríos y los síntomas que pueden indicar presión arterial alta son: náuseas, dolor de cabeza, zumbido en los oídos, dificultad para respirar [Gasteiz, 2002].

- Esta variable es trimodal en los valores de 60, 70 y 80 mmHg, es decir, gran porcentaje de población analizada tiene uno de estos valores.

Es importante indicar que para mantenerse en niveles normales en todas estas 7 variables, es importante: mantener hábitos alimenticios saludables, es decir, no comer más calorías de las que se quema, el sobrepeso aumenta los niveles de colesterol, diabetes, triglicéridos, también puede ocasionar inconvenientes en la presión arterial. La falta de actividad física, con mucho sedentarismo y poco ejercicio también aumenta el riesgo de adquirir estas enfermedades [MedlinePlus, 2019a].

Distribución de la población de estudio por Actividades Económicas

Dado que la población de estudio consiste en personas que mantienen una relación de dependencia laboral, un parámetro importante ha considerar es la actividad económica en la que labora el paciente, estas actividades las podemos observar en la Tabla 18.

Tabla 18
Distribución de la población de estudio, por actividad económica.

Actividades económicas en la Data	Población	Porcentaje
Agricultura, ganadería, silvicultura y pesca	35	0.28
Explotación de minas y canteras	3 329	26.93
Industrias manufactureras	336	2.72
Suministro de electricidad, gas, vapor y aire acondicionado	55	0.44
Distribución de agua; alcantarillado, gestión de desechos	44	0 .36
Construcción	1598	12.93
Comercio al por mayor y al por menor	963	7.79
Transporte y almacenamiento	326	2.64
Información y comunicación	583	4.71
Actividades financieras y de seguros	2 199	17.79
Actividades inmobiliarias	5	0.04
Actividades profesionales, científicas y técnicas	1558	12.60
Actividades de servicios administrativos y de apoyo	708	5.73
Administración pública y defensa	15	0.12
Enseñanza	38	0.31
Actividades de atención de la salud humana	386	3.12
Otras actividades de servicios	161	1.30
Actividades de organizaciones y órganos extraterritoriales	24	0.19
Población total	12 363	100.00

En la población de estudio las actividades económicas relevantes, con porcentajes mayores al 10 %, se presentan en la Tabla 19 y se visualizan en la Figura 9,

Tabla 19
Principales actividades económicas de la población.

Grupo por Actividad Económica	Población	Porcentaje
Explotación de minas y canteras	3 329	26.93
Actividades financieras y de seguros	2 199	17.79
Construcción	1 598	12.93
Actividades profesionales científicas y técnicas	1 558	12.60
Otras Actividades	3 679	29.75
Población total	12 363	100.00

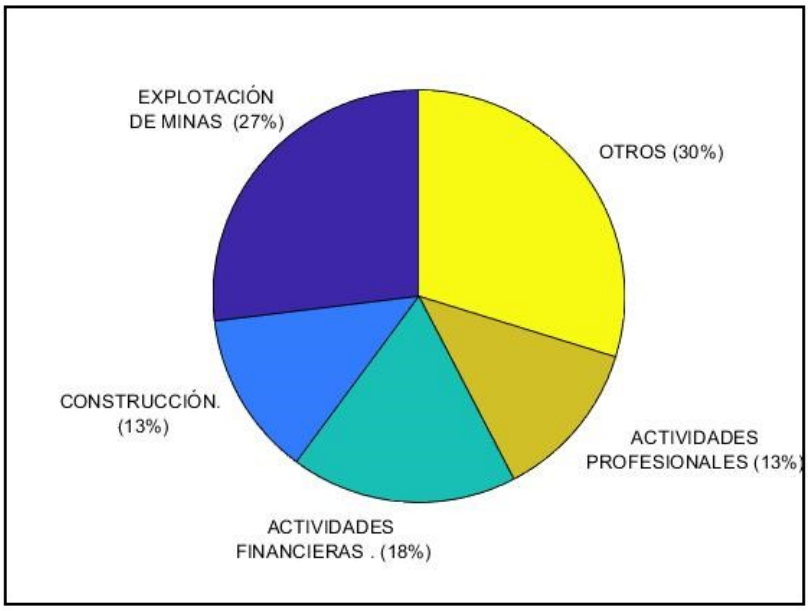


Figura 9. Distribución de la población por Actividades Económicas relevantes, con porcentajes mayores al 10 %.

3.3. Análisis inferencial predictivo

El análisis de los datos se realiza de tres maneras:

- El primero determinando **Índices de Salud**. Aquí a las variables se las subdivide en categorías, según intervalos proporcionados por la literatura médica consultada.
- El segundo hallando **factores de riesgo**. Aquí se transforma las variables cuantitativas en cualitativas, en particular, dicotómicas, de acuerdo a valores límite; es decir, valores superiores de la normalidad propia de cada variable.
- El tercero determina las **similaridades** (distancias) entre los distintos individuos de un grupo

de riesgo determinado. Los grupos de riesgo se forman por la combinación de los factores de riesgo construidos en el segundo método de análisis.

3.3.1. Índice de Salud

Para establecer el estado de salud de un paciente en función de los resultados de diversos exámenes clínicos, correspondientes a un perfil médico específico, será importante construir un criterio, en base a la información de miles de casos similares, que ayude a visualizar el estado de salud. Con este propósito se ha definido un Índice de Salud para el SM+, es decir, para el riesgo de tener un paro cardiovascular o cerebrel. Este índice constituye un referente para que el individuo modifique malos hábitos de vida como excesivo sedentarismo, alcoholismo, tabaquismo, etc.; y/o de alimentación como consumo excesivo de grasas, consumo de comida chatarra, consumo excesivo de carnes rojas, etc.

Las variables cuantitativas objeto de este estudio presentan diversos niveles de gravedad, por lo que se considera adecuado dividirlas en distintas categorías, de acuerdo al nivel de gravedad o riesgo; así, se asignará el valor de:

- 1, cuando el valor esta bajo el límite de normalidad,
- 2, cuando el valor esta dentro de los límites de normalidad,
- 3, cuando el valor esta sobre los valores de normalidad;
- 4 o 5, cuando el valor representa condiciones críticas para el paciente.

Para las variables consideradas, las Tablas desde la 6 a la 10 (Sec 3.2.1) muestran su categorización y valoración correspondiente.

Definición 3. *El "Índice de Salud' de un paciente, es la suma de los valores asignados a cada uno de las distintas variables, de acuerdo al nivel de gravedad o riesgo; variables correspondientes a los exámenes clínicos, de un perfil médico determinado.*

Constituído el Índice de Salud, su valor definirá el estado de salud del paciente. Para sus diversos valores se construirán categorías que nos permiten clasificar a la población de acuerdo a su estado de salud y así; por ejemplo, identificar los pacientes de grupos que requieran atención especial.

Las variables IMC, Edad, Sexo, Educación, Actividad Económica, permiten realizar análisis a grupos específicos de interés, o hacer comparaciones entre diversos grupos.

Las categorías definidas de acuerdo al Índice de salud se indica en la Tabla 20.

Tabla 20
Estado de salud de acuerdo al Índice de Salud.

Estado de salud	Índice de salud
Bajo nivel de normalidad	8 a 9
Normal	10
Sobre nivel de normalidad	11 a 12
Crítico	13 a 16

En la Tabla 21 se observa un segmento de data; por ejemplo, para la observación remarcada con ID 18 los valores para las variables Colesterol, Glucosa y Presión Sistólica, es 3, para Triglicéridos, 2 y; para la Presión Diastólica, 4. La columna "Índice de Salud" indica la suma total de estos valores. Para el ejemplo esta suma es 15 que corresponde a un "Estado de Salud" Crítico, lo cual implica que el paciente debe extremar cuidados para revertir su condición de salud.

Tabla 21
Segmento de la Data con variables valoradas: Índice de Salud y Estado de Salud.

ID	Colesterol	Trigliceridos	Glucosa	Ps	Pd	Índice de Salud	Estado de Salud
1	3	3	3	2	2	13	Crítico
2	3	2	2	2	2	11	Sobre normal
3	2	3	3	2	2	12	Sobre normal
5	3	2	2	2	2	11	Sobre normal
6	2	2	2	2	2	10	Normal
10	3	2	3	2	2	12	Sobre normal
11	3	3	2	2	3	13	Crítico
12	2	2	2	2	2	10	Normal
13	3	3	2	2	2	12	Sobre normal
16	2	3	3	2	2	12	Sobre normal
17	3	2	2	2	2	11	Sobre normal
18	**3**	**2**	**3**	**3**	**4**	**15**	**Crítico**
19	3	3	3	2	2	13	Crítico
20	3	3	2	3	3	14	Crítico
23	3	3	2	2	2	12	Sobre normal
24	2	3	3	2	3	13	Crítico
26	3	2	3	2	2	12	Sobre normal
⋮							

La Figura 10 muestra la distribución de la población total de acuerdo a las diversas categorías de

las variables de mayor influencia en el estado de salud del paciente, que son: Colesterol, Triglicéridos, Sedentarismo e IMC ubicadas en el ***eje horizontal*** y; las Actividades Económicas relevantes en el ***eje vertical***. En esta figura la elipse señala al grupo que posee Colesterol Alto, Triglicéridos Alto, es Sedentario y tiene Sobrepeso. La distribución por Actividad Económica para este grupo es: 338 pacientes en Explotación de Minas, 168 en Construcción, 131 en Actividades Financieras, 121 en Actividades Profesionales y 206 en otras actividades. La población de este grupo es de 964 pacientes.

El Anexo 2 presenta la distribución de la población total cambiando la variable IMC de la Figura 10 por las variables Sexo, Edad y Educación; respectivamente. Además se considera el número de pacientes de cada grupo así como el porcentaje que representa cada grupo respecto a la población total.

Las actividades económicas con mayor número de pacientes corresponden a: Explotación de Minas y Actividades Financieras, cada uno de estos grupos tienen particularidades respecto al desarrollo de su actividad, lo que se refleja en el análisis de cada uno de estos grupos por separado.

La Figura 11, presenta la distribución de la población de la actividad económica Explotación de Minas, de acuerdo a las diversas categorías de las variables: Colesterol, Trigliceridos y Sedentarismo en el ***eje horizontal*** y las categorías de la variable IMC en el ***eje vertical***. En esta figura por ejemplo la elipse nos muestra el grupo que posee Colesterol Alto, Triglicéridos Altos, es Sedentario y tiene Sobrepeso. La población de este grupo es 338 pacientes.

La Figura 12, presenta la distribución de la población de Actividades Financieras con las mismas variables de la Figura 11. La elipse en esta figura nos indica el grupo de 414 pacientes con Colesterol Normal, Triglicéridos Normal, es Sedentario y tiene IMC Normal.

En las Figuras 10, 11 y 12, las categorías de la variable Sedentarismo son: 1 cuando el paciente es sedentario y 0 en caso contrario.

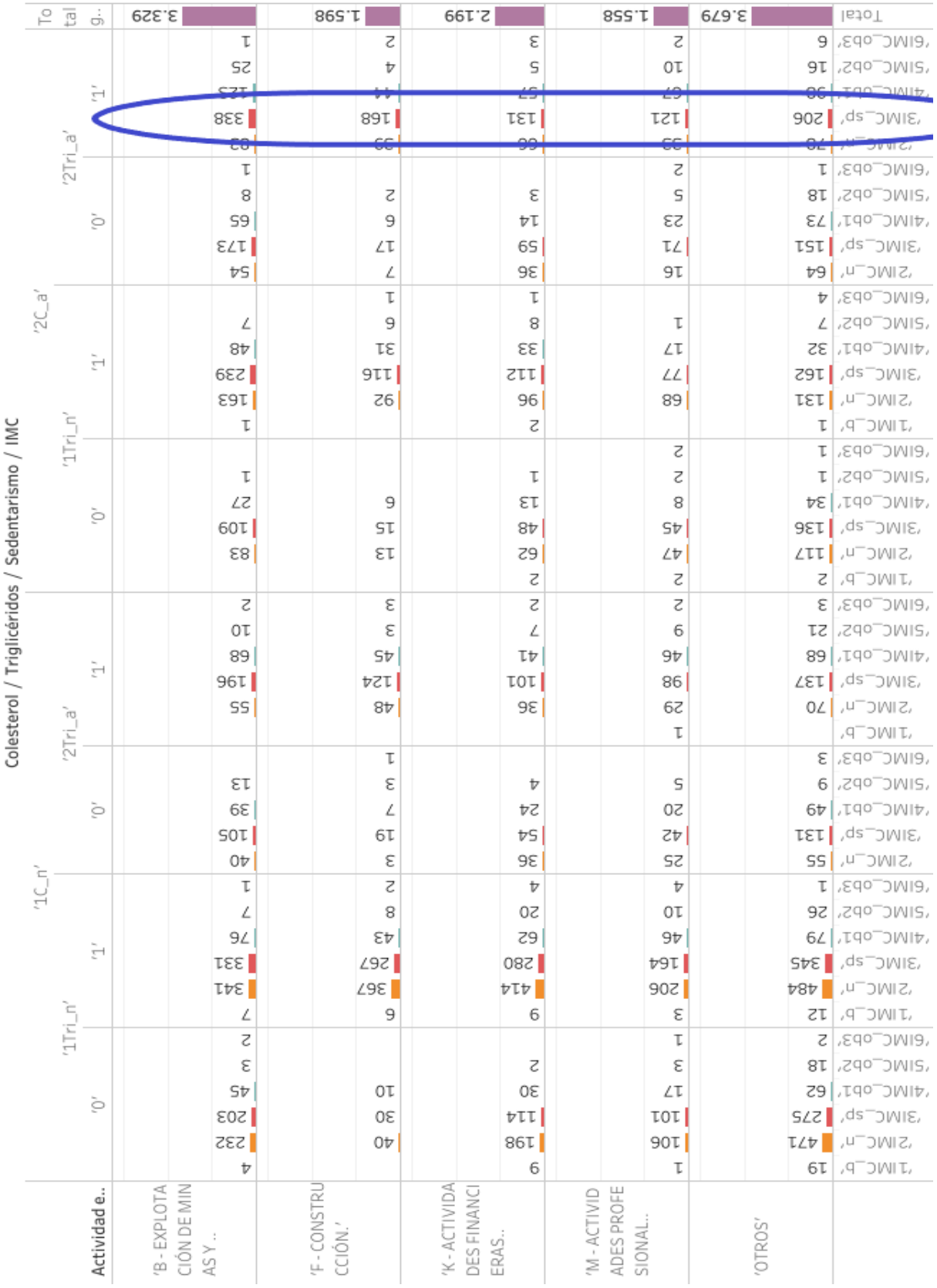


Figura 10. Distribución de la población total de acuerdo a las diversas categorías: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo e IMC.

La elipse señala al grupo que posee Colesterol Alto, Triglicéridos Alto, es sedentario y tiene sobrepeso. La distribución por Actividad Económica es: 338 pacientes en Explotación de Minas, 168 en Construcción, 131 en Actividades Financieras, 121 en Actividades Profesionales y 206 en otras actividades; la población total de este grupo es 964 pacientes, de una población total de 12 363 pacientes.

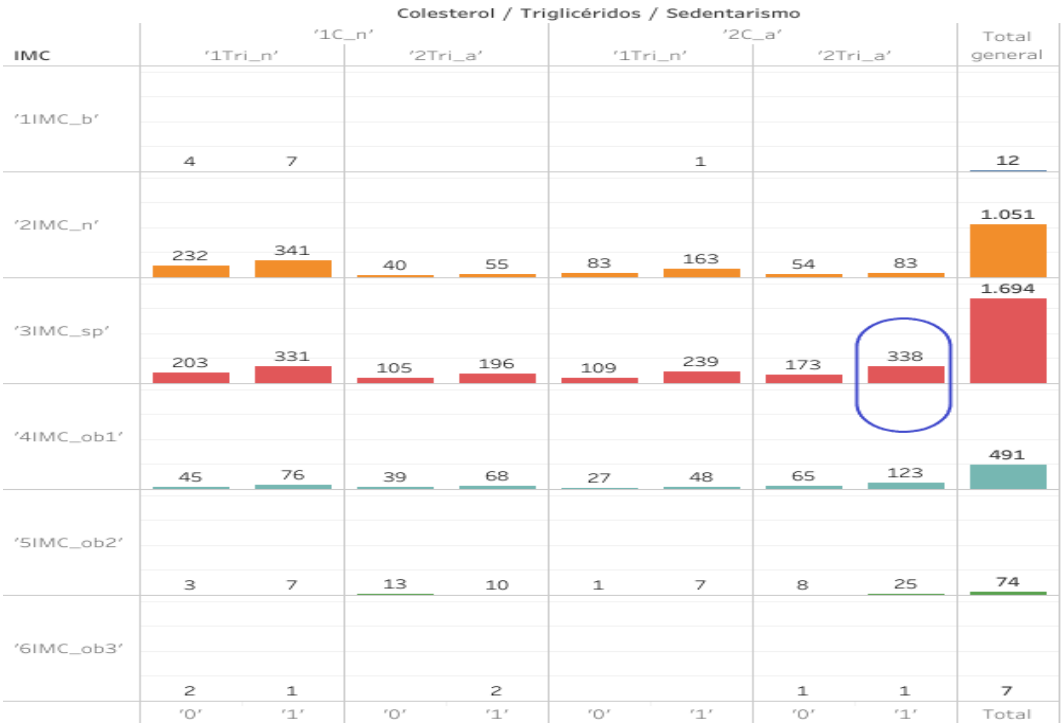


Figura 11. Distribución de la población del grupo Explotación de Minas de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo e IMC.

La elipse señala al grupo de 338 pacientes, que poseen Colesterol Alto, Triglicéridos Alto, es Sedentario y tiene Sobrepeso..

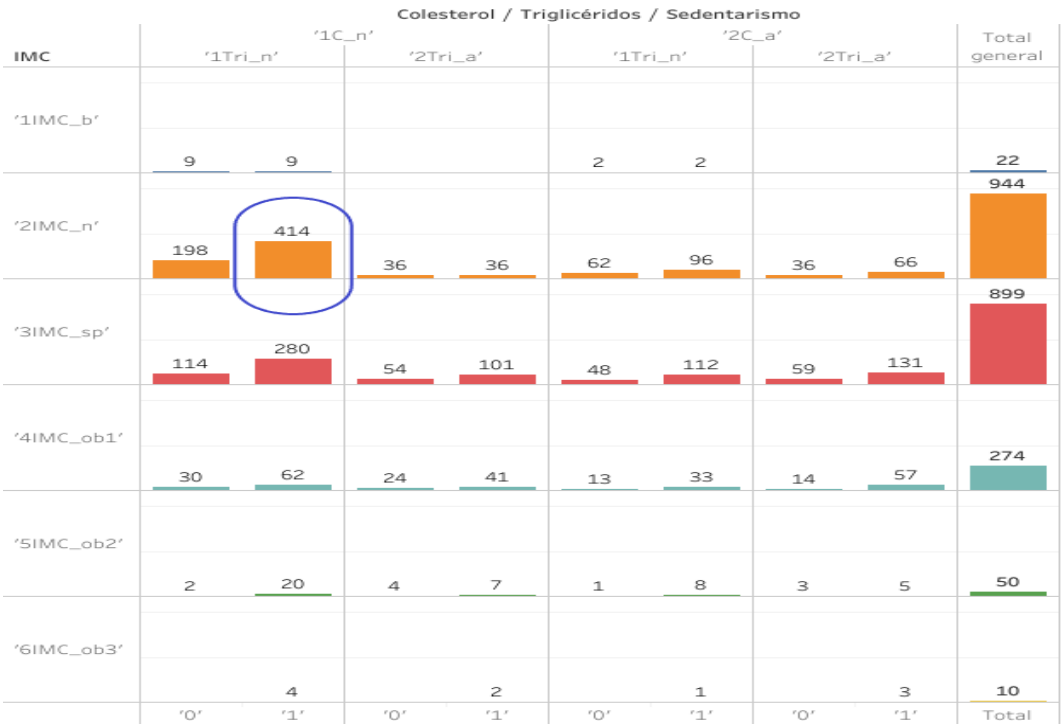


Figura 12. Distribución de la población del grupo Actividades Financieras de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo e IMC.

La elipse señala al grupo de 414 pacientes que poseen Colesterol Normal, Triglicéridos Normal, es Sedentario, e IMC Normal.

El Anexo 3 presenta la distribución de la población de Explotación de Minas, cambiando la variable IMC de la Figura 11 por las variables Sexo, Edad y Educación; respectivamente. Además se considera el número de pacientes de cada grupo así como el porcentaje que representa cada grupo respecto a la población de Explotación de Minas.

El Anexo 4 presenta la distribución de la población de Actividades Financieras, cambiando la variable IMC de la Figura 12 por las variables Sexo, Edad y Educación; respectivamente. Además se considera el número de pacientes de cada grupo así como el porcentaje que representa cada grupo respecto a la población Actividades Financieras.

3.3.2. Factores de riesgo

Considerando el límite de normalidad de cada variable numérica, se construye la correspondiente variable dicotómica; de tal forma que 0, si está dentro del campo de normalidad (no existe riesgo para la salud en esta variable) y, 1 si sobrepasa este límite (esta variable representa un riesgo a la salud). Los límites de normalidad de las variables consideradas se observan en la Tabla 22.

Tabla 22
Límites de normalidad para las variables numéricas.

Variable	Límite de normalidad	Unidades	Categiría 0	Categiría 1
E*	35	años	< 35	≥ 35
IMC	25	kg/m^2	< 25	≥ 25
G	100	mg/dl	< 100	≥ 100
C	200	mg/dl	< 200	≥ 200
Tri	150	mg/dl	< 150	≥ 150
Ps	140	mmHg	< 140	≥ 140
Pd	90	mmHg	< 90	≥ 90

Fuente: [OMS, 2018], [Múnera and Escobar, 2007], [Calle, 2011], [Rotaeche del Campo, 2002].
$*$ Para la variable edad se considera 35 años como límite, ya que a partir de esta edad es necesario tener un control más adecuado de la salud y; además, la media poblacional esta alrededor de ese valor.

A las siete variables dicotómicas construidas, sumamos las variables de Hábitos: Sedentarismo y Tabaquismo, formando de esta manera los nueve factores de riesgo que se analizan. Luego, a partir de esta categorización, cuantificamos el número de factores de riesgo presentes en cada observación, lo que permite formar grupos de pacientes con el mismo número de factores de riesgo. En los grupos formados, se puede determinar cuáles son los factores de riesgo más comunes o relevantes.

Las variables cualitativas como Sexo, Educación, Actividad Económica, etc., permiten realizar análisis a grupos específicos de interés, lo que faculta hacer comparaciones entre diversos grupos.

Un segmento de la data, donde se indica los factores de riesgo, se observa en la Tabla 23; por ejemplo, para la observación remarcada con ID 18, las variables: Edad, Tabaquismo, IMC, Colesterol, Glucosa, Presión Sistólica y Presión Diastólica tienen el valor de 1 y las variables Sedentarismo y Triglicéridos presentan el valor de 0. La columna *Suma* indica el total de factores de riesgo de cada observación, para el ejemplo, la observación 18 posee 7 factores de riesgo pues su suma es 7.

Tabla 23
Segmento de la Data con variables en forma dicotómica.

ID	Edad	S	T	IMC	Colesterol	Triglicéridos	Glucosa	Ps	Pd	Suma
1	1	1	1	0	1	1	1	0	0	6
2	1	0	1	1	1	0	0	0	0	4
3	1	1	1	1	0	1	1	0	0	6
5	1	1	1	1	1	0	0	0	0	5
6	1	0	1	1	0	0	0	0	0	3
10	1	1	1	1	1	0	1	0	0	6
11	1	1	1	1	1	1	0	0	1	7
12	1	1	1	0	0	0	0	0	0	3
13	1	1	1	0	1	1	0	0	0	5
16	1	0	1	1	0	1	1	0	0	5
17	1	0	1	0	1	0	0	0	0	3
18	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**7**
19	1	1	1	0	1	1	1	0	0	6
20	1	1	1	1	1	1	0	1	1	8
23	1	1	1	1	1	1	0	0	0	6
24	1	1	1	1	0	1	1	0	1	7
26	1	1	1	1	1	0	1	0	0	6
⋮										

ID=Identificación, S=Sedentarismo, T=Tabaquismo, IMC=Índice de Masa Corporal, Ps= Presión Sistólica, Pd=Presión Diastólica.

La determinación del número de factores de riesgo que tiene cada observación, permite clasificar a la población total de estudio de acuerdo a este parámetro, obteniéndose grupos de observaciones con el mismo número de factores de riesgo.

Definición 4. *Los grupos de riesgo denominados* $C_i, \quad i = 0, 1, ..., 9$*, corresponden a grupos poblacionales que poseen el mismo número de factores de riesgo, los mismos que pueden ser cualquiera de las nueve variables analizadas; de tal forma que:*

- C_1 *representa al grupo que posee un factor de riesgo,*
- C_2 *representa al grupo que posee dos factores de riesgo,*

  ⋮
- C_9 *representa al grupo que posee nueve factores de riesgo.*

La distribución de la población total de acuerdo a los grupos de riesgo C_i, se visualiza en la Figura 13, la que indica que el grupo C_3 es el que posee mayor población.

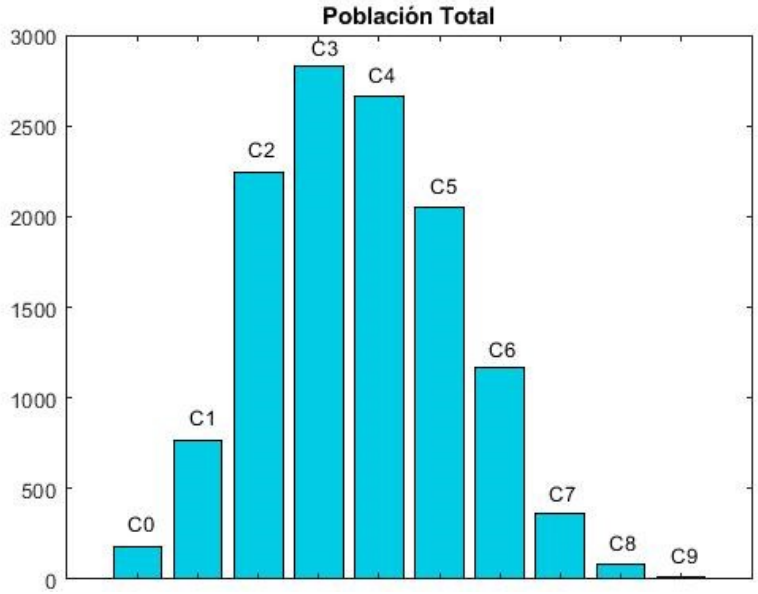


Figura 13. Distribución de la población total de acuerdo a los grupos de riesgo (C_i).

La Tabla 24 presenta la distribución de la población total por actividad económica y por número de factores de riesgo. Los porcentajes indicados se refieren al total de la población por cada actividad económica; así, por ejemplo, el grupo C_3, de la actividad económica Explotación de Minas, contiene 661 pacientes, que equivale al 19.85 % respecto al total de esta actividad (3 329 pacientes), y cada elemento de este grupo posee tres factores de riesgo.

La Figura 14 presenta los histogramas de la distribución de la población por número de factores de riesgo para cada una de las actividades económicas de la Tabla 24.

Tabla 24
Distribución de la población por actividad económica y número de factores de riesgo.

Grupos de riesgo	Explotacion de Minas		Actividades Financieras		Construcción		Actividades Profesionales		Otras actividades		**Total**
	población	%	población	%	población	%	población	%	población	%	
C_0	5	0.15	16	0.72	1	0.06	9	0.57	149	4.05	**180**
C_1	155	4.65	142	6.45	22	1.37	76	4.87	369	10.03	**764**
C_2	437	13.12	490	22.28	289	18.08	267	17.13	764	20.77	**2 247**
C_3	661	19.85	579	26.33	396	24.78	353	22.65	844	22.94	**2 833**
C_4	771	23.16	477	21.69	344	21.52	346	22.20	722	19.62	**2 660**
C_5	674	20.24	326	14.82	290	18.14	292	18.74	470	12.78	**2 052**
C_6	458	13.75	122	5.54	178	11.13	143	9.17	267	7.26	**1 168**
C_7	142	4.26	32	1.45	63	3.94	55	3.53	74	2.01	**366**
C_8	23	0.69	12	0.54	12	0.75	15	0.96	18	0.49	**80**
C_9	3	0.09	3	0.13	3	0.18	2	0.12	2	0.05	**13**
Total	**3 329**	**100.00**	**2 199**	**100.00**	**1 598**	**100.00**	**1 558**	**100.00**	**3 679**	**100.00**	**12 363**

C_i = grupos con i factores de riesgo.

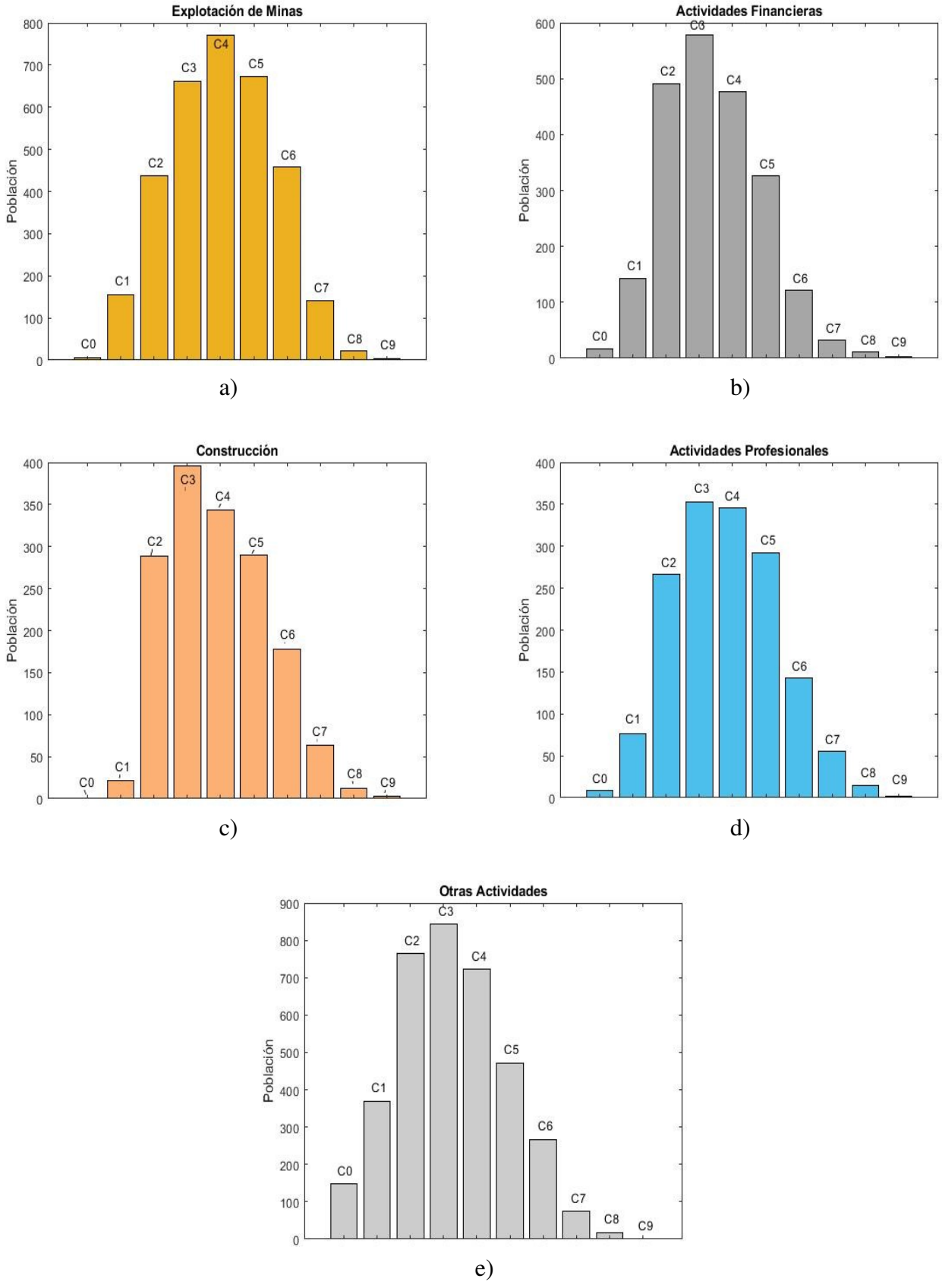


Figura 14. Histogramas de las actividades económicas por número de factores de riesgo.

El número de posibles subgrupos con tres factores de riesgo de los nueve definidos, esta dado por el número de combinaciones

$$\binom{9}{3} = 84,$$

todas estas opciones se visualizan en la Tabla 25.

Tabla 25

Subgrupos definidos por tres factores de riesgo.

1	2	3	4	5	6	7	Total
E IMC C	E C Tri	E Tri G	E G Ps	E Ps Pd	E Pd S	E S T	7
E IMC Tri	E C G	E Tri Ps	E G Pd	E Ps S	E Pd T		6
E IMC G	E C Ps	E Tri Pd	E G S	E Ps T			5
E IMC Ps	E C Pd	E Tri S	E G T				4
E IMC Pd	E C S	E Tri T					3
E IMC S	E C T						2
E IMC T							1
IMC C Tri	IMC Tri G	IMC G Ps	IMC Ps Pd	IMC Pd S	IMC S T		6
IMC C G	IMC Tri Ps	IMC G Pd	IMC Ps S	IMC Pd T			5
IMC C Ps	IMC Tri Pd	IMC G S	IMC Ps T				4
IMC C Pd	IMC Tri S	IMC G T					3
IMC C S	IMC Tri T						2
IMC C T							1
C Tri G	C G Ps	C Ps Pd	C Pd S	C S T			5
C Tri Ps	C G Pd	C Ps S	C Pd T				4
C Tri Pd	C G S	C Ps T					3
C Tri S	C G T						2
C Tri T							1
Tri G Ps	Tri Ps Pd	Tri Pd S	Tri S T				4
Tri G Pd	Tri Ps S	Tri Pd T					3
Tri G S	Tri Ps T						2
Tri G T							1
G Ps Pd	G Pd S	G S T					3
G Ps S	G Pd T						2
G Ps T							1
Ps Pd S	Ps S T						2
Ps Pd T							1
Pd S T							1
			Total				84

Grupos formados con tres factores de riesgo, considerando el orden E, IMC, C, Tri, G, Ps, Pd, S, T. Las filas y columnas indican las secuencias de formación de estos grupos para el orden indicado.

Aplicando la metodología para la determinación de los factores de riesgo en la población de estudio se obtienen grupos con diversos factores, lo cual nos permite determinar como se encuentra el estado de salud de grupos con características específicas, así; por ejemplo, la Figura 15, representan la distribución de la población total considerando las variables dicotómicas Colesterol, Triglicéridos, Sedentarismo, Tabaquismo, IMC y la variable Actividad Económica de múltiples categorías.

En la Figura 15, el primer nivel de ceros y unos, en la parte superior del eje horizontal, corresponde a la variable Colesterol, el segundo nivel a la variable Triglicéridos, el tercer nivel corresponde a la variable Sedentarismo, el cuarto nivel a la variable Tabaquismo y el quinto nivel ubicado en la parte inferior, corresponde a la variable IMC, es decir, los niveles mantienen el orden indicado en la parte superior de esta figura.

En la Figura 15 se observa la distribución de la población de acuerdo a la presencia o no de los factores de riesgo: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC representadas en el ***eje horizontal***, según las actividades económicas relevantes expresadas en el ***eje vertical***. En esta figura, por ejemplo, la elipse señala al grupo de condición más crítica donde todas las variables analizadas son factores de riesgo. La distribución de la población de este grupo por actividades económicas es: 487 pacientes en Explotación de Minas, 218 en Construcciones, 196 en Actividades Financieras, 200 en Actividades Profesionales y 326 en otras actividades, dando un total de la población de este grupo de 1 427 pacientes de una población total de 12 363.

El Anexo 5, indica la distribución de la población total de acuerdo a la presencia o no de los factores de riesgo para las variables Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Actividad Económica, intercambiando las variables IMC, Sexo, Edad y Educación; tanto por número de pacientes como por porcentajes respecto a la población total.

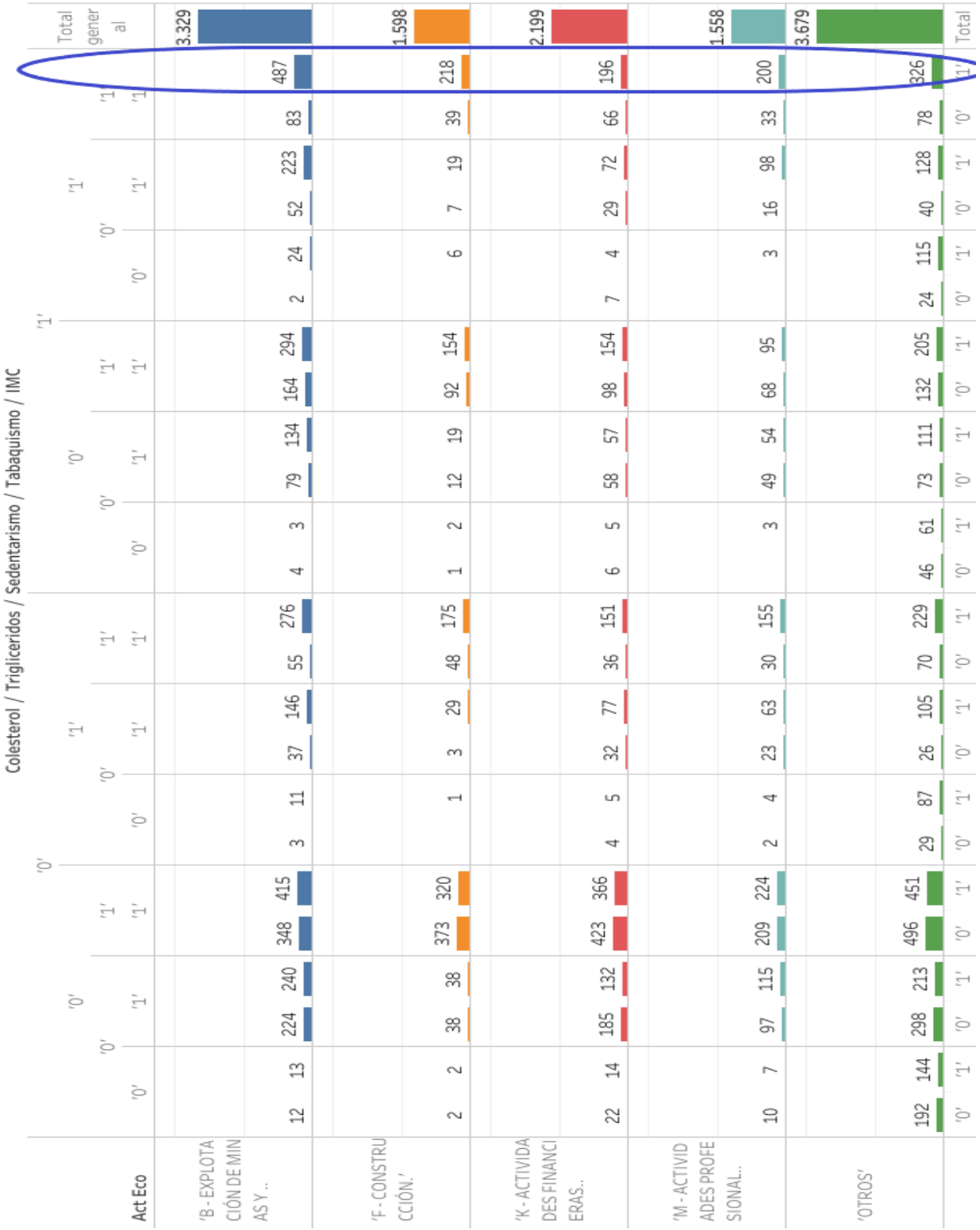


Figura 15. Distribución de la población total con variables dicotómicas: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC.

La elipse señala al grupo de condición más crítica donde todas las variables analizadas son factores de riesgo. La distribución de la población de este grupo por actividades económicas relevantes es: 487 pacientes en Explotación de Minas, 218 en Construcciones, 196 en Actividades Financieras, 200 en Actividades Profesionales y 326 en otras actividades, dando un total de la población de este grupo de 1 427 pacientes de una población total de 12 363.

Factores de riesgo por actividad económica

Consideramos que la actividad económica en la que labora el paciente puede influir en su estado de salud; por tanto, se realiza un análisis de dos de ellas: Explotación de Minas, que tiene principalmente trabajo de campo y, Actividades Financieras que es, mayoritariamente, un trabajo de oficina. Estas actividades además de poseer diferentes condiciones de trabajo, presentan las mayores poblaciones.

Grupo Explotación de Minas

La actividad económica Explotación de Minas, representa el grupo de mayor población, con 3 329 pacientes (ver Tabla 24), sobre este grupo se presenta un análisis más detallado acerca de la cantidad de factores de riesgo que afectan a los pacientes, así como su influencia en el estado de salud y el riesgo de tener una determinada patología.

La distribución de la población por número de factores de riesgo se observa en la Figura 16, donde los grupos más representativos corresponden a los que poseen 3, 4 y 5 factores de riesgo, en particular, representan al 20 %, 23 % y 20 % de los pacientes, respectivamente.

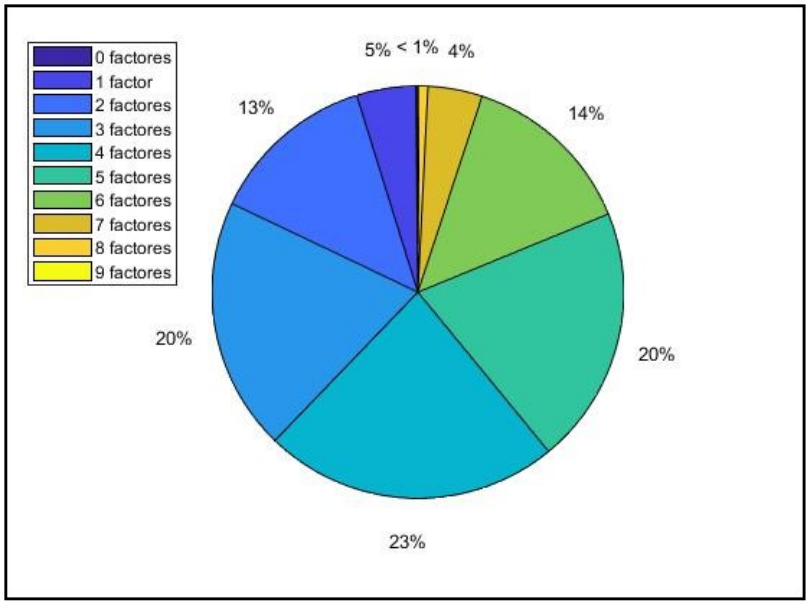


Figura 16. Distribución por factores de riesgo de la población correspondiente al grupo Explotación de Minas.

La Figura 17, presentan la distribución de la población de este grupo de acuerdo a los factores de riesgo: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC; en esta figura se observa, por ejemplo que 487 personas el 14.63 % tienen valores elevados de Colesterol y Triglicéridos, así

como hábitos de Sedentarismo y Tabaquismo; adicionalmente, tienen un IMC superior al normal.

En la Figura 17, la leyenda de la parte superior indica el orden de los diversos niveles de ceros y unos del eje horizontal; así, el primer nivel corresponde a la variable Colesterol, el segundo a Triglicéridos, el tercer nivel al hábito Sedentarismo y el cuarto nivel, en la parte inferior del gráfico a la variable Tabaquismo. En el eje vertical se observa las categorías de IMC.

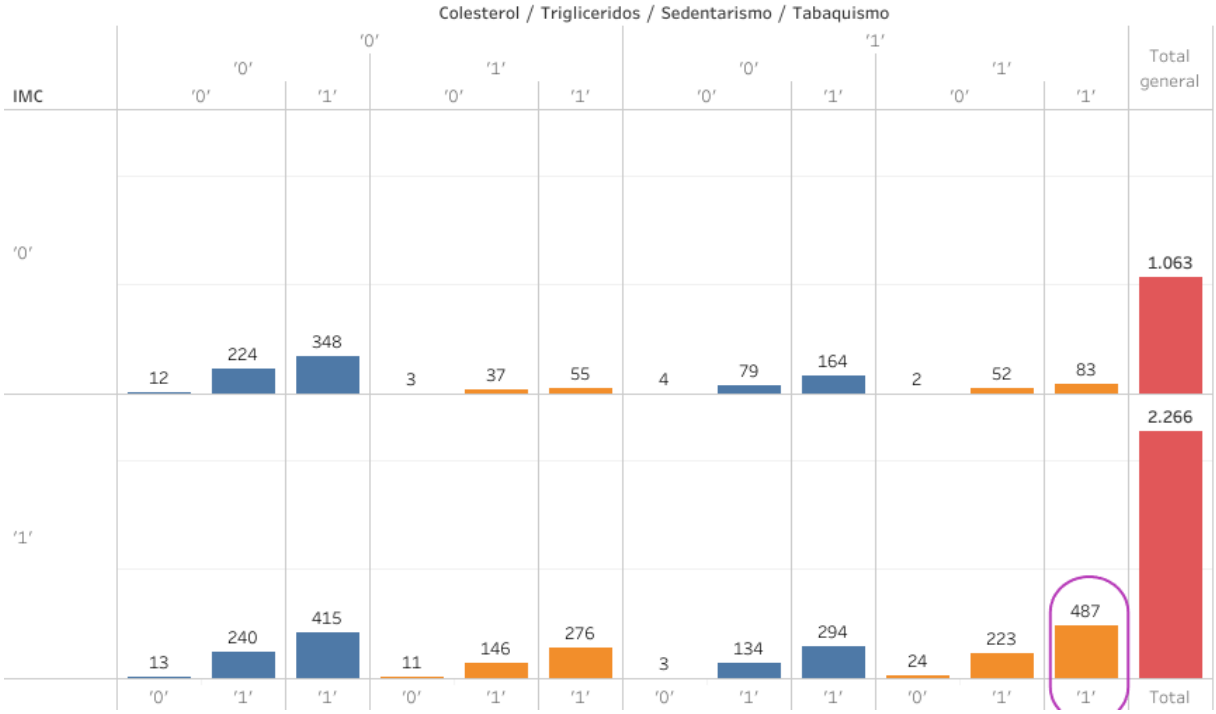


Figura 17. Distribución del grupo Explotación de Minas con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC.
La elipse señala al grupo de condición más crítica donde todas las variables analizadas son factores de riesgo. La población de este grupo es: 487 pacientes.

Grupo Actividades Financieras

La población del grupo Actividades Financieras es 2 199 pacientes, su distribución de acuerdo al número de factores de riesgo se observa en la Figura 18.

La Figura 19, presenta la distribución de la población del grupo Actividades Financieras de acuerdo a los factores de riesgo Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC. En esta figura se observa por ejemplo que el Sedentarismo y el Tabaquismo son factores de riesgo a tener en cuenta en este grupo poblacional; además, se indica que 1 233 pacientes que representa el 56 % de su población han rebasado el límite de peso normal.

En la Figura 19, la leyenda de la parte superior indica el orden de los diversos niveles de ceros

y unos del eje horizontal; así, el primer nivel corresponde a la variable Colesterol, el segundo a Triglicéridos, el tercer nivel al hábito Sedentarismo y el cuarto nivel, en la parte inferior del gráfico, a la variable Tabaquismo. En el eje vertical se observa las categorías de IMC.

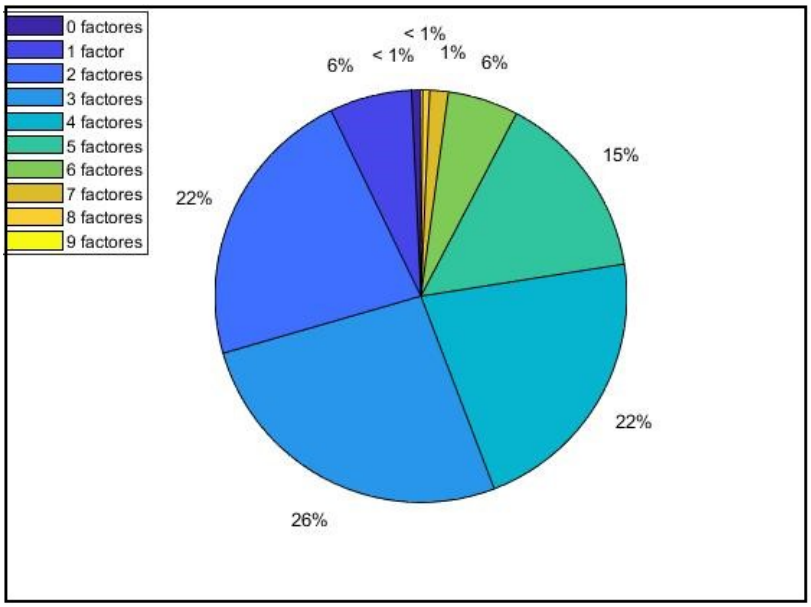


Figura 18. Distribución por factores de riesgo de la población correspondiente al grupo Actividades Financieras.

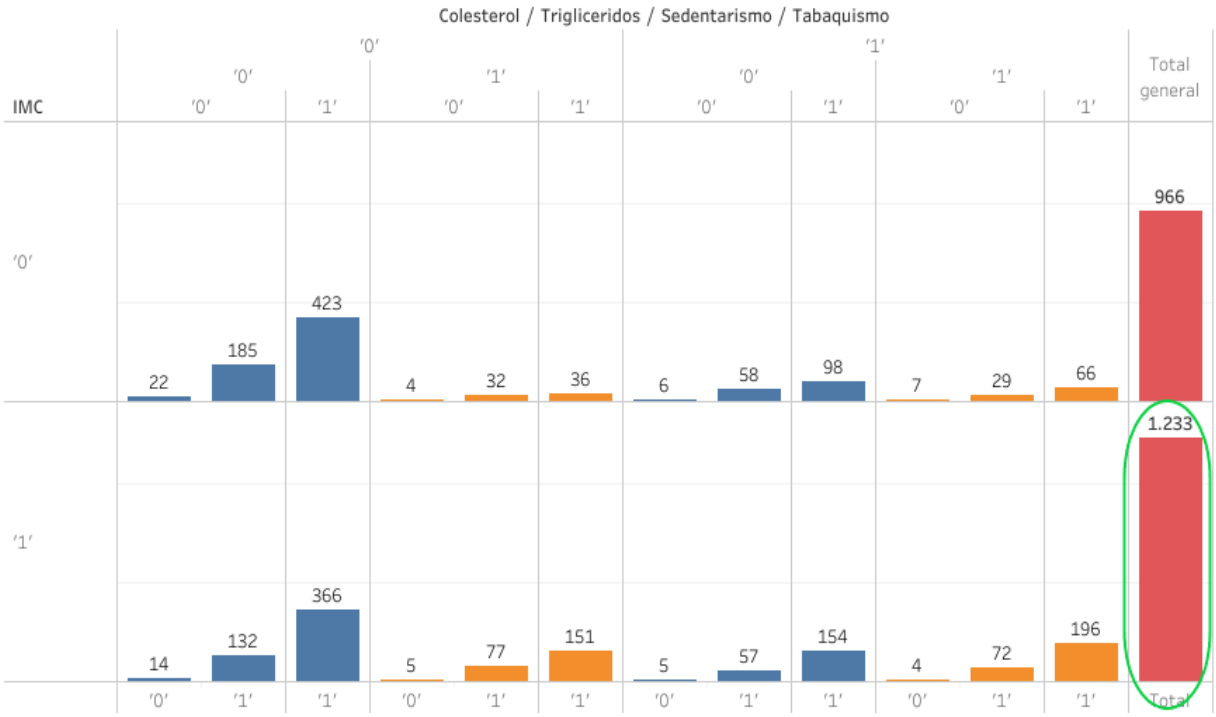


Figura 19. Distribución del grupo Actividades Financieras con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC.

La elipse señala que 1 233 pacientes de un total de 2 199 tiene alguna condición de sobrepeso.

3.3.3. Escalado multidimensional

Para aplicar la técnica de escalado multidimensional se escogen dos grupos de actividades económica, Explotación de Minas y Actividades Financieras; ambos con tres factores de riesgo. El hecho de seleccionar estos grupos se debe a que poseen la mayor población, dentro de la base analizada, el motivo de considerar solo tres factores de riesgo, se debe a que los mismos ya representan un número adecuado de factores (más de dos factores ya hay riesgo) y que ese número nos permite un mejor análisis y visualización.

En estos grupos las variables numéricas se transforman en variables dicotómicas, más las variables Sedentarismo y Tabaquismo, que también serán dicotómicas; así se forman las nueve variables de interés, las que se detallan en la Tabla 26. Estas se han agrupado de acuerdo al criterio del Doctor Fuster, es decir, se agruparán en factores: Mecánicos, Químicos, de Hábitos y Edad.

Tabla 26
Factores de riesgo y su clasificación.

Número	Factor de Riesgo	Simbología	Tipo	Categoría
1	Índice de Masa Corporal	IMC	Mecánicos	F1
2	Presión Sistólica	Ps	Mecánicos	F1
3	Presión Diastólica	Pd	Mecánicos	F1
4	Colesterol	C	Químicos	F2
5	Triglicéridos	Tri	Químicos	F2
6	Glucosa	G	Químicos	F2
7	Sedentarismo	S	Hábitos	F3
8	Tabaquismo	T	Hábitos	F3
9	Edad	E	Edad	F4

Sobre las variables indicadas, en el grupo de la actividad económica Explotación de Minas con tres factores de riesgo, se aplican los coeficientes de similaridad de: i) Sokal-Michener, ii) Russell-Rao y iii) Rogers-Tanimoto; por ser estos coeficientes los recomendados cuando la data contiene la totalidad de las variables de comparación, como lo es en éste estudio; así se obtendrán matrices de similaridades y a partir de estas, matrices de distancias [Baíllo and Grané, 2008].

En las matrices de distancias aplicando el proceso de escalado multidimensional, se representa cada una de las observaciones del grupo mediante dos nuevas variables ortogonales, pudiendo visualizar a cada una de las observaciones como puntos en un plano. El gráfico presenta las similaridades entre los elementos del grupo; así, mientras más cercanos se encuentren las observaciones, mayor es su similaridad. Cada una de estas nuevas variables ortogonales son en realidad una com-

binación lineal de las variables originales sin que ninguna de estas tenga una única preponderancia sobre el resto, por lo que en este estudio las llamaremos como Eje horizontal y Eje vertical.

Las Figuras 20, 21 y 22, representan los gráficos de similaridades obtenidos a partir del coeficiente de Sokal-Michener, Russell-Rao y de Rogers-Tanimoto; en estos gráficos muchas observaciones tienen los mismos factores críticos y por tanto están representados por un mismo punto en estos gráficos. Los puntos numerados del uno al cinco indican en orden decreciente los subgrupos con mayor porcentaje de observaciones en cada uno de los gráficos. Las Tablas 27, 28 y 29, indican los factores críticos de esos subgrupos y su porcentaje respecto a la población total del grupo estudiado.

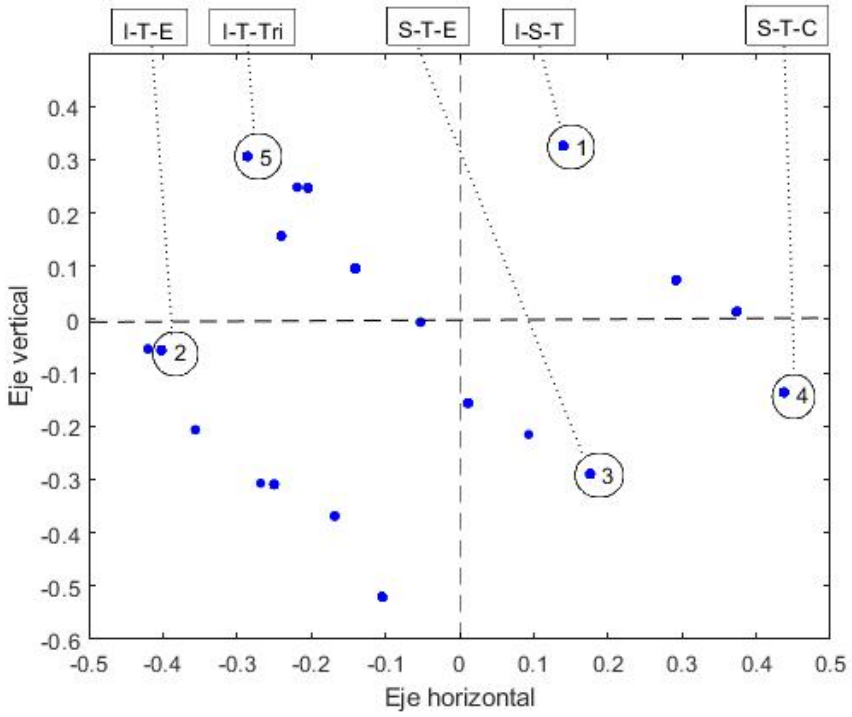


Figura 20. Similaridades del grupo Explotación de Minas, utilizando Sokal-Michener I=IMC, S=Sedentarismo, T=Tabaquismo, E=Edad, C=Colesterol, Tri=Triglicéridos.

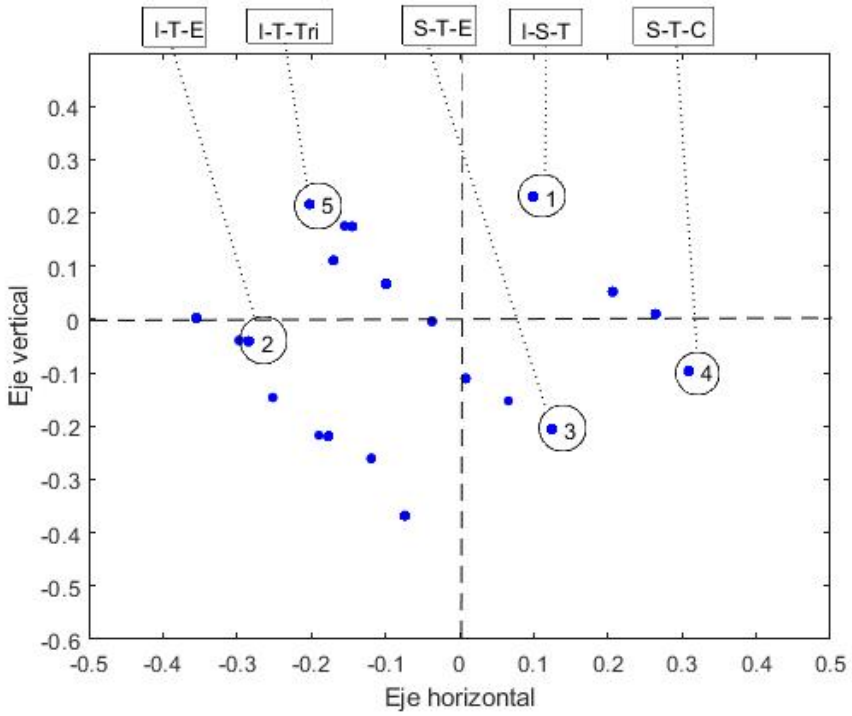


Figura 21. Similaridades del grupo Explotación de Minas, utilizando Russell-Rao
I=IMC, S=Sedentarismo, T=Tabaquismo, E=Edad, C=Colesterol, Tri=Triglicéridos.

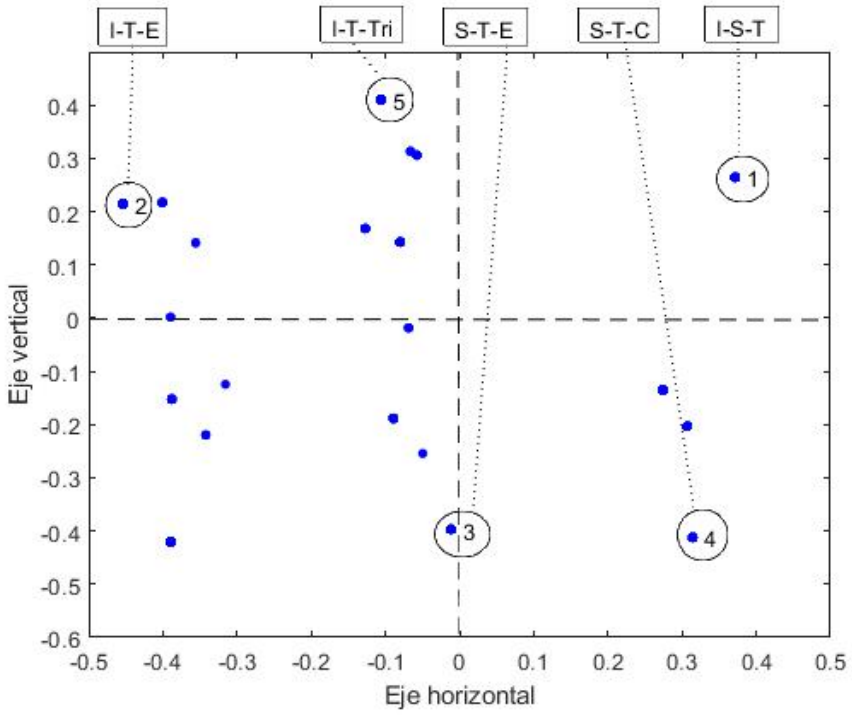


Figura 22. Similaridades del grupo Explotación de Minas, utilizando Rogers-Tanimoto
I=IMC, S=Sedentarismo, T=Tabaquismo, E=Edad, C=Colesterol, Tri=Triglicéridos.

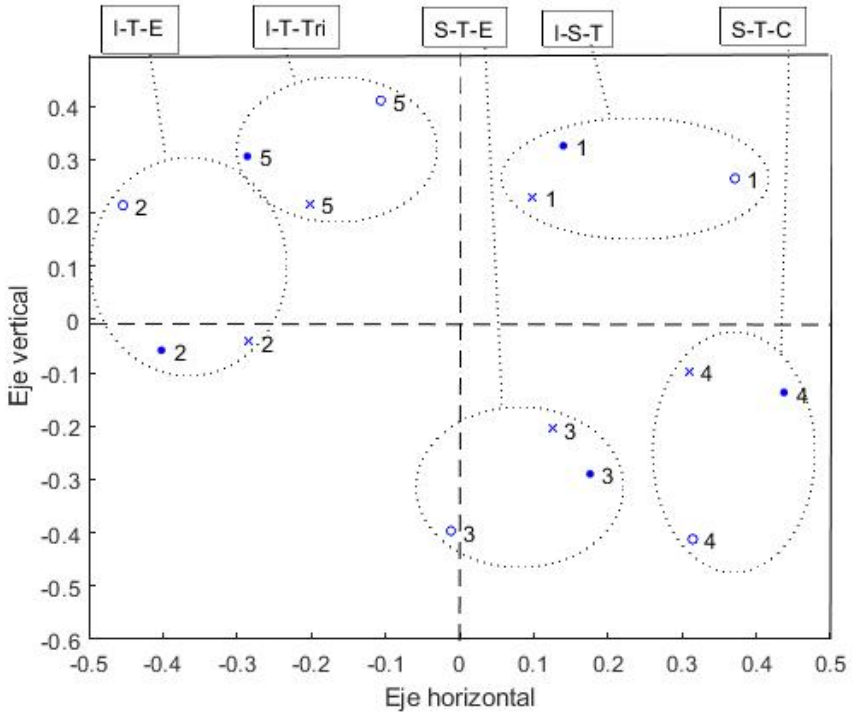


Figura 23. Similaridades del grupo Explotación de Minas, utilizando Sokal-Michener, Russell-Rao y Rogers-Tanimoto para los subgrupos más representativos.

• = Sokal-Michener, o = Russell-Rao, x = Rogers-Tanimoto.

I=IMC, S=Sedentarismo, T=Tabaquismo, E=Edad, C=Colesterol, Tri=Triglicéridos.

Tabla 27

Factores de riesgo de los subgrupos de la actividad Explotación de Minas usando Sokal-Michener.

Observación	Factor 1	Factor 2	Factor 3	Porcentaje
1	IMC	Sedentarismo	Tabaquismo	24
2	IMC	Tabaquismo	Edad	17
3	Sedentarismo	Tabaquismo	Edad	15
4	Colesterol	Sedentarismo	Tabaquismo	10
5	IMC	Triglicéridos	Tabaquismo	5

Tabla 28

Factores de riesgo de los subgrupos de la actividad Explotación de Minas usando Russell-Rao.

Observación	Factor 1	Factor 2	Factor 3	Porcentaje
1	IMC	Sedentarismo	Tabaquismo	24
2	IMC	Tabaquismo	Edad	17
3	Sedentarismo	Tabaquismo	Edad	15
4	Colesterol	Sedentarismo	Tabaquismo	10
5	IMC	Triglicéridos	Tabaquismo	8

Tabla 29

Factores de riesgo de los subgrupos de la actividad Explotación de Minas usando Rogers-Tanimoto.

Observación	Factor 1	Factor 2	Factor 3	Porcentaje
1	IMC	Sedentarismo	Tabaquismo	24
2	IMC	Tabaquismo	Edad	17
3	Sedentarismo	Tabaquismo	Edad	15
4	Colesterol	Sedentarismo	Tabaquismo	10
5	IMC	Triglicéridos	Tabaquismo	8

Los gráficos obtenidos con los tres diferentes coeficientes de similaridad son muy parecidos entre sí, lo que se ratifica con la información proporcionada por las Tablas 27, 28 y 29; que indican los factores críticos de los subgrupos más representativos y su porcentaje respecto a la población total, las cuales para los tres coeficientes de similaridad son identicas a excepción del porcentaje de la observación 5. Por ejemplo, en estas tablas el punto uno, indica que el subgrupo que tiene los factores críticos: IMC, Sedentarismo, Tabaquismo, representa el 24 % de la población del grupo Explotación de Minas con tres factores de riesgo.

Por lo expuesto anteriormente, se observa que el análisis de similaridades es una prueba robusta respecto a los diversos coeficientes; por tanto, para los análisis posteriores se utilizará unicamente el coeficiente de Sokal-Michener.

La Figura 24 y la Tabla 30 indican el gráfico de similaridades y la tabla de los factores críticos de los subrupos más representarivos correspondientes al grupo Actividades Financieras con tres factores de riesgo

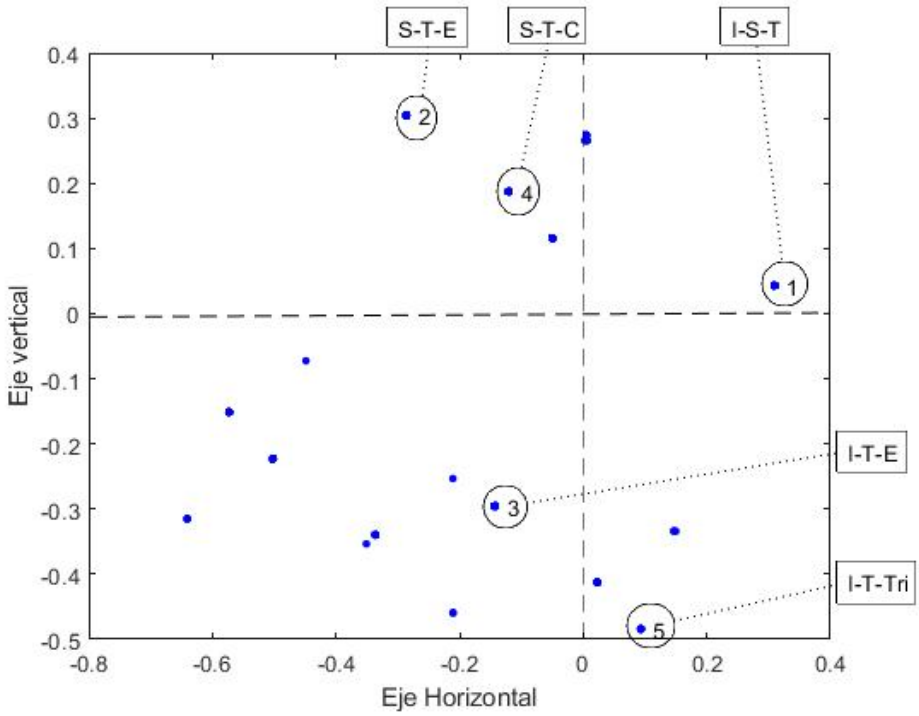


Figura 24. Similaridades del grupo Actividades Financieras, con 3 factores de riesgo. I=IMC, S=Sedentarismo, T=Tabaquismo, E=Edad, C=Colesterol, Tri=Triglicéridos.

Tabla 30
Factores de riesgo de los subgrupos de Actividades Financieras.

Observación	Factor 1	Factor 2	Factor 3	Porcentaje
1	IMC	Sedentarismo	Tabaquismo	39
2	Sedentarismo	Tabaquismo	Edad	17
3	IMC	Tabaquismo	Edad	8
4	Colesterol	Sedentarismo	Tabaquismo	8
5	IMC	Triglicéridos	Tabaquismo	5

Los gráficos de similaridades de los grupos Explotación de Minas y Actividades Financieras presentan ciertas diferencias, las que se visualizan de mejor modo con la información de de las tablas correspondientes a los factores de riesgo de los subgrupos más representativos y el porcentaje de estos respecto a la población total.

Para determinar la influencia de los diversos factores de riesgo, en los gráficos de similaridades de las Figuras 20 y 24, se procede a incluir individuos que únicamente poseen determinadas características. Por ejemplo en las Figuras 25 y 26, el triángulo ▽ representa al paciente que posee únicamente factores mecánicos (IMC, Presión Sistólica, Presión Diastólica), el cuadrado □ al que posee factores químicos (Colesterol, Triglicéridos y Glucosa), el diamante ◇ al que posee solo el factor de Edad, la estrella ★ al que posee factores de hábitos (Sedentarismo y Tabaquismo) y el

asterisco $*$ al individuo que no posee factores de riesgo, es decir, esta "sano" o, de manera particular, los valores de las variables analizadas están dentro de lo considerado normal.

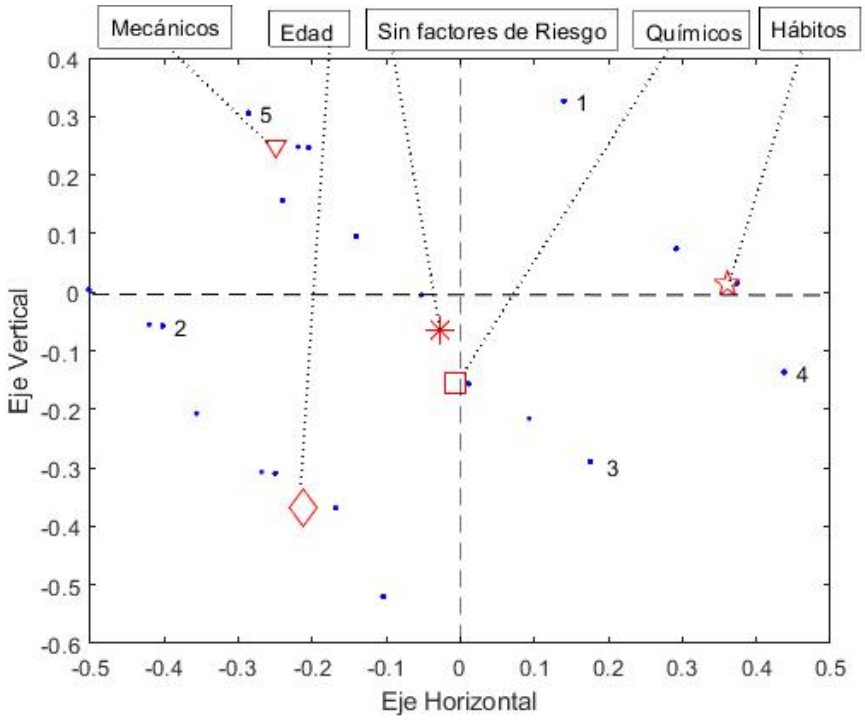


Figura 25. Ubicación de los factores de riesgo para el grupo de Explotación de Minas.

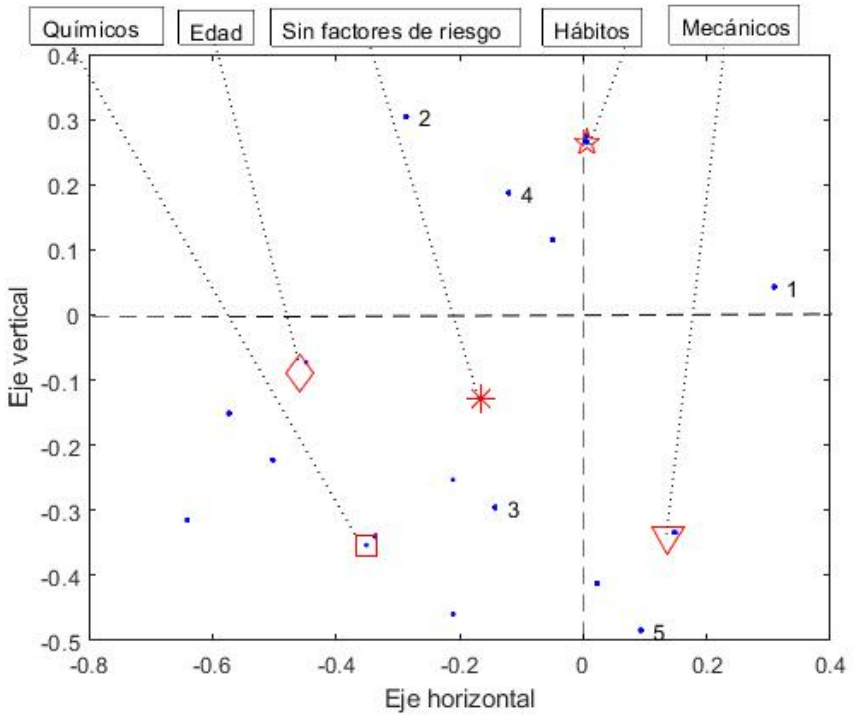


Figura 26. Ubicación de los factores de riesgo para el grupo de Actividades Financieras.

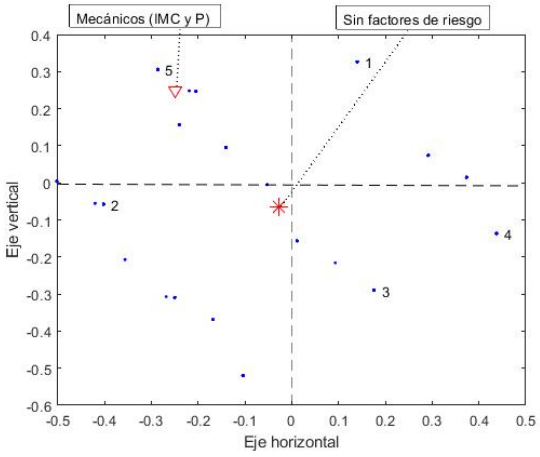
En las Figuras 25 y 26, la observación que no posee factores de riesgo, representa un estado de salud óptimo, se ubica en las zonas centrales de los gráficos, muy cerca del origen de coordenadas, mientras que las observaciones que representan los individuos con diversos tipos de factores como Mecánicos, Químicos, de Hábito, y de Edad, se ubican distantes del centro. Si consideramos la influencia de las variables que conforman este tipo de factores, se determinará cuál de ellas presenta mayor influencia en tener un estado de salud óptimo.

- **Estudio de las variables que conforman los factores Mecánicos**

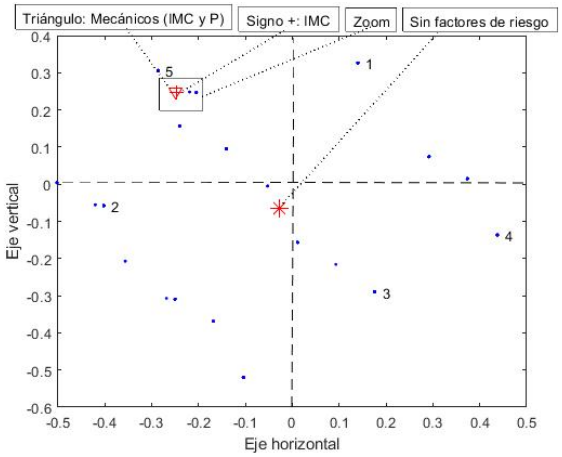
Los factores Mecánicos están constituidos por las variables: IMC, Presión Sistólica y Presión Diastólica, la influencia conjunta de estas dos últimas le llamaremos Presión Arterial; en las Figuras 27 a y 28 a, un paciente que tiene los factores de riesgo Mecánicos está representada por el triángulo $\triangledown$. El asterisco $*$ en estas figuras representa al paciente sin factores de riesgo, es decir, una condición óptima de salud.

En el estudio de los factores de riesgo, se utiliza la expresión *"se elimina el factor de riesgo ..."*, en este contexto, eliminar un factor de riesgo significa que el valor de esa variable retorna al rango de normalidad y por tanto, esta variable ya no representa un riesgo para la salud del paciente.

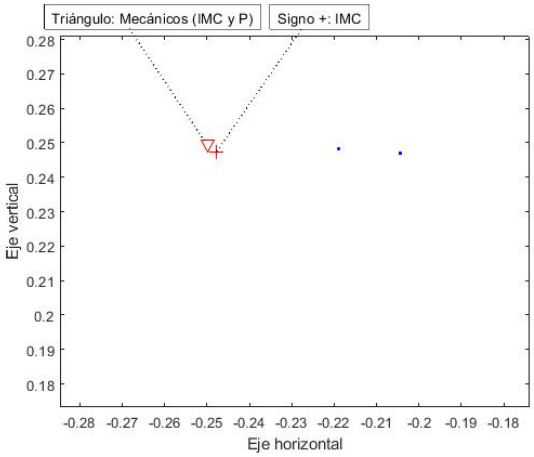
Las Figuras 27 b, 27 c, 28 b y 28 c indican que al eliminar el factor de riesgo Presión Arterial, manteniendo el factor de riesgo IMC, existe un leve desplazamiento del paciente respecto a su posición inicial; en cambio las Figuras 27 d, 27 e, 28 d y 28 e, muestran que al eliminar el factor de riesgo IMC, es decir, lograr que alcance valores menores a 25 kg/m^2, pero manteniendo como factor de riesgo a la Presión Arterial, existe un desplazamiento del paciente hacia la condición de óptima salud. Lo observado hace suponer que en un paciente que tenga los factores de riesgo IMC y Presión Arterial alterados, controlar la variable IMC es más eficiente que hacerlo con la Presión Arterial.



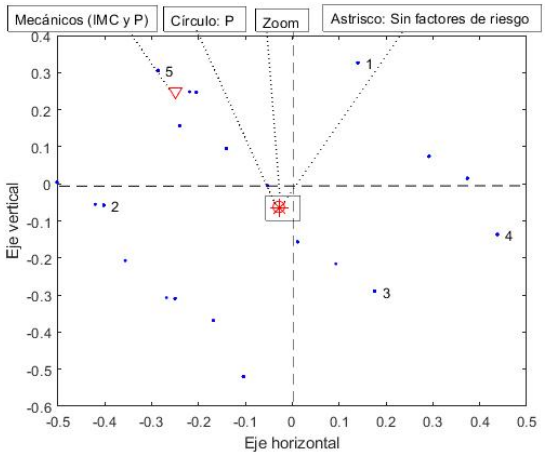
a) Factores Mecánicos



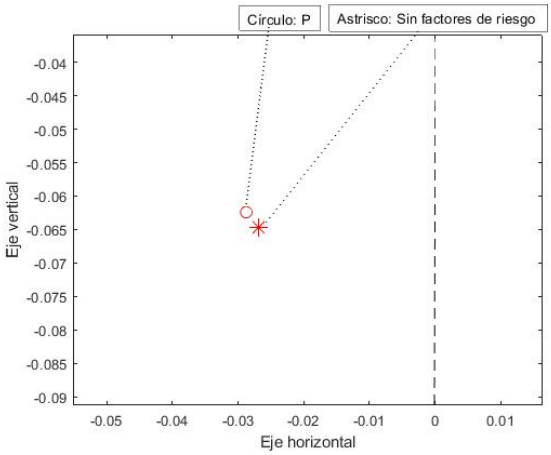
b) Se elimina Presión Arterial



c) Ampliación de b

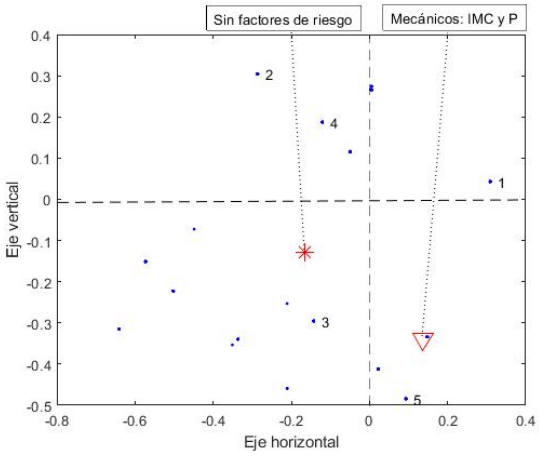


d) Se elimina IMC

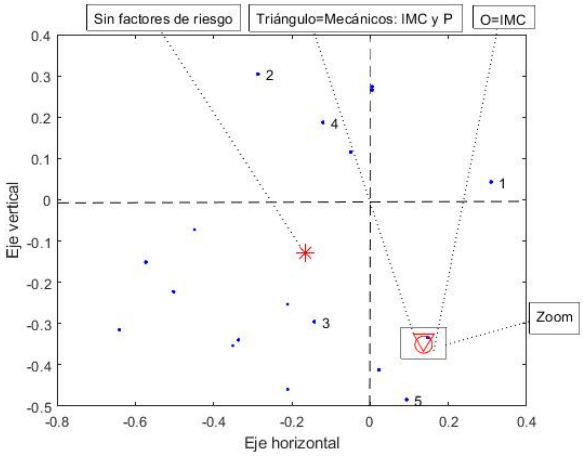


e) Ampliación de d

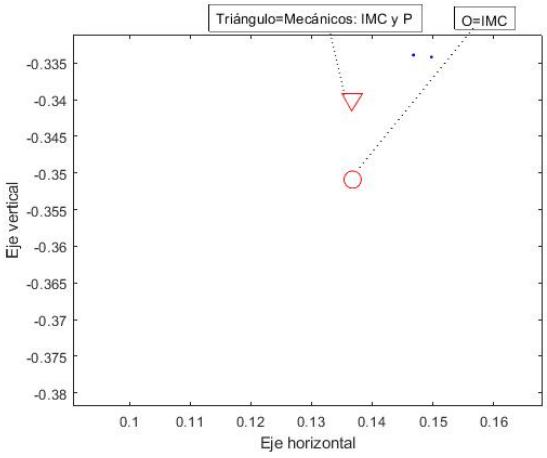
Figura 27. Estudio de las variables que conforman los factores Mecánicos para el grupo de Explotación de Minas.



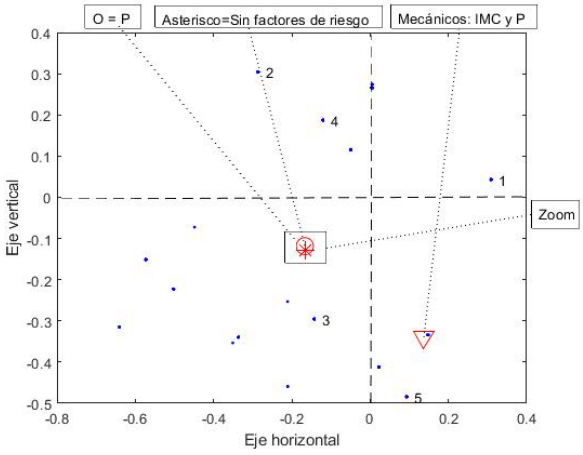
a) Factores Mecánicos



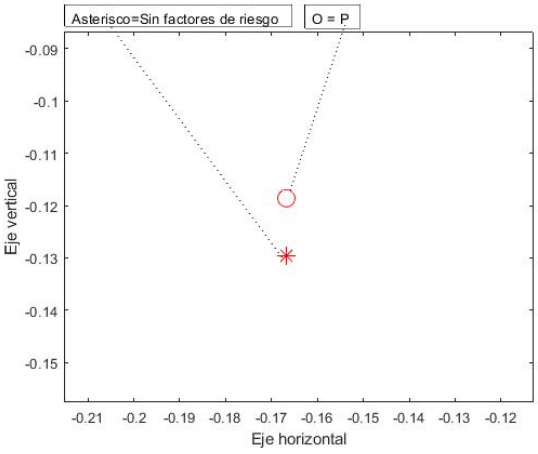
b) Se elimina Presión Arterial



c) Ampliación de b



d) Se elimina IMC



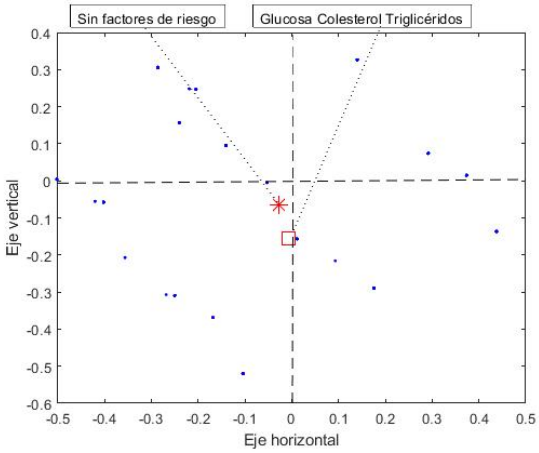
e) Ampliación de d

Figura 28. Estudio de las variables que conforman los factores Mecánicos para el grupo de Actividades Financieras .

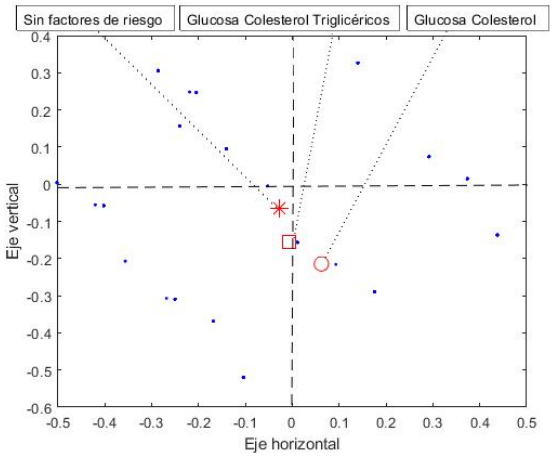
- **Estudio de las variables que conforman los factores Químicos**

Las variables que conforman los factores Químicos son: Colesterol, Triglicéridos y Glucosa. En las Figuras 29 a y 31 a, un paciente que posee los tres factores Químicos está representado por el cuadrado □. El asterisco $*$ en esta figura representa al paciente sin factores riesgo, es decir, una condición óptima de salud. Al suprimir en este paciente, la variable Triglicéridos, Figuras 29 b y 31 b, o la variable Colesterol, Figuras 29 c y 31 c; o las variables Triglicéridos y Glucosa, Figuras 29 d y 31 d; o las variables Colesterol y Glucosa, Figuras 29 e y 31 e; no se alcanza el nivel de salud óptimo.

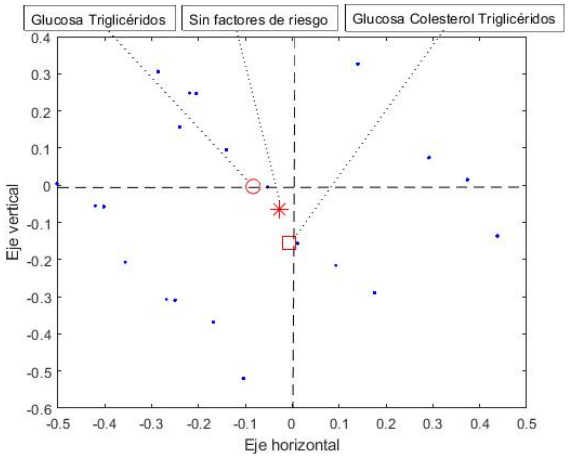
Las Figuras 30 f y 32 f, indican que al eliminar la variable Glucosa, hay un mínimo desplazamiento del paciente, lo que hace suponer que esta variable tiene muy poca influencia para lograr un estado de salud ideal; por otro lado, las Figuras 30 h y 32 h, indican que al eliminar el efecto de las variables Colesterol y Triglicéridos, hay un desplazamiento del paciente, con la variable glucosa, hacia el punto de salud óptima, indicado por el círculo O. Lo observado nos permite suponer que para lograr un estado de salud adecuado en un individuo que presenta los factores de riesgo Químicos, es más eficiente controlar en forma conjunta las variables Colesterol y Triglicéridos.



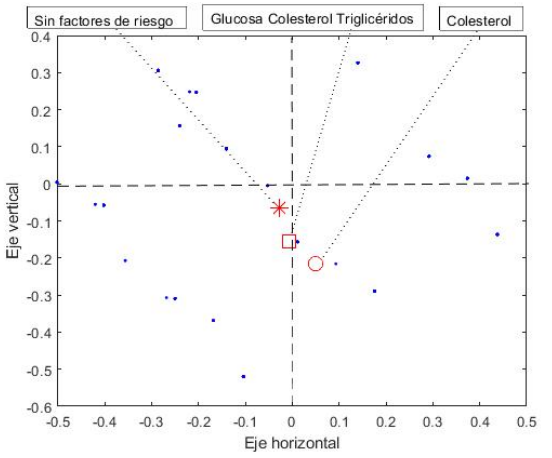
a) Factores Químicos



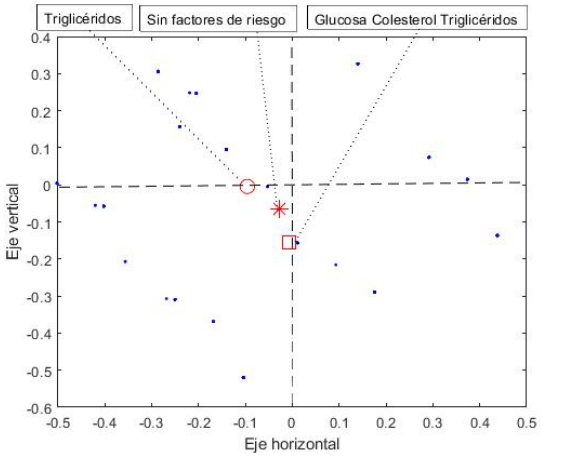
b) Se elimina Triglicéridos



c) Se elimina Colesterol

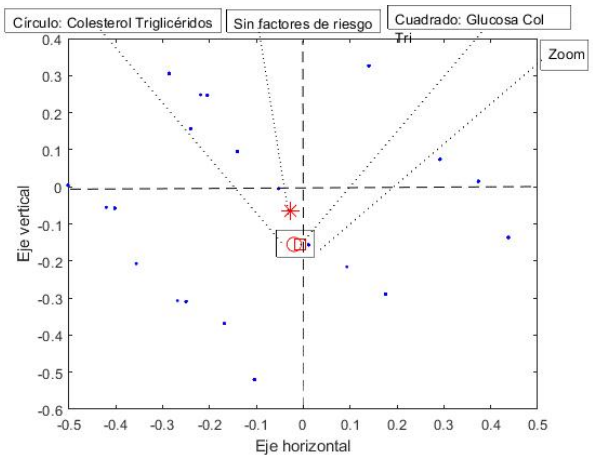


d) Se elimina Triglicéridos y Glucosa

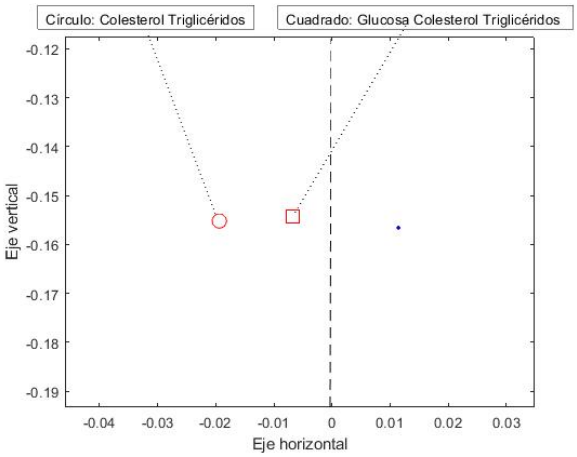


e) Se elimina Colesterol y Glucosa

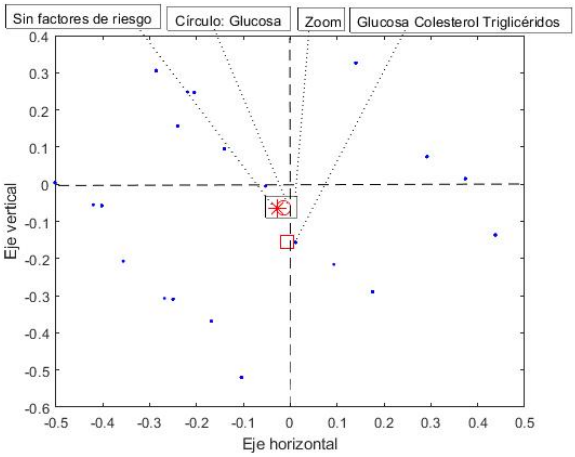
Figura 29. Estudio de las variables que conforman los factores Químicos para el grupo de Explotación de Minas (parte 1).



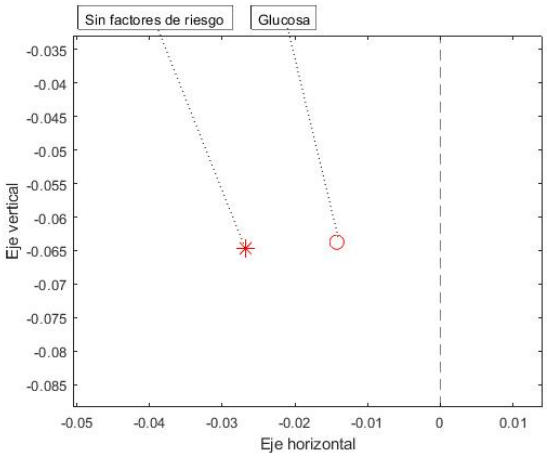
f) Se elimina Glucosa



g) Ampliación de f

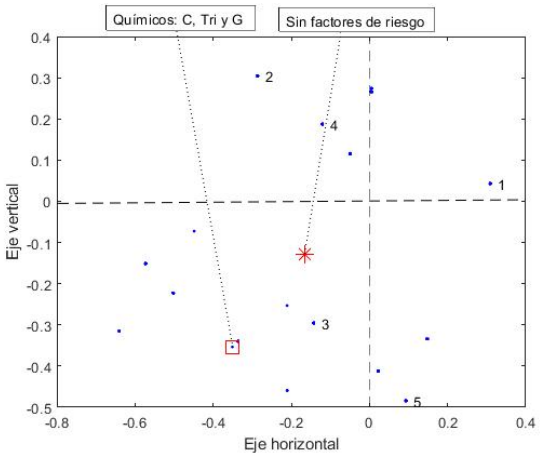


h)Se elimina Colesterol y Triglicéridos

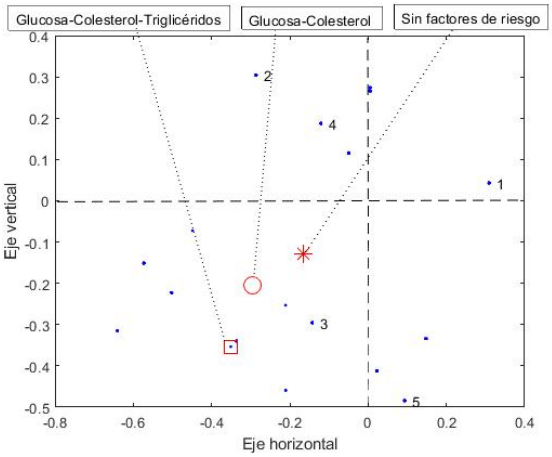


i) Ampliación de h

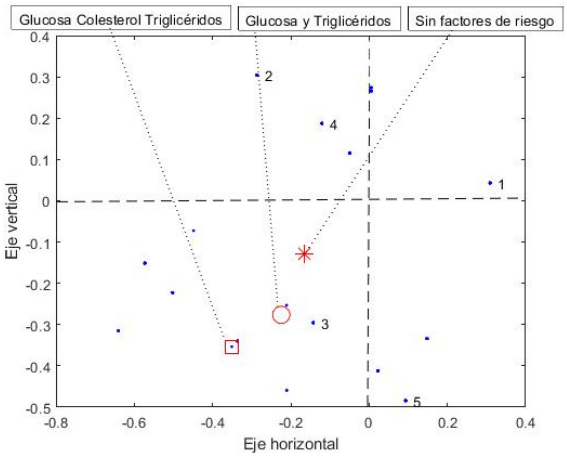
Figura 30. Estudio de las variables que conforman los factores Químicos para el grupo de Explotación de Minas (parte 2).



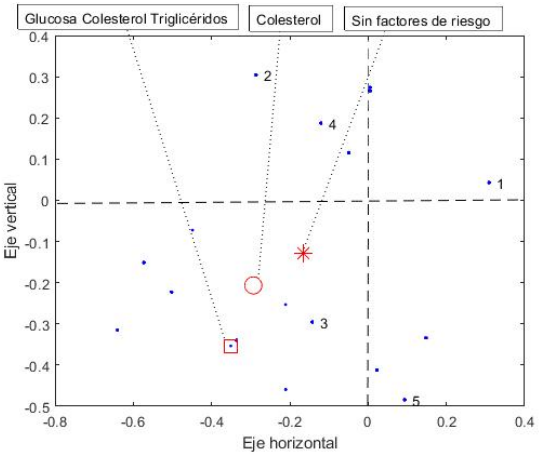
a) Factores Químicos



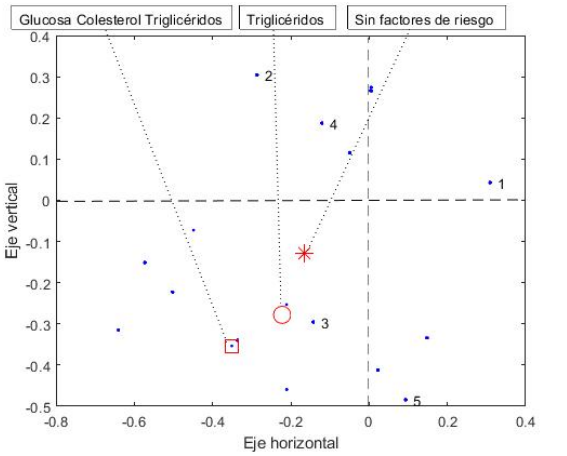
b) Se elimina Triglicéridos



c) Se elimina Colesterol

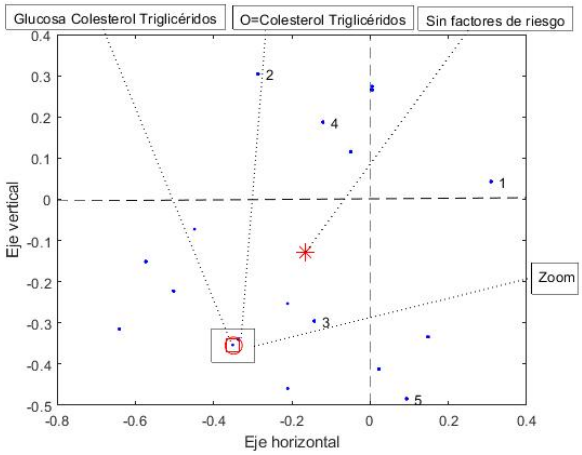


d) Se elimina Triglicéridos y Glucosa

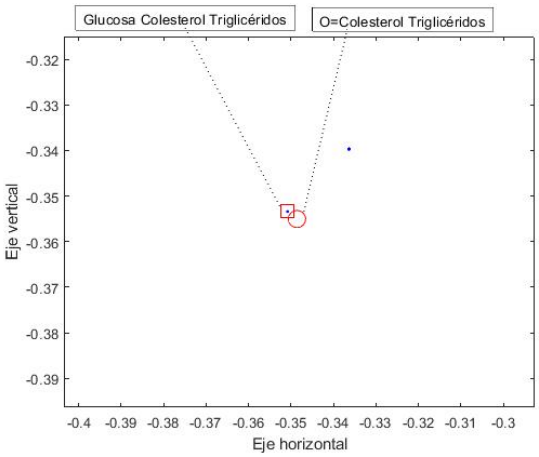


e) Se elimina Colesterol y Glucosa

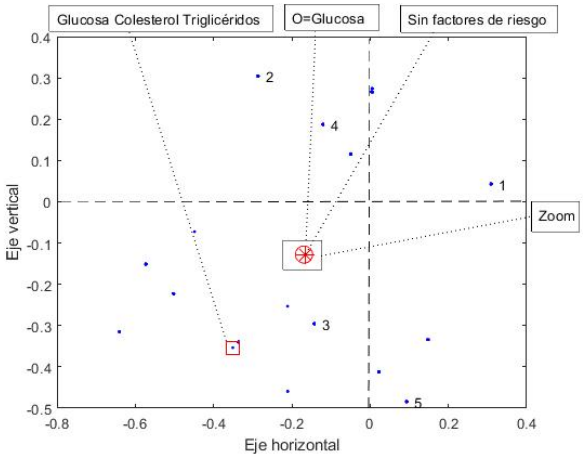
Figura 31. Estudio de las variables que conforman los factores Químicos para el grupo de Actividades Financieras (parte 1).



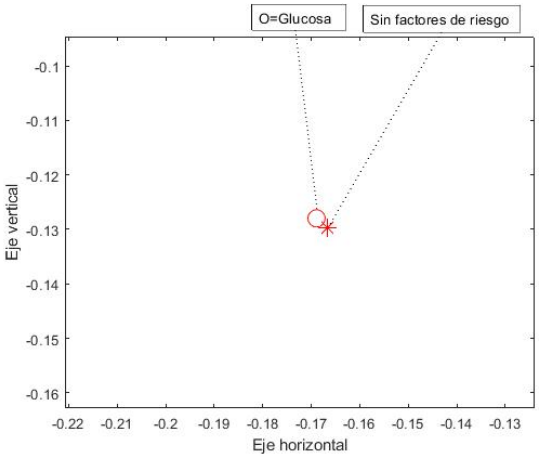
f) Se elimina Glucosa



g) Ampliación de f



h)Se elimina Colesterol y Triglicéridos



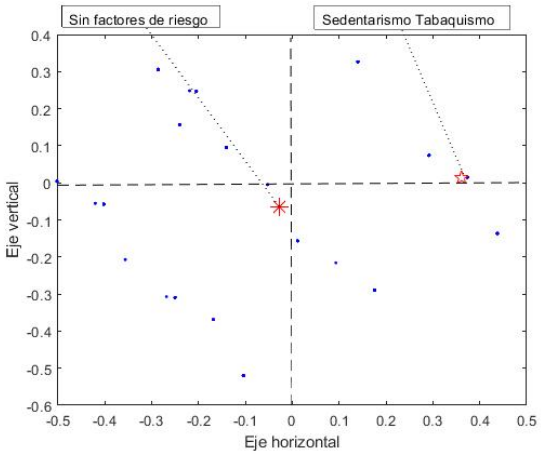
i) Ampliación de h

Figura 32. Estudio de las variables que conforman los factores Químicos para el grupo de Actividades Financieras (parte 2).

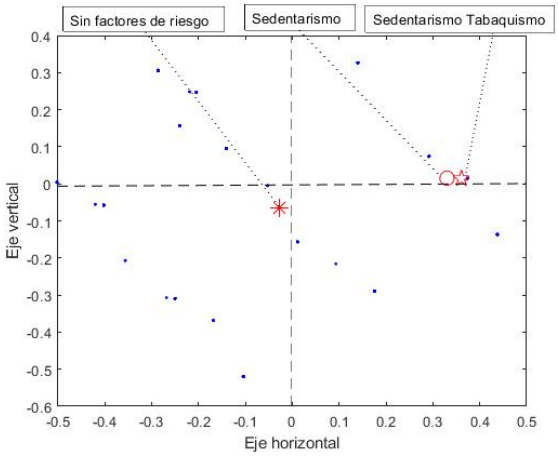
- **Estudio de las variables que conforman los factores de riesgo Hábitos**

Los factores de riesgo Hábitos, conforman las variables Sedentarismo y Tabaquismo. En las Figuras 33 a y 34 a, un paciente que posee los dos factores de riesgo Hábitos esta representado por la estrella ★. El asterisco $*$ en esta figura representa al paciente sin factores riesgo, es decir, una condición óptima de salud. Al suprimir en este paciente, la variable Tabaquismo, como lo indican las Figuras 33 b y 34 b, hay un mínimo desplazamiento del paciente hacia el punto que representa la salud óptima, lo que indica que esta variable tiene poca influencia en lograr un buen estado de salud; por otro lado, al suprimir la variable sedentarismo como lo muestran las Figuras 33 c y 34 c, el paciente alcanza una posición muy cercana al punto que representa un estado de salud óptimo.

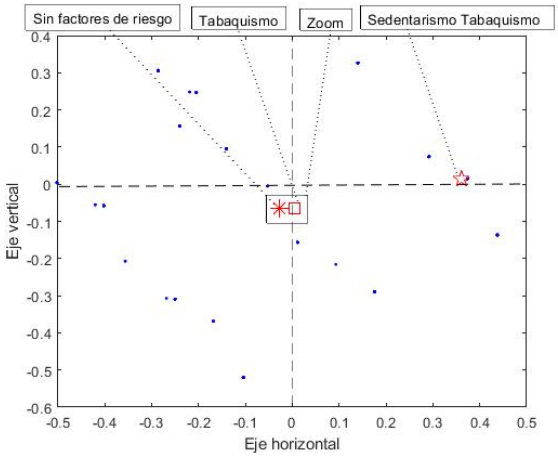
Los resultados antes descritos hacen suponer que, para un paciente que posee los factores de riesgo Hábitos, para tener un adecuado estado de salud, es más eficiente evitar el sedentarismo mediante un adecuado régimen de ejercicios o actividad física.



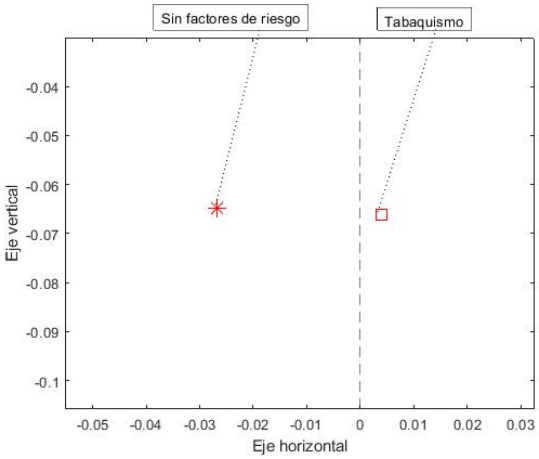
a) Factores de Hábitos



b) Se elimina Tabaquismo



c) Se elimina Sedentarismo



d) Ampliación de c

Figura 33. Estudio de las variables que conforman los factores de riesgo Hábitos para el grupo de Explotación de Minas.

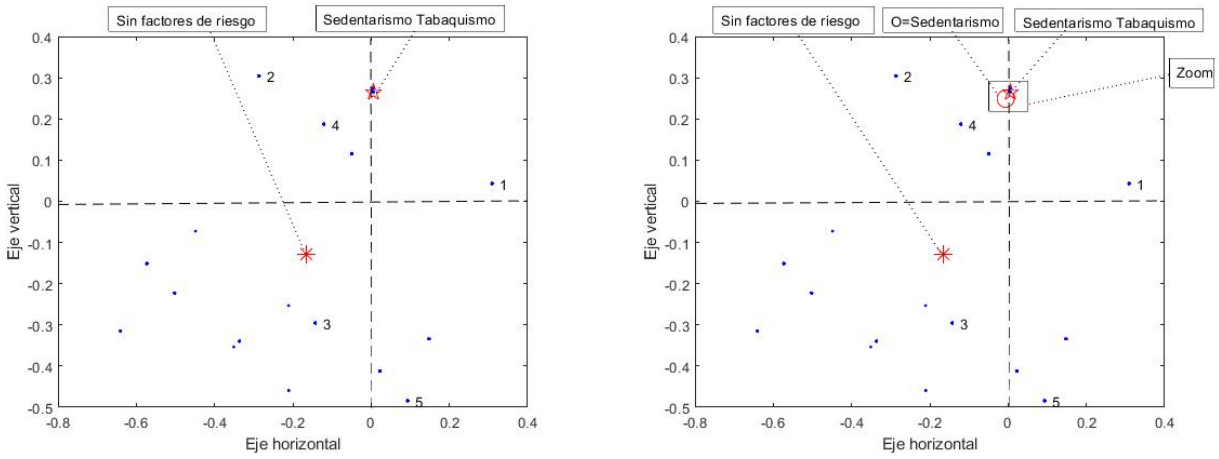


Figura 34. Estudio de las variables que conforman los factores de riesgo Hábitos para el grupo de Actividades Financieras.

- **Estudio de la variable Edad**

Dado que la variable Edad es una condición que no se puede revertir, es importante propiciar un estilo de vida que nos permita controlar los otros factores de riesgo, es decir, mantener un adecuado régimen alimenticio, evitar el sedentarismo, evitar caer en hábitos perjudiciales a la salud, controlar el estrés, etc., y de esta manera lograr un estado de salud adecuado con el pasar de los años, evitando que la edad sea otro factor de riesgo para nuestra salud.

En la Figuras 35 y 36, un paciente que tiene la variable Edad como factor de riesgo, esta representado por el diamante ◇. La variable edad por si sola no representa un riesgo, pero acompañada de otros factores como sobrepeso, altos niveles de glucosa, etc., complican el panorama de la salud de un paciente.

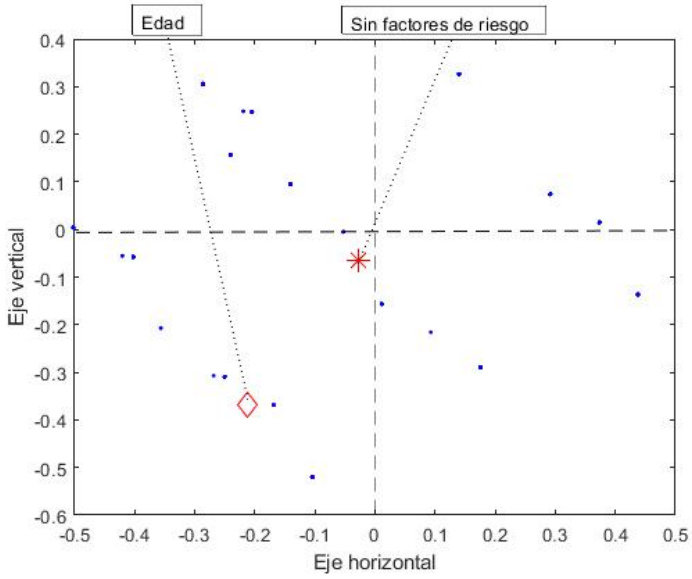


Figura 35. Factor de riesgo Edad para el grupo de Explotación de Minas.

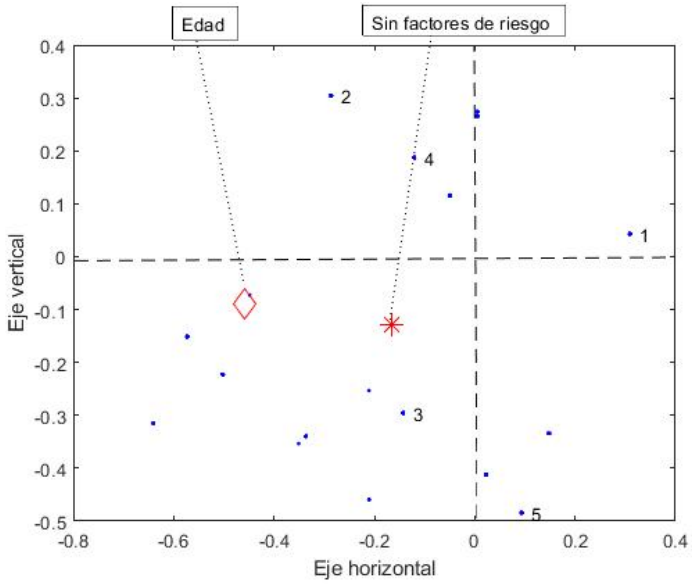


Figura 36. Factor de riesgo Edad para el grupo de Actividades Financieras.

- **Análisis predictivo mediante los factores de riesgo**

Una situación deseable es poder predecir el estado de salud de un paciente en base a los resultados de sus exámenes médicos, resultados que con ayuda del escalado multidimensional se pueden representar como un punto en el plano de similaridades, generado por un grupo determinado; por tanto, la ubicación de dicho punto nos dará información de posibles factores de riesgo presentes y sugerirá opciones para contrarrestarlos.

Para poder predecir un estado de salud en función de la ubicación del paciente en el gráfico de similaridades, es necesario conocer la zona de influencia de cada uno de los factores de riesgo, zona de influencia que queda definida por la ubicación del extremo del vector de cada factor de riesgo.

La representación de un paciente con cierto factor de riesgo, por medio de un vector con el origen en la observación sin factores de riesgo, por ejemplo el vector $\vec{OA}$; define la zona de influencia de cada factor de riesgo, situación que la podemos visualizar en los círculos de las Figuras 37 y 38.

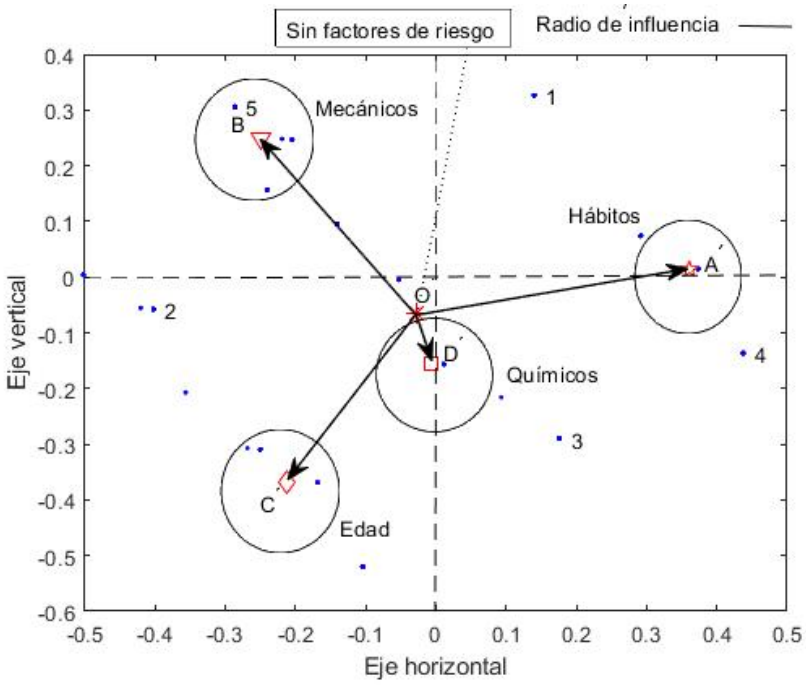


Figura 37. Zonas de influencia de los factores: Mecánicos, Químicos, Hábitos y Edad, para el grupo de Explotación de Minas.

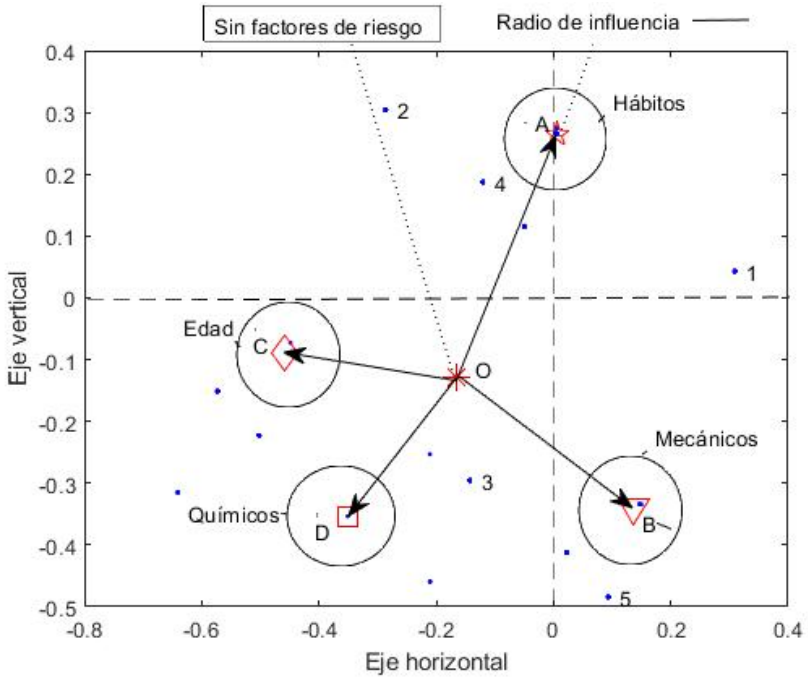
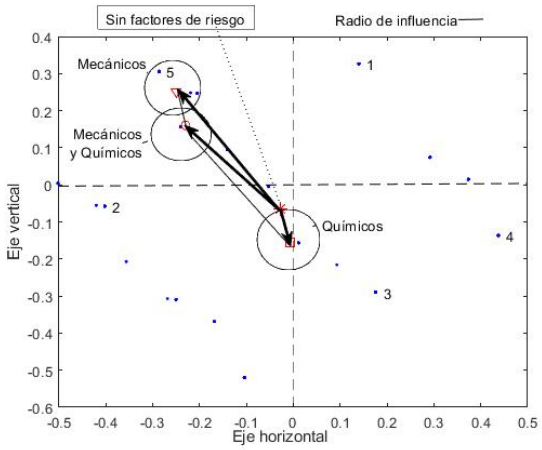


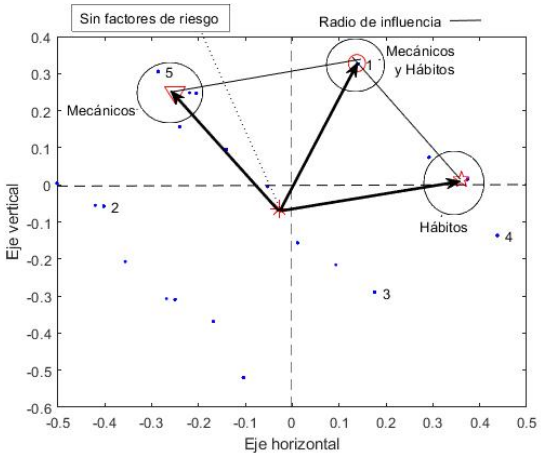
Figura 38. Zonas de influencia de los factores: Mecánicos, Químicos, Hábitos y Edad, para el grupo de Actividades Financieras.

Las Figuras 37 y 38, indican las zonas de influencia de los factores de riesgo para los dos grupos de análisis; sin embargo, se observa grandes áreas de los gráficos donde no se conoce la influencia de ningún factor. Para determinar las características de los individuos de estas áreas, sumamos los vectores que representan a los factores de riesgo, lo cual definirá las zonas de influencia de dos factores en conjunto. Lo indicado se observa en las Figuras 39 y 40.

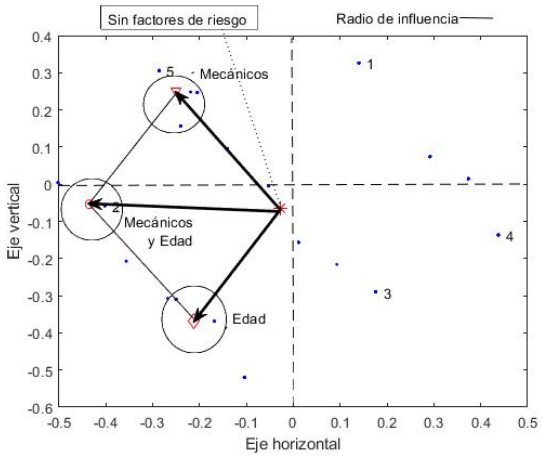
Las Figuras 39 a y 40 a, indican la suma de los vectores correspondientes a los factores de riesgo Mecánicos y Químicos; las Figuras 39 b y 40 b, muestran la suma vectorial de los factores de riesgo Mecánicos y de Hábitos; en las Figuras 39 c y 40 c, se tiene la suma de los vectores que representan los factores de riesgo Mecánicos y Edad; las Figuras 39 d y 40 d, presentan la suma vectorial de los factores de riesgo de Hábitos y Químicos; mientras que las Figuras 39 e y 40 e, indican la suma de los vectores que representan los factores de riesgo Edad y Químicos. Finalmente en las Figuras 39 f y 40 f, se observa la suma vectorial de los factores de riesgo Edad y Hábitos.



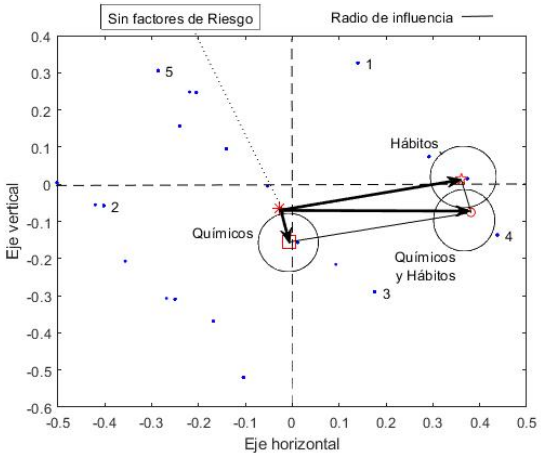
a) Mecánicos y Químicos



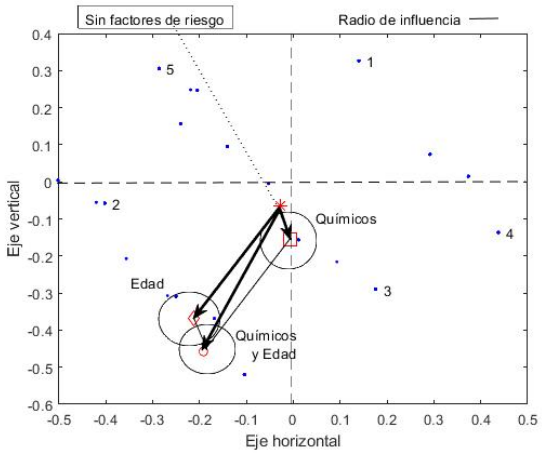
b) Mecánicos y Hábitos



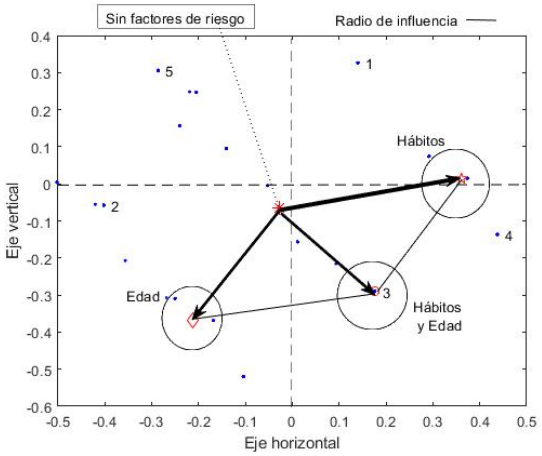
c) Mecánicos y Edad



d) Hábitos y Químicos

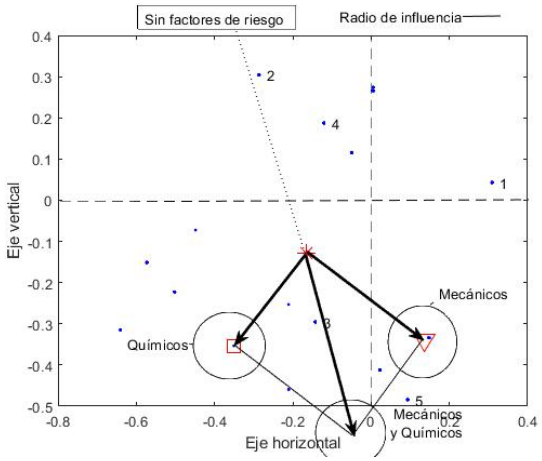


e) Edad y Químicos

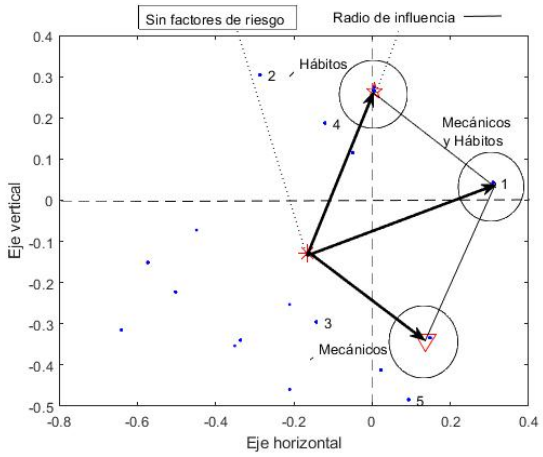


f) Edad y Hábitos

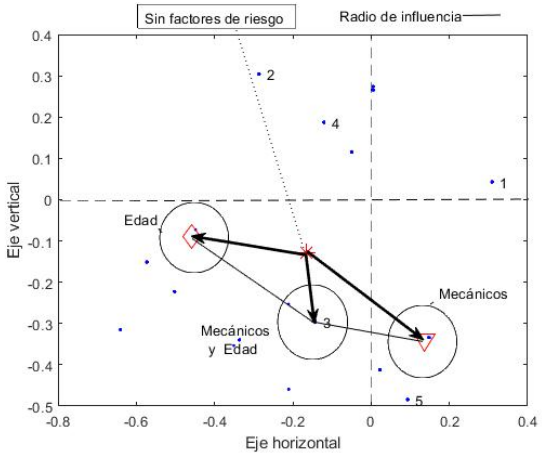
Figura 39. Suma de los factores de riesgo para el grupo de Explotación de Minas.



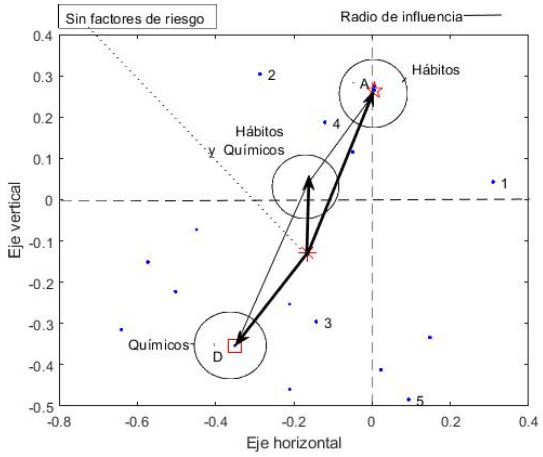
a) Mecánicos y Químicos



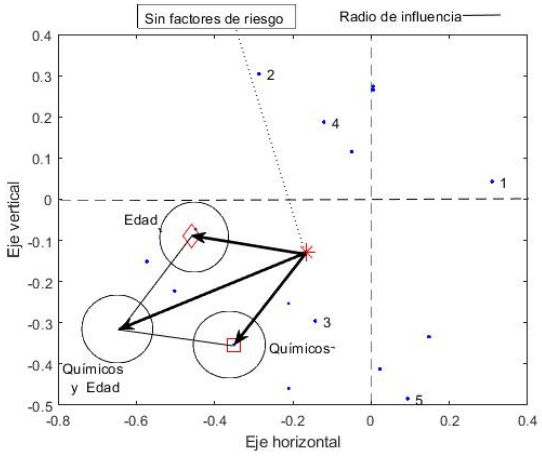
b) Mecánicos y Hábitos



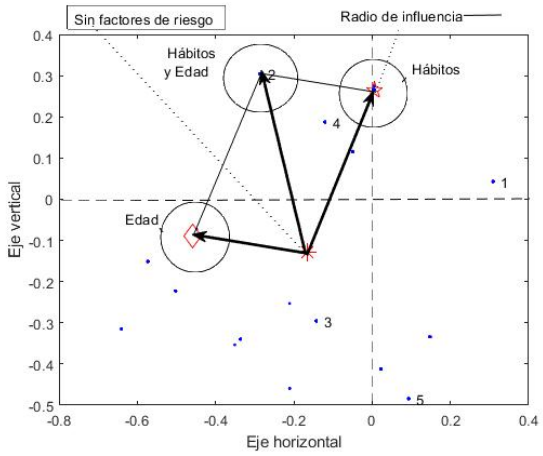
c) Mecánicos y Edad



d) Hábitos y Químicos



e) Edad y Químicos



f) Edad y Hábitos

Figura 40. Suma de los factores de riesgo para el grupo de Actividades Financieras.

Al sobreponer los gráficos de las Figuras 37 y 39, y de las Figuras 38 y 40, se obtienen las Figuras 41 y 42, respectivamente las mismas contienen los gráficos de similaridades con las zonas de influencia de uno o dos factores de riesgo para los grupos de Explotación de Minas y de Actividades Financieras.

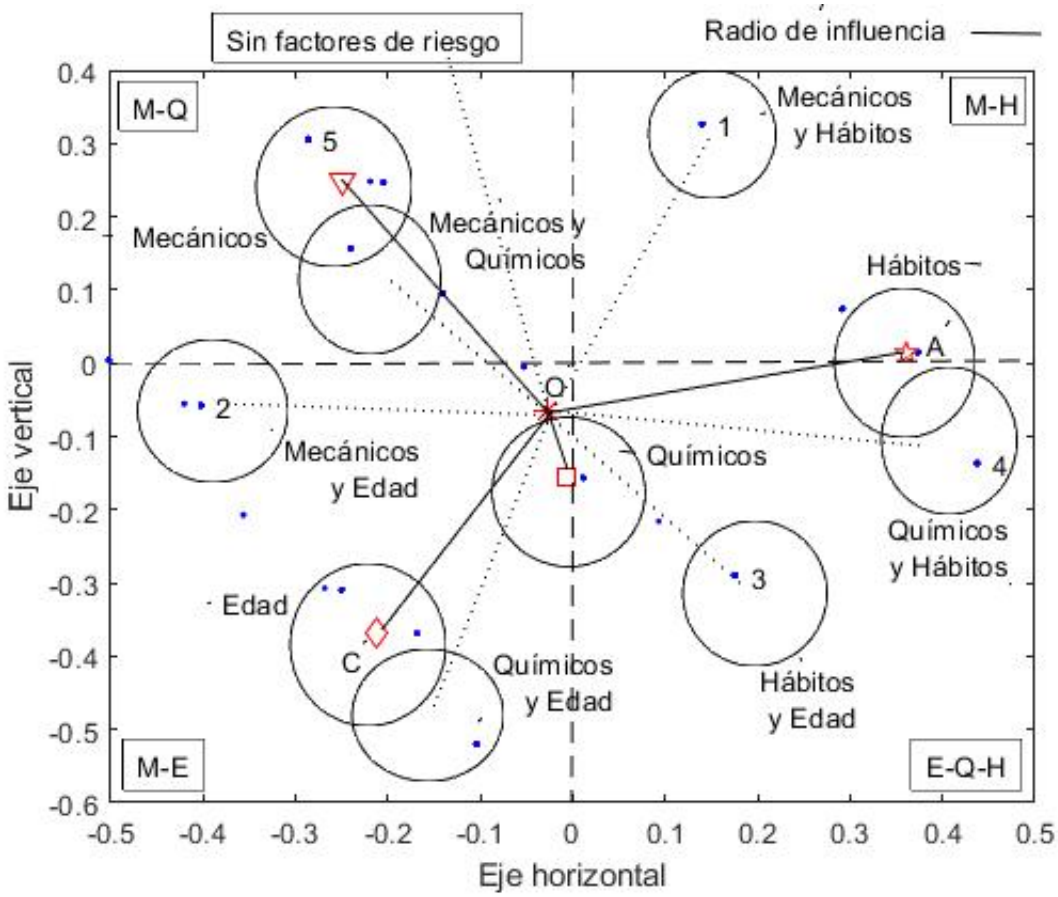


Figura 41. Zonas de influencia de uno o dos factores de riesgo para el grupo de Explotación de Minas.

La Figura 41 permite observar que el primer cuadrante del gráfico de similaridades esta bajo la influencia de los factores Mecánicos y Hábitos; en el segundo cuadrante la influencia es de de los factores Mecánicos y Químicos; a los factores Mecánicos y Edad les corresponde el cuadrante tres; y, en el cuarto cuadrante la influencia es de los factores Edad, Químicos y Hábitos. Lo indicado se puede verificar con el paciente número uno; pues está ubicado en el primer cuadrante bajo la influencia de factores Mecánicos y de Hábitos que coinciden con su situación, ya que sus factores de riesgo son IMC (Mecánico), Sedentarismo y Tabaquismo (Hábitos), según lo indica la Tabla 27.

La Figura 41 además puede indicar que para el grupo de Explotación de Minas, el eje horizontal representa al **factor Hábitos** y el eje vertical represente los **factores Mecánicos, Químicos y Edad**.

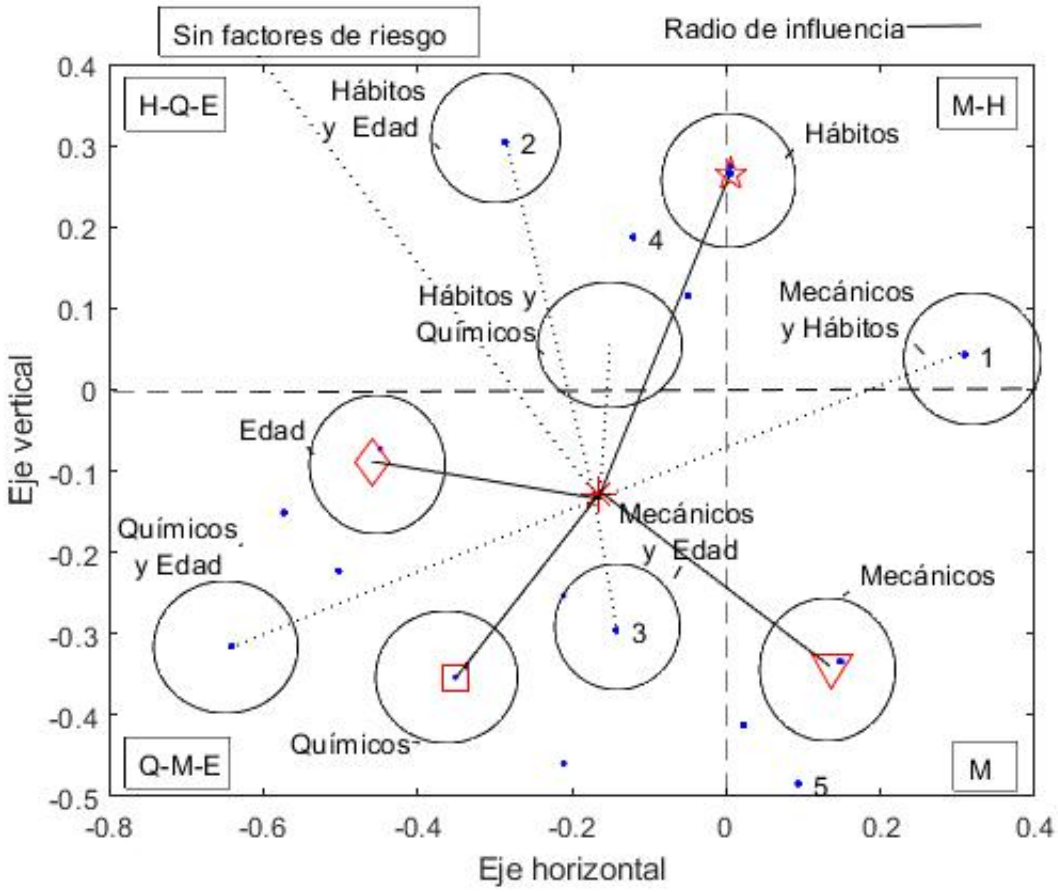


Figura 42. Zonas de influencia de uno o dos factores de riesgo para el grupo de Actividades Financieras.

En la Figura 42 se observa que el primer cuadrante del gráfico de similaridades esta bajo la influencia de los factores Mecánicos y Hábitos; en el segundo cuadrante la influencia es de de los factores Químicos, Hábitos y Edad; a los factores Mecánicos, Químicos y Edad les corresponde el cuadrante tres; y, en el cuarto cuadrante la influencia es de los factores Mecánicos. Lo indicado se puede verificar con el paciente número uno; pues está ubicado en el primer cuadrante bajo la influencia de factores Mecánicos y de Hábitos que coinciden con su situación, ya que sus factores de riesgo son IMC (Mecánico), Sedentarismo y Tabaquismo (Hábitos), según lo indica la Tabla 30.

La Figura 42 además puede indicar que para el grupo de Actividades Financieras, el eje horizontal representa a los **factores Mecánicos, Químicos y Edad** y el eje vertical represente al **factor Hábitos** .

Capítulo 4

4. DISCUSIÓN DE RESULTADOS

4.1. Análisis de la data

El presente trabajo se sustenta en un grupo de datos de exámenes clínicos realizados a pacientes por un laboratorio clínico de la ciudad de Quito. Para su análisis, estos datos se transformaron en una matriz donde las filas representan los pacientes y las columnas presentan información correspondiente a signos vitales, hábitos y resultados de exámenes clínicos.

De la base de datos, se seleccionó un grupos de variables que conforman el Síndrome Metabólico Plus SM+, es decir, las variables que definen el Sindrome Metabólico: IMC, Colesterol, Triglicéridos, Glucosa, Presión Sistólica y Presión Diastólica; más las variables Edad, Sedentarismo y Tabaquismo. Complementan la data de estudio, variables cualitativas como Educación, Sexo y Actividad Económica, las cuales permiten caracterizar la población y formar grupos de interés; adicionalmente, se crea la variable ID, para la identificación anónima de los pacientes. La data resultante con la información antes indicada, forma una matriz de 12 363 observaciones y 13 variables.

La información procesada permitió obtener modelos matemáticos descriptivos, predictivos y de toma de decisiones, con el objeto de describir y predecir el estado de salud de un paciente.

El análisis de la data tuvo las siguientes fases:

- i) Transformación de la matriz de datos proporcionada al lenguaje Matlab,
- ii) eliminación de datos erróneos, inconsistentes o redundantes,
- iii) creacón de nuevas variables,
- iv) categorización de variables numéricas,
- v) selección de variables para análisis específicos, con el objeto de identificar observaciones (pacientes) en situaciones críticas.

4.2. Análisis descriptivo

El estudio estadístico descriptivo de las variables relevantes permite caracterizar a la población de estudio, los principales resultados son:

- Para la variable Edad, la media poblacional es 35.49 años. Las categorías relevantes en esta variable son: Menores de 30 años con 31,10 %, de 30 a 40 años con 39.10 % y de 40 a 50 años con el 19.90 %; en conjunto, estos grupos representan el 90.10 % de la población.

- La media de la variable IMC, de la población total es de 26.48 kg/m^2, valor que corresponde al grupo de sobrepeso. En esta variable los grupos predominantes son: normal con 37.20 %, sobrepeso con el 45.40 % y obeso 1 con el 13.70 %. En conjunto representan el 96.30 % de la población. El 62.10 % de la población total sobrepasan los límites de normalidad de la variable IMC (aproximadamente 2 de cada 3 personas).

- La media de la variable Colesterol es de 191.78 mg/dl, la misma que está bajo el valor del límite de normalidad, pero muy cercano a este (el límite de normalidad es 200 mg/dl). El 39.10 % de la población posee Colesterol Alto (aproximadamente 2 de cada 5 personas).

- Para la variable Triglicéridos, la media es 149.21 mg/dl, valor muy cercano al límite de normalidad de esta variable, que es 150 mg/dl. El 36.46 % de la población tiene triglicéridos altos (aproximadamente 1 de cada 3 personas).

- La media de la variable Glucosa es de 91.03 mg/dl, valor que corresponde a glucosa normal; sin embargo, el 14.10 % de la misma población, tiene problemas de glucosa alta (aproximadamente 1 de cada 7 personas).

- La media de la variable Presión Sistólica es 112.29 mmHg, valor por debajo del límite de normalidad de 140 mmHg. La mayoría de la población, el 90.40 % tiene presión sistólica normal, el 2.90 % tiene presión sistólica alta y 6.70 % tiene presión sistólica baja.

- La media de la variable Presión Diastólica es 72.40 mmHg, valor por debajo del límite de normalidad de 90 mmHg. El 93.00 % de la población esta dentro del rango normal, 4.1 % tiene la presión diastólica alta y 2.90 % tiene presión diastólica baja.

4.3. Análisis del Síndrome Metabólico

El Síndrome Metabólico es una serie de desórdenes o anormalidades de cualquiera de estas variables: IMC, Colesterol, Triglicéridos, Glucosa, Presión Sistólica y Presión Diastólica; considerando que la edad y los hábitos del paciente inciden en su estado de salud, se añaden las variables Edad, Sedentarismo y Tabaquismo. A este conjunto de variables se denomina Sindrome Metabólico Plus SM+, sobre el cual se realiza un análisis inferencial con tres enfoques: la creación de un Índice de Salud, la determinación de los factores de riesgo y el análisis del escalado multidimensional sobre dos grupos específicos de la población.

Análisis respecto a Índices de Salud

Una forma para conocer el estado de salud de un paciente es la determinación de su Índice de Salud, valor que lo categoriza dentro de un grupo de salud específico, definidos en la Tabla 20. La distribución de la población de acuerdo a este índice es para los grupos: Bajo 695, Normal 4 536, Sobre 6 244 y Crítico 888 pacientes.

Las actividades económicas con mayor población son Explotación de Minas con 3 329 y Actividades Financieras con 2 199 pacientes; la distribución de la población de acuerdo al Estado de Salud, considerando estos grupos es:

- i) Para Explotación de Minas:
  - Bajo 142 pacientes (4.27 %),
  - Normal 996 pacientes (29.92 %),
  - Sobre 1 869 pacientes (56.14 %) y
  - Crítico 323 pacientes (9.67 %).
- ii) Para Actividades Finacieras:
  - Bajo 112 pacientes (5.09 %),
  - Normal 968 pacientes (44.03 %),
  - Sobre 1 021 pacientes (46.43) % y
  - Crítico 98 pacientes (4.45 %).

Comparando la distribución de las dos poblaciones se observa que la actividad económica Explotación de Minas posee mayor porcentaje de pacientes en los Estado de Salud Sobre y Crítico, en comparación con Actividades Financieras.

La distribución de la población según las variables categóricas definidas en las Tablas 4-10, permiten determinar grupos poblacionales con características específicas. Así, por ejemplo se tiene; en la Figura 10, que la actividad económica Explotación de Minas tiene 338 pacientes (2.73 % de la población total) y en Actividades Financieras existen 131 pacientes (1.06 % de la población total), en el grupo definido por las categorias Sobrepeso, Colesterol Alto, Triglicéridos Alto y Sedentarios.

En forma similar se puede hacer comparaciones entre diversos grupos poblacionales de acuerdo a las categorías de las variables IMC, Sexo, Educación, Edad; para ello se puede hacer uso de las figuras del Anexo 2, para la población total; del Anexo 3, para la población de Explotación de Minas y del Anexo 4, para la población de Actividades Financieras

Análisis a partir de los factores de riesgo

Se considera como factores de riesgo para la salud a las variables del SM+ cuando sus valores sobrepasan el límite de normalidad. Categorizando la población por actividad económica y factores de riesgo; en primer lugar, se analiza en forma general a toda la población y luego se particulariza a los grupos Explotación de Minas y Actividades Financieras, por tener las mayores poblaciones, 3 329 y 2 199 pacientes, respectivamente.

La distribución de la población de acuerdo a los factores de riesgo y actividades económicas, permiten determinar grupos con características específicas, para implementar acciones encaminadas a mejorar el estado de salud del paciente. Así por ejemplo se tiene:

- la Figura 67, indica que el número de personas que tiene los factores de riesgo Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Sobrepeso; corresponden al grupo de Explotación de Minas 487 pacientes (3.94 % del total de la población) y al grupo Actividades Financieras 196 (1.59 % del total de la población).

- En la Figura 69, el número de pacientes de sexo masculino y femenino, con cuatro factores de riesgo Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y que pertenecen al grupo Explotación de Minas son 518 y 52 respectivamente; por otro lado, la población de las Acti-

vidades Financieras son 140 pacientes del sexo masculino y 122 del femenino .

- Respecto a la edad, la Figura 71 indica 403 pacientes del grupo Explotación de Minas y 137 del grupo Actividades Financieras, estos poseen los cuatro factores de riesgo: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y sobrepasan los 35 años.

- En cuanto al nivel de instrucción, la Figura 73 muestra 351 pacientes del grupos Explotación de Minas y 183 del grupo Financiero que poseen los cuatro factores de riesgo Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y pertenecen al tercer nivel.

En forma similar se puede encontrar el número de pacientes bajo la influencia de determinados factores de riesgo o las variables categóricas: Sexo, Edad o nivel de Educación. (Ver Anexos del 2 al 7).

La categorización de la población de acuerdo al número de factores de riesgo que lo afectan, permite identificar grupos en diversas situaciones de salud.

Análisis del Escalado Multidimensional

El Escalado Multidimensional se aplica sobre los grupos poblacionales de las actividades económicas Explotación de Minas y Actividades Financieras con tres factores de riesgo, el estudio de las similitudes entre los elementos de estos grupos, permite observar el comportamiento de los factores de riesgo y su influencia en el estado de salud del paciente, los principales resultados obtenidos son:

- El análisis es robusto respecto al uso de diversos coeficientes de similaridad que consideren las ausencias relativas (la frecuencia d), ya que al aplicar al grupo Explotación de Minas con tres factores críticos, los coeficientes de similaridad de Sokal-Michener, Rogers-Tanimoto y Russel-Rao se obtienen resultados similares.

- En lo referente a los factores Mecánicos, que incluyen las variables IMC y Presión Arterial, resulta más eficiente realizar un control de la variable IMC para alcanzar un estado óptimo de salud.

- Los factores Químicos son: Glucosa, Colesterol y Triglicéridos, para un paciente que tenga estos factores de riesgo, la simulación indica que para lograr un estado de salud óptimo resulta eficiente controlar las variables Triglicéridos y Colesterol.

- Si el paciente tiene factores de riesgo Hábitos, Sedentarismo y Tabaquismo; lo conveniente para lograr un estado de salud óptimo, es controlar la variable Sedentarismo.

- La ubicación de los factores Mecánicos, Químicos, Hábitos y Edad; considerados en los gráficos de similaridades, de los grupos Explotación de Minas y Actividades Financieras, (Figuras 37 y 38), se diferencian debido a que los factores de riesgo presentan diferentes porcentajes en cada una de estas actividades.

- El paciente sin factores de riesgo, en el grupo Actividades Financieras se ubica en el mismo cuadrante del grupo Explotación de Minas, pero más alejado del centro de coordenadas.

- **El vector que representa los factores Mecánicos en el grupo Explotación de Minas es de mayor longitud a su similar del grupo Actividades Financieras, lo que se puede interpretar que los factores Mecánicos son más importantes en el grupo Explotación de Minas.**

- **El vector que representa los factores Químicos en el grupo de Actividades Financieras es de mayor longitud a su similar del grupo Explotación de Minas, lo que indica que los factores Químicos son más relevantes en Actividades Financieras.**

- **Los vectores que representan a los factores Edad y Hábitos son de similar longitud, lo que indica que estos factores tienen igual influencia en los dos grupos.**

Capítulo 5

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

- Las características de la población objeto del presente estudio son:
  - La población tiende al sobrepeso.
  - El 39.10 % de la población tiene colesterol alto.
  - El 36.46 % de la población tiene triglicéridos alto.
  - El 14.10 % de la población tiene glucosa alta.
  - El 3.00 % de la población tiene presión arterial alta.

  Todo esto, indica que será recomendable un cambio de hábitos y una mejor alimentación en la población.

- Respecto a la Actividad Económica de la población, la mayoría se concentra en dos: Explotación de Minas y, en Actividades Financieras.

- El índice de masa corporal es un factor determinante en la salud de una persona, ya que si éste se mantiene en valores normales las otras variables también se mantienen bajo límites de normalidad y si este índice aparece indistintamente las otras variables también aparecen como el Colesterol, Triglicéridos, Glucosa, etc. De igual manera una persona que adolece de una de las enfermedades del perfil metabólico, debería dar prioridad en bajar los valores de IMC, dando como consecuencia la baja de valores en las otras variables.

- Respecto a las variables Tabaquismo y Sedentarismo, estas variables indudablemente influyen, especialmente la segunda. Toda persona que no es sedentaria tiene mejores posibilidades de desarrollar una vida sana; sin embargo, estas dos variables son fruto de encuestas y no son obtenidas bajo ningún proceso de laboratorio, por lo que su análisis hay que tomarlo con discreción

- Esta investigación es producto de las competencias adquiridas en la MEMAT y se alinea con los objetivos planteados, tales como: elevar el nivel académico y científico mediante la aplicación de nuevas metodologías y herramientas tecnológicas; la aplicación de la matemática en problemas de la vida real y la solución de los mismos; utilizar programas computacionales acordes a los requerimientos y a las necesidades de la enseñanza actual de las matemáticas con el fin de potenciar la actividad del docente investigador.

5.2. Recomendaciones

En países desarrollados, en el área de la salud, los laboratorios clínicos, en los últimos años han implementado tecnologías que permiten almacenar y recopilar datos por ellos generados, obteniendo múltiples ventajas. En Ecuador no existe información sobre la implementación de algún método técnico que permita aprovechar la información generada por los sistemas de salud, es más no existe siquiera alguna política de estado que de las directrices o normas para trabajar en esta área; por tal motivo, se recomienda normalizar las diferentes mediciones de exámenes clínicos.

Las tecnologías basadas en Big Data están revolucionando todos los campos, uno de ellos es el de la salud, el aporte de este trabajo esta encaminado a construir una metodología de análisis de un perfil médico, y se aplica al SM+, pudiendo aplicarse a otros perfiles en trabajos futuros.

ANEXOS

Anexo 1

Lista total de variables proporcionadas

Tabla 31

Variables de la data, con porcentaje de pacientes, información y tipo.

Número	Variable	Porcentaje	Tipo
1	Identificación	100.00	N
2	Fecha admisión	100.00	N
3	n	100.00	N
4	Hábito A	100.00	L
5	Hábito D	100.00	L
6	Hábito S	100.00	L
7	Hábito T	100.00	L
8	Fecha nacimiento	99.99	L
9	Género	100.00 7	L
10	Tipo sangre	95.76	L
11	Educación	100.00	L
12	Empresa	100.00	L
13	Peso	94.50	N
14	Estatura	94.48	N
15	Frecuencia cardiaca	89.96	N
16	imc	91.43	N
17	Presión arterial	94.52	N
18	Presión arterial 2	94.48	N
19	Saturación oxigeno	83.18	N
20	Respiración	79.42	N
21	Temperatura	77.78	N
22	Ácido úrico	64.15	N
23	Neutrofilos	98.92	N
24	Basofilos	97.72	N
25	Eosinofilos	97.72	N
26	Concentración corpuscular media hemoglobina	98.92	N
27	Hemoglobina	98.92	N
28	Volumen plaquetario medio	98.91	N

Sigue en la siguiente página

Número	Variable	Porcentaje	Tipo
29	Eritrocitos	98.90	N
30	Linfocitos	98.92	N
31	Plaquetas	98.92	N
32	Volumen corpuscular medio	98.92	N
33	Monocitos	97.74	N
34	Leucocitos	98.92	N
35	Hemoglobina corpuscular media	96.93	N
36	Hematocrito	98.92	N
37	Ancho de distribución G.R	98.92	N
38	Creatinina	88.82	N
39	Glucosa basal	90.97	N
40	Triglicéridos	79.71	N
41	Colesterol total	79.94	N
42	Colesterol HDL	72.61	N
43	Colesterol LDL (calculado)	70.28	N
44	Densidad	10.06	N
45	Ph	10.06	N
46	Urea	43.20	N
47	Hepatitis B, anticuerpos anti-HBS	1.35	N
48	Bilirrubina directa	5.09	N
49	Bilirrubina indirecta	5.09	N
50	Bilirrubina total	1.29	N
51	TGO/AST	43.34	N
52	TGP/ALT	43.35	N
53	Gammaglutamil tranpeptidasa (GGT)	13.28	N
54	PSA total (antígeno prostático específico)	8.15	N
55	CEA(antígeno carcinoembrionario)	1.45	N
56	Fosfata alcalina	3.26	N
57	PSA libre	0.82	N
58	% PSA total/PSA libre	0.74	N

Sigue en la siguiente página

Número	Variable	Porcentaje	Tipo
59	Tiempo de protrombina(TP)	0.04	N
60	Tiempo de tromboplastina parcial (TTP)	5.23	N
61	TSH	5.72	N
62	Hemoglobina glicada (HBA1C)	0.32	N
63	Bun	0.96	N
64	Velocidad de cedimentación 1 hora	0.16	N
65	Colesterol LDL (cuantificado)	1.98	N
66	Hepatitis B, anticuerpos anti-HBS AG	0.18	N
67	FT4 libre	0.06	N
68	Plomo en sangre	0.01	N
69	FT3 libre	0.05	N
70	H1V1Lanticuerpos+antígeno P24	0.02	N
71	Albúmina	0.17	N
72	Calcio sérico total	0.00	N
73	CPK	0.06	N
74	Globulina	0.14	N
75	Hierro sérico	0.00	N
76	Potasio(K)	0.00	N
77	Proteinas totales	0.21	N
78	Sodio (Na)	0.00	N
79	Transferrina	0.00	N
80	Lípidos totales	0.02	N
81	Alfafeto proteina (AFP)	0.05	N
82	Ca 125	0.05	N
83	Ca 72-4	0.05	N
84	LK-MB (masa)	0.06	N
85	Homocisteina	0.05	N
86	LDH	0.15	N
87	PCR cuantitativo	0.09	N
88	T3 total	0.05	N

Sigue en la siguiente página

Número	Variable	Porcentaje	Tipo
89	T4 total	0.07	N
90	Colinesterasa acetil eritrocitaria (sangre total)	0.03	N
91	Reticulositos	0.02	N
92	Apolipoproteina B	0.14	N
93	Glucosa postprandial	0.00	N
94	Insulina basal	0.00	N
95	Apolipoproteina A1	0.00	N
96	Asto cuantitativo	0.00	N
97	Amilasa	0.09	N
98	Cromo en sangre	0.00	N
99	Magnesio en sangre	0.00	N
100	Colesterol VLDL	0.00	N
101	Anticuerpos anti-chlamydia trachomatis IGG	0.00	N
102	Anticuerpos anti-chlamydia trachomatis IGM	0.00	N
103	Anticuerpos anti-herpes II IGG	0.00	N
104	Anticuerpos anti-herpes II IGM	0.00	N

N: Variable cuantitativa; L: Variable cualitativa.

Anexo 2

Histogramas: población total y variables con múltiples categorías

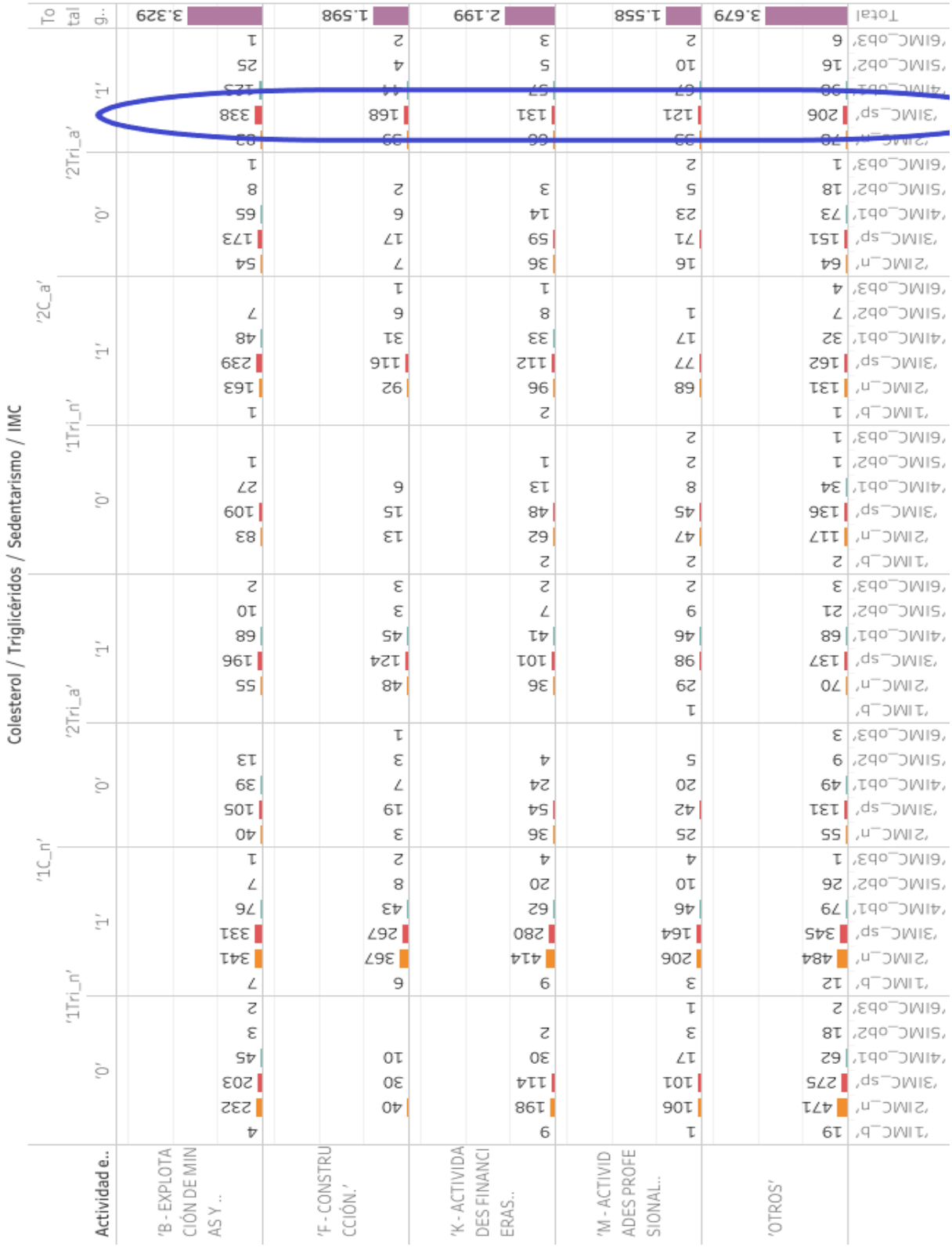


Figura 43. Distribución de la población total de acuerdo a diversas categorías: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo e IMC.

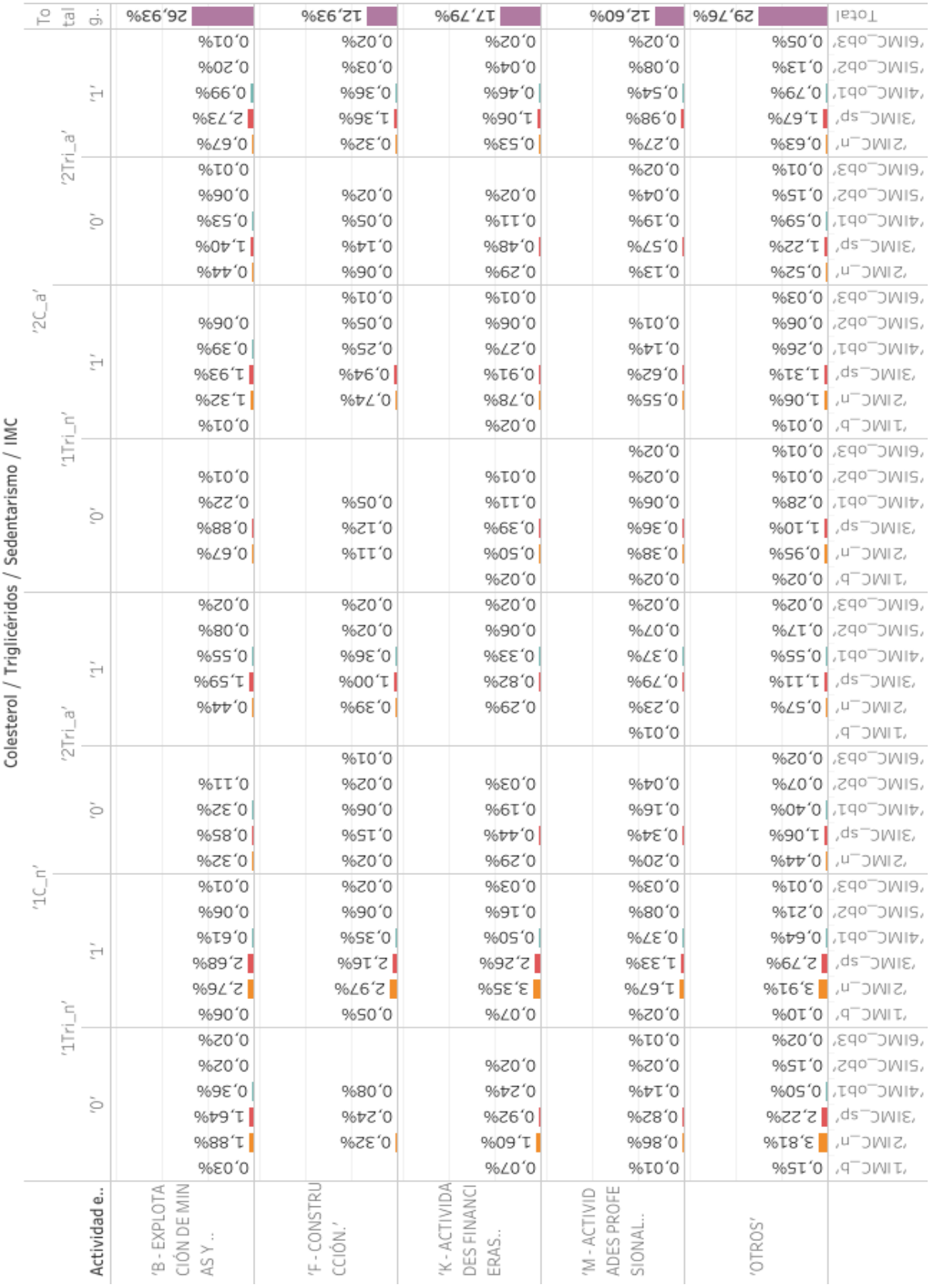


Figura 44. Distribución en porcentajes de la población total de acuerdo a diversas categorías: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo e IMC.

Figura 45. Distribución de la población total de acuerdo a diversas categorías: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo y Sexo.

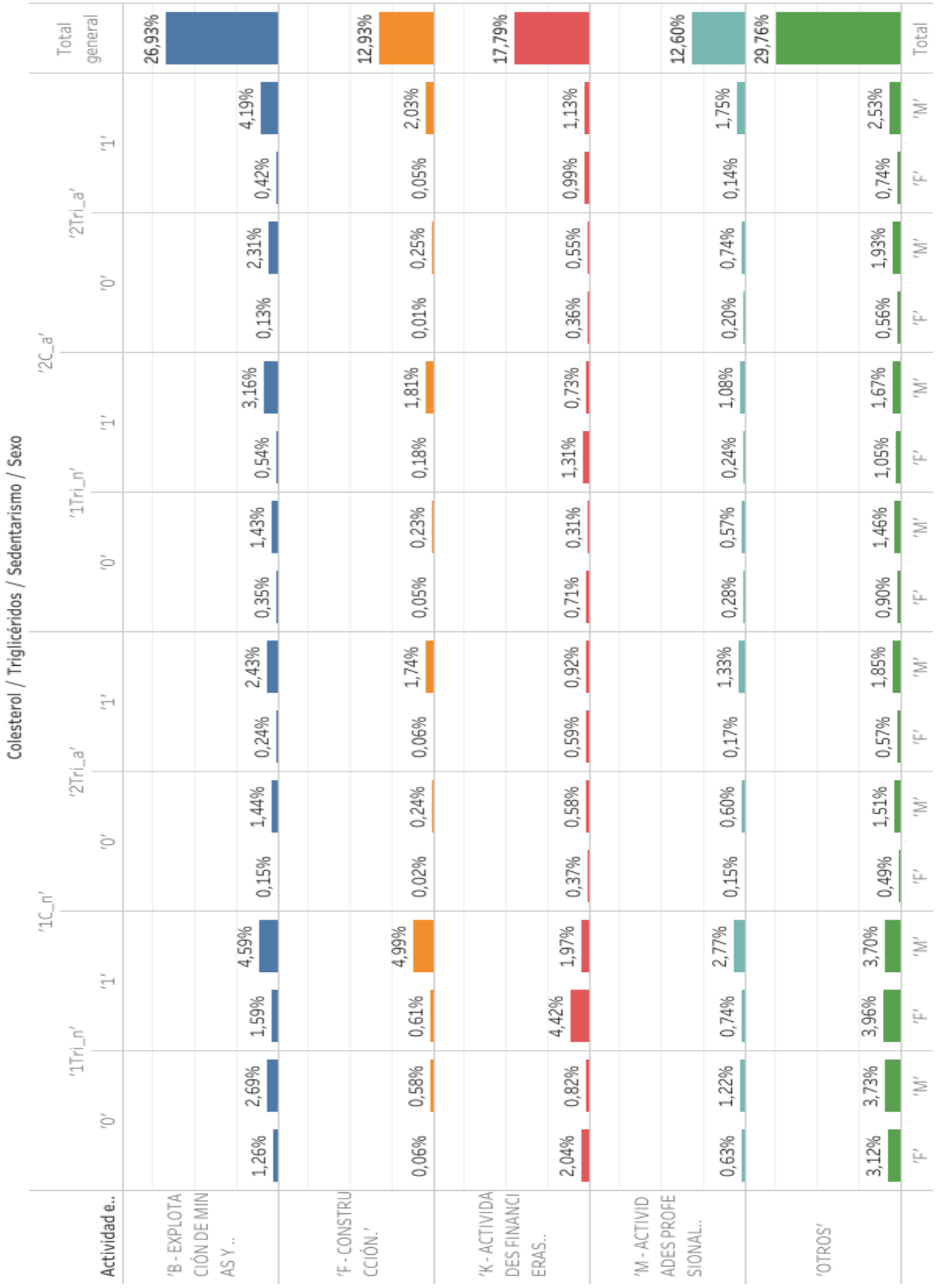


Figura 46. Distribución en porcentajes de la población total de acuerdo a diversas categorías: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo y Sexo.

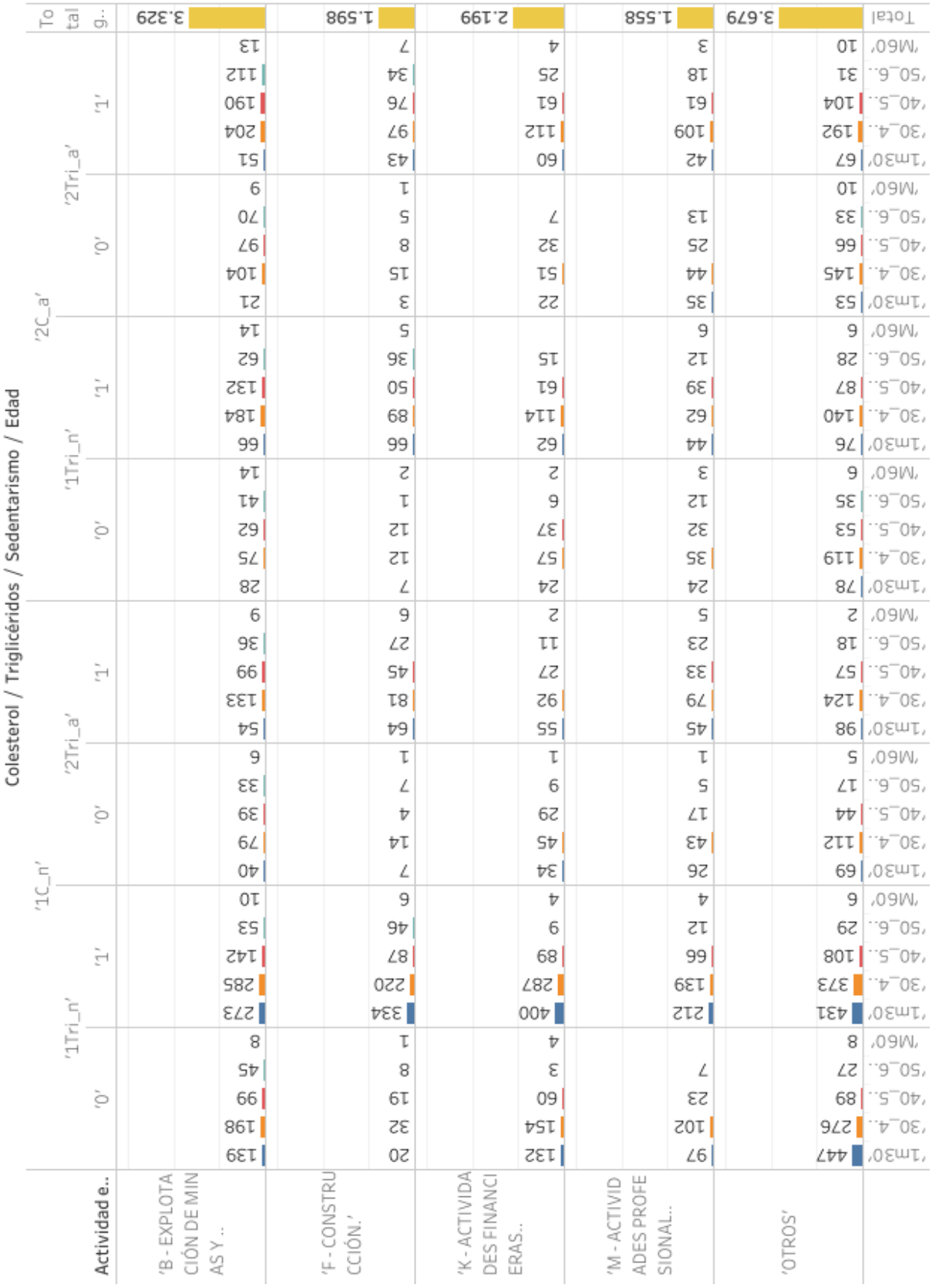


Figura 47. Distribución de la población total de acuerdo a diversas categorías: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo y Edad.

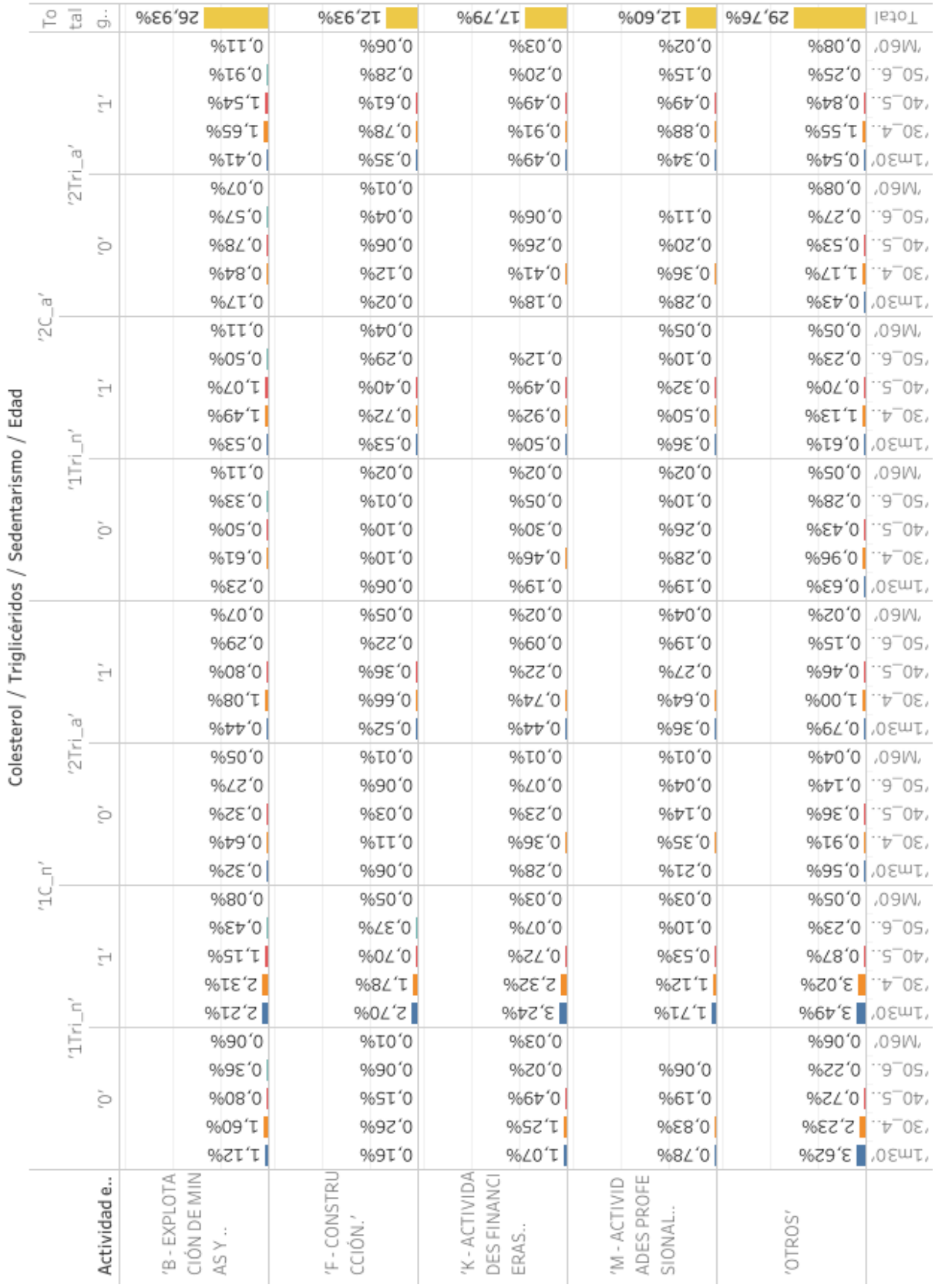


Figura 48. Distribución en porcentajes de la población total de acuerdo a diversas categorías: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo y Edad.

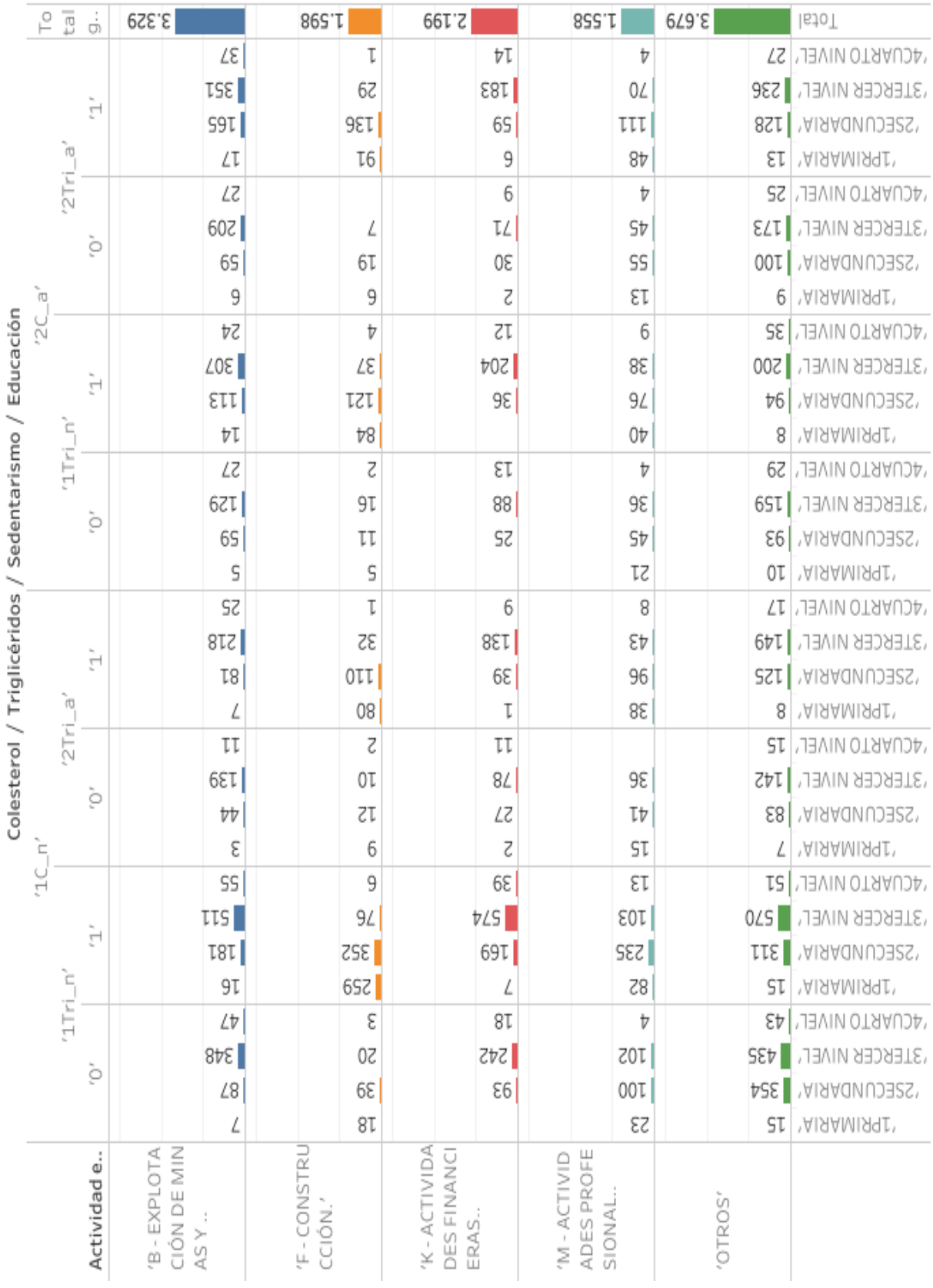


Figura 49. Distribución de la población total de acuerdo a diversas categorías: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo y Educación.

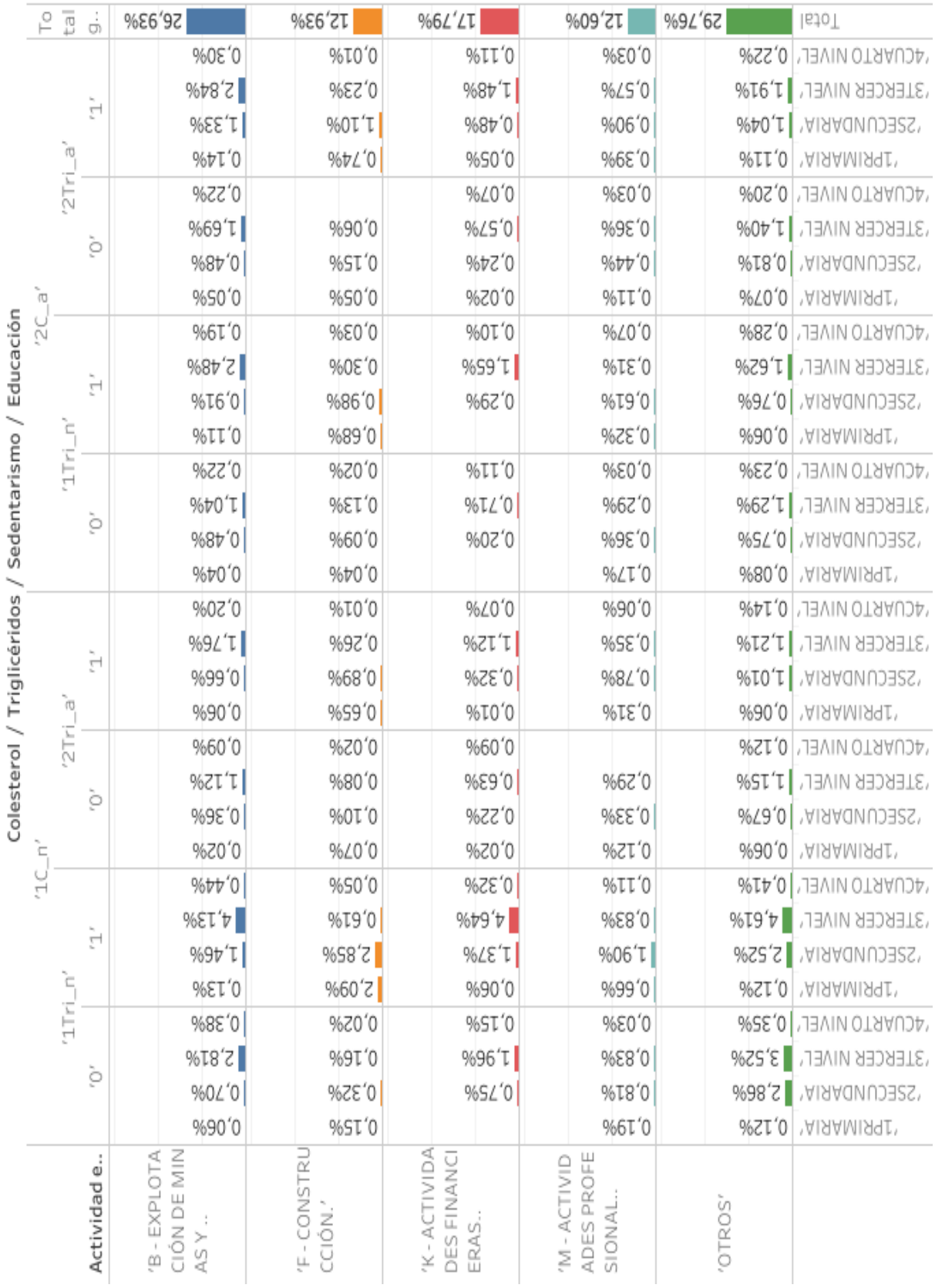


Figura 50. Distribución en porcentajes de la población total de acuerdo a diversas categorías: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo y Educación.

Anexo 3

Histogramas: Explotación de Minas y variables con múltiples categorías

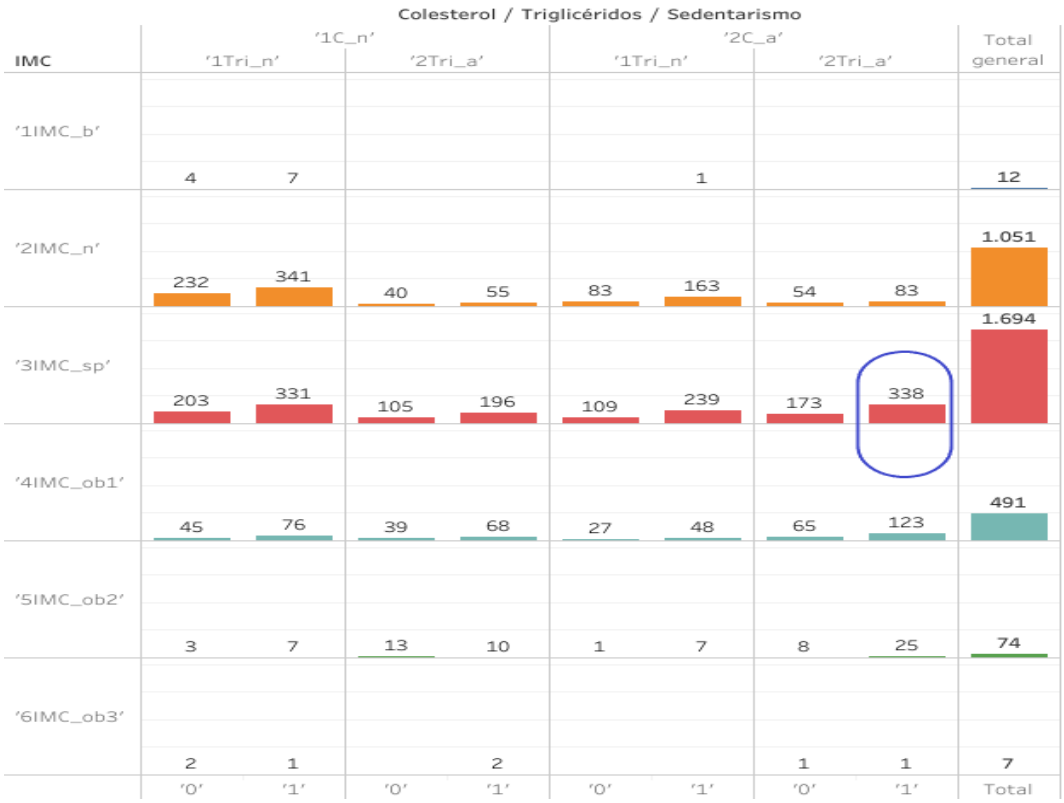


Figura 51. Distribución de la población del grupo Explotación de Minas de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo e IMC.

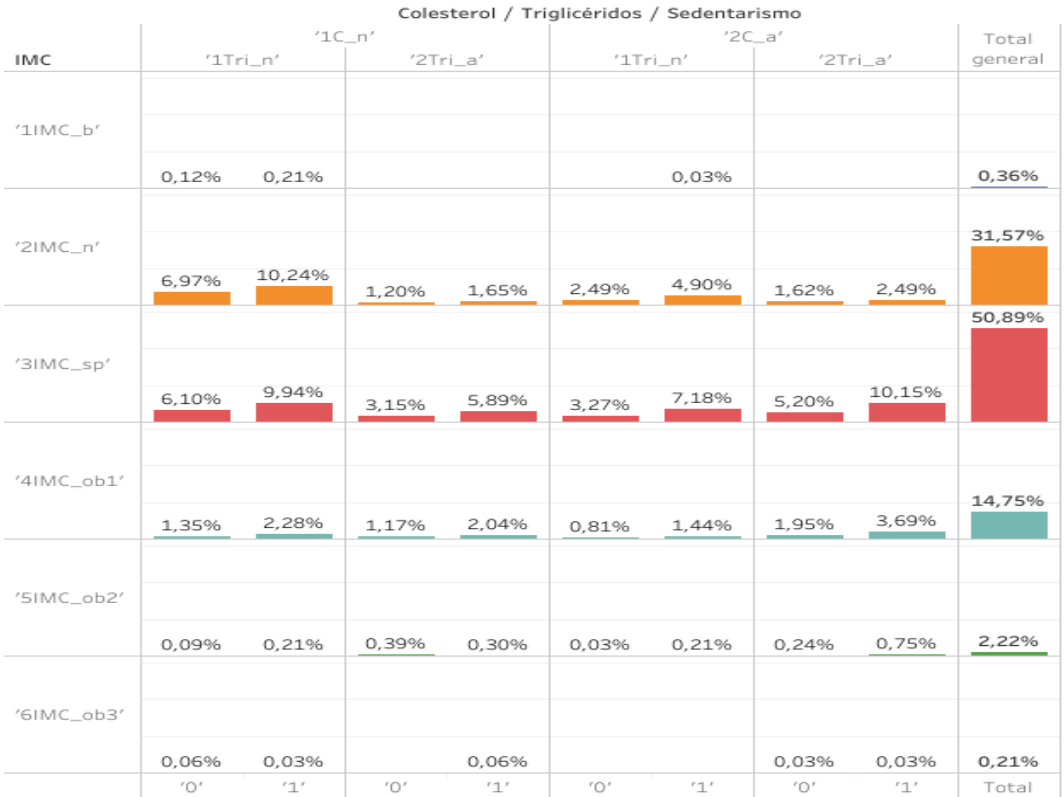


Figura 52. Distribución en porcentajes de la población del grupo Explotación de Minas de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo e IMC.

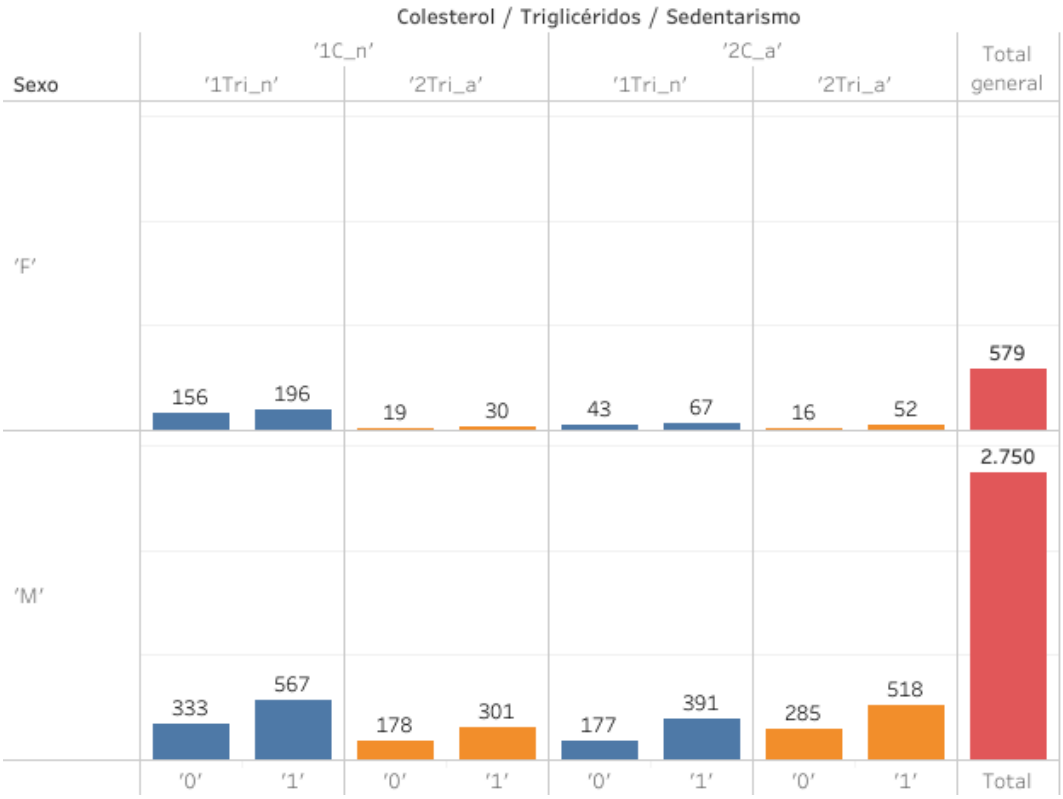


Figura 53. Distribución de la población del Explotación de Minas de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo y Sexo.

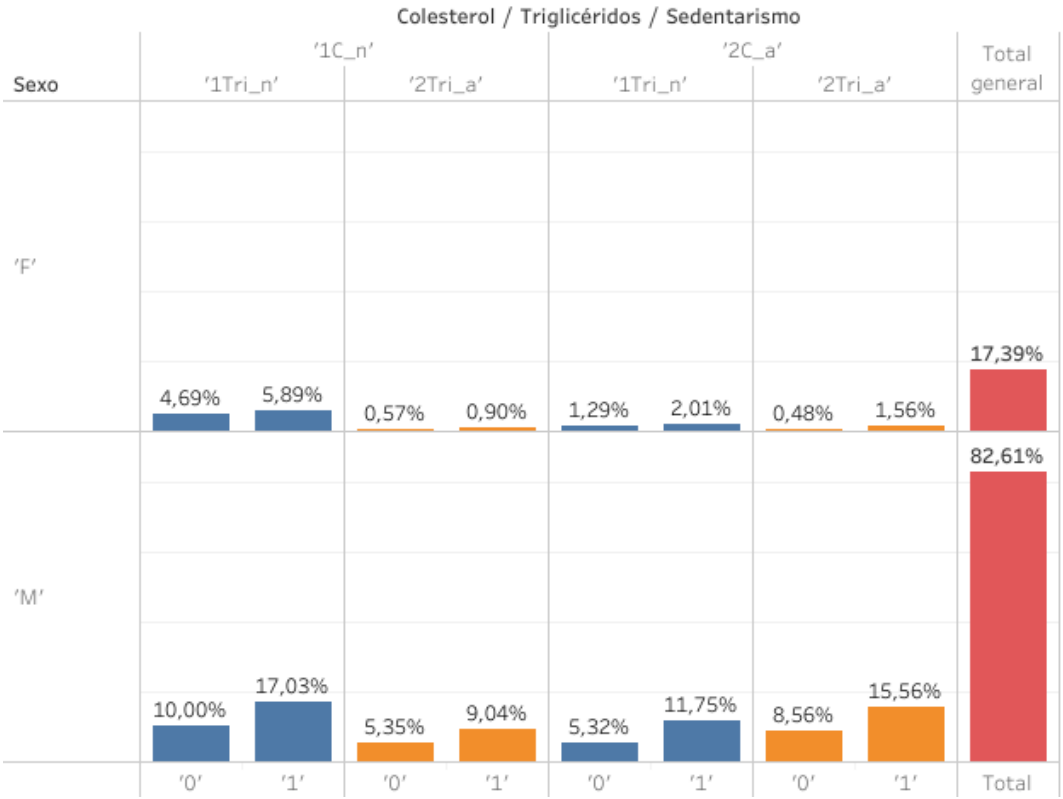


Figura 54. Distribución en porcentajes de la población del grupo Explotación de Minas de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo y Sexo.

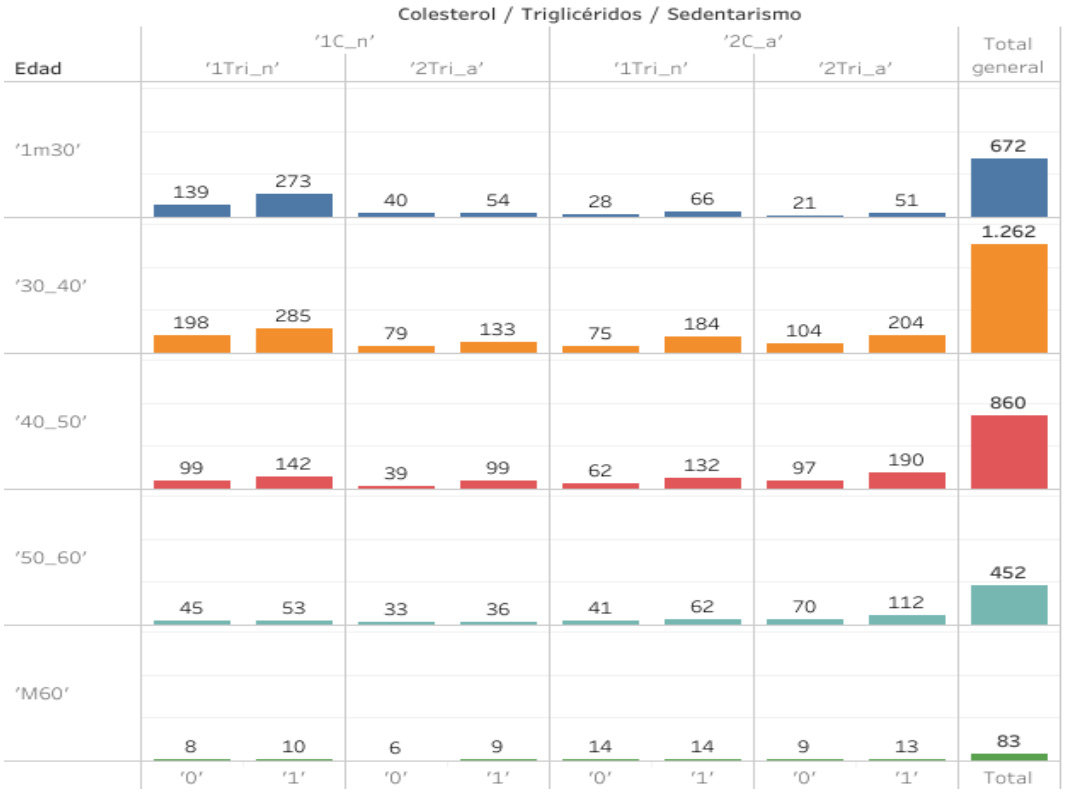


Figura 55. Distribución de la población de Explotación de Minas de acuerdo a diversas categorías: Coleterol, Triglicéridos, Sedentarismo y Edad.

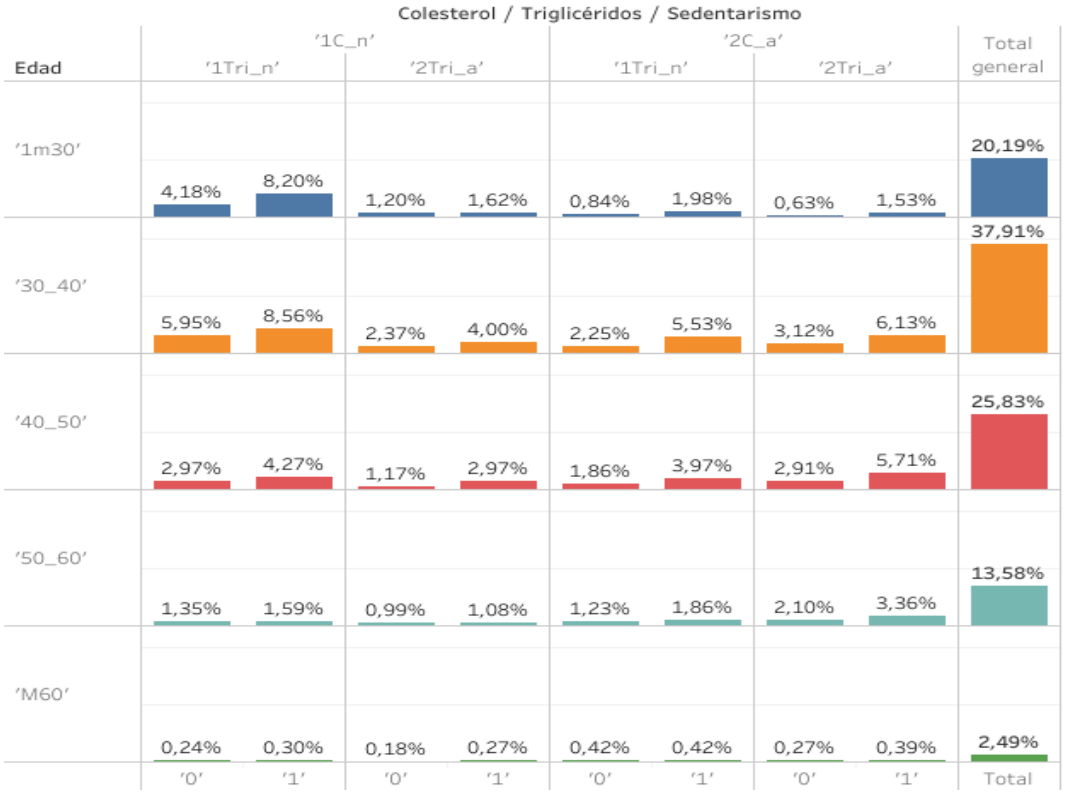


Figura 56. Distribución en porcentajes de la población del grupo Explotación de Minas de acuerdo a diversas categorías: Coleterol, Triglicéridos, Sedentarismo y Edad.

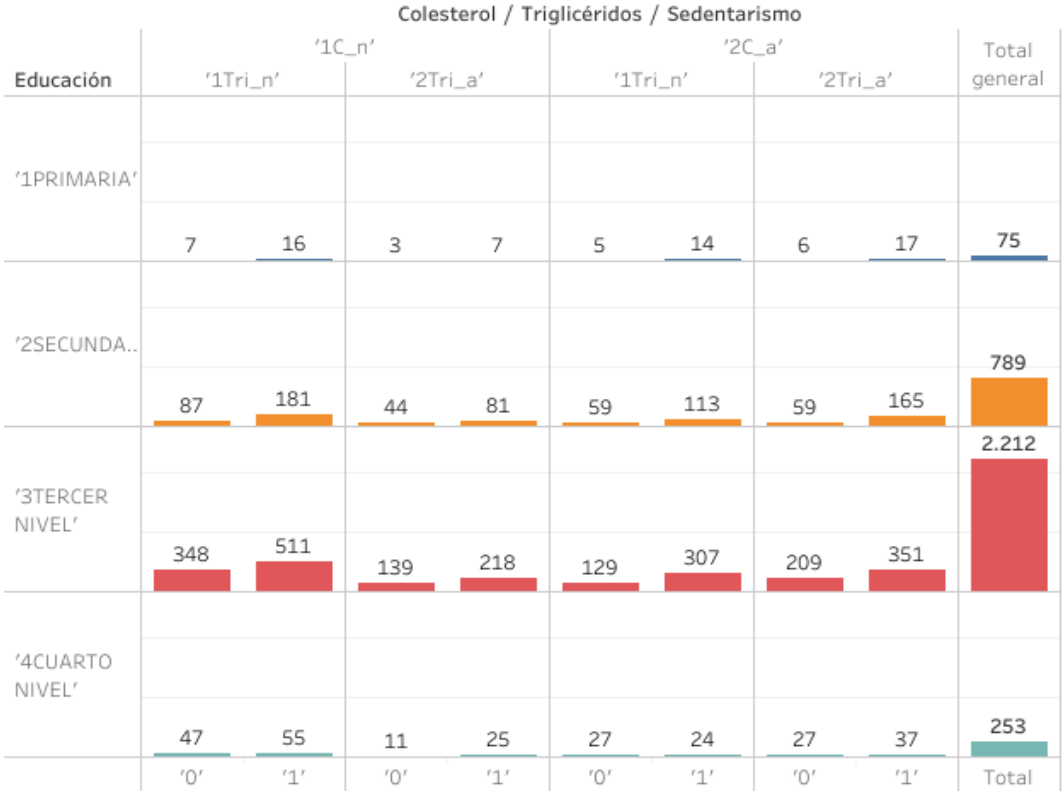


Figura 57. Distribución de la población del grupo Explotación de Minas de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo y Educación.

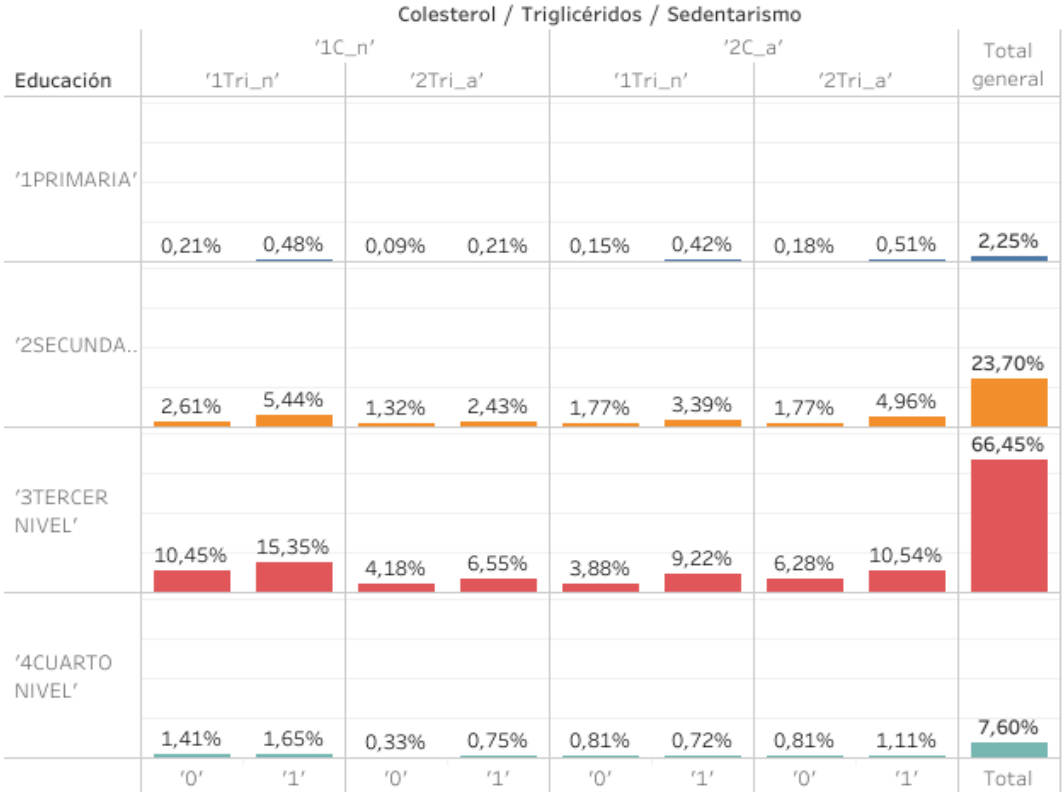


Figura 58. Distribución en porcentajes de la población del grupo Explotación de Minas de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo y Educación.

Anexo 4

Histogramas: Actividades Financieras y variables con múltiples categorías

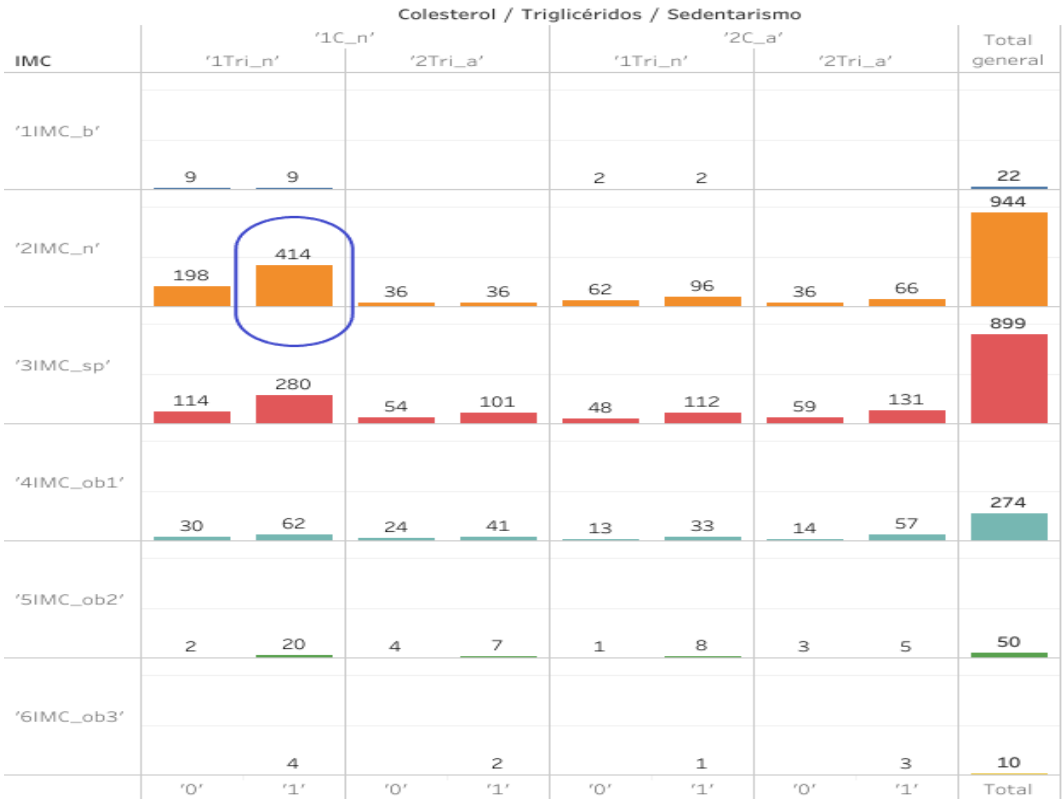


Figura 59. Distribución de la población del grupo Actividades Financieras de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo e IMC.

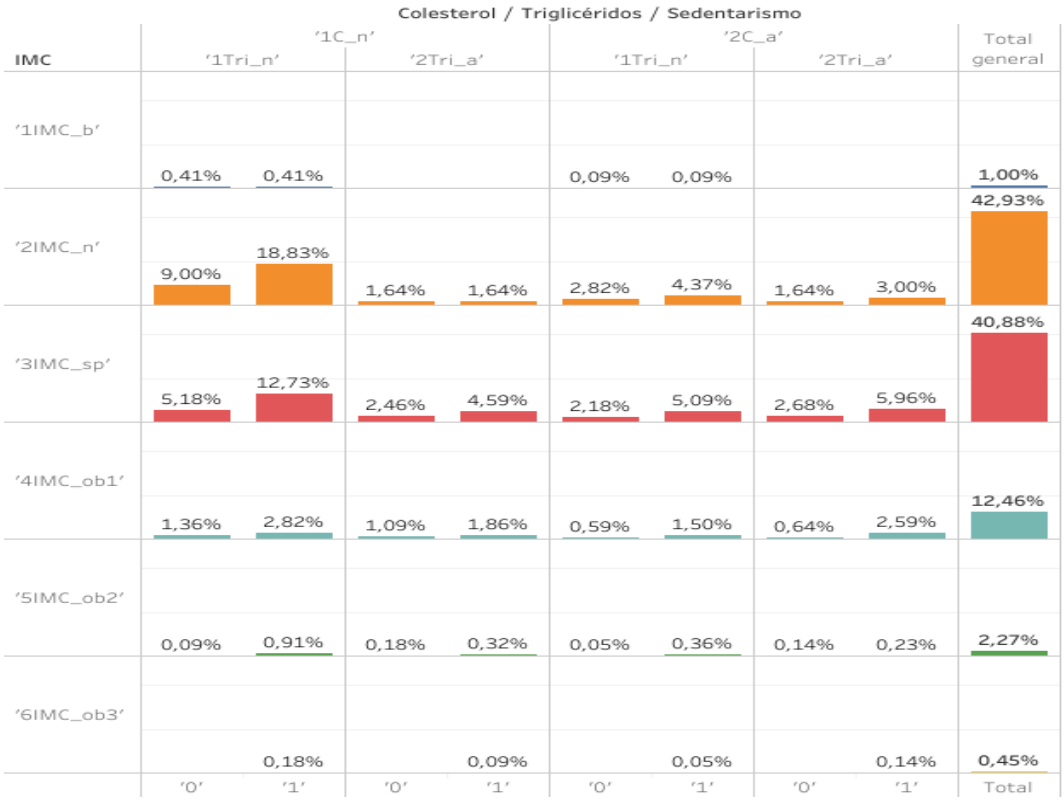


Figura 60. Distribución en porcentajes de la población del grupo Actividades Financieras de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo e IMC.

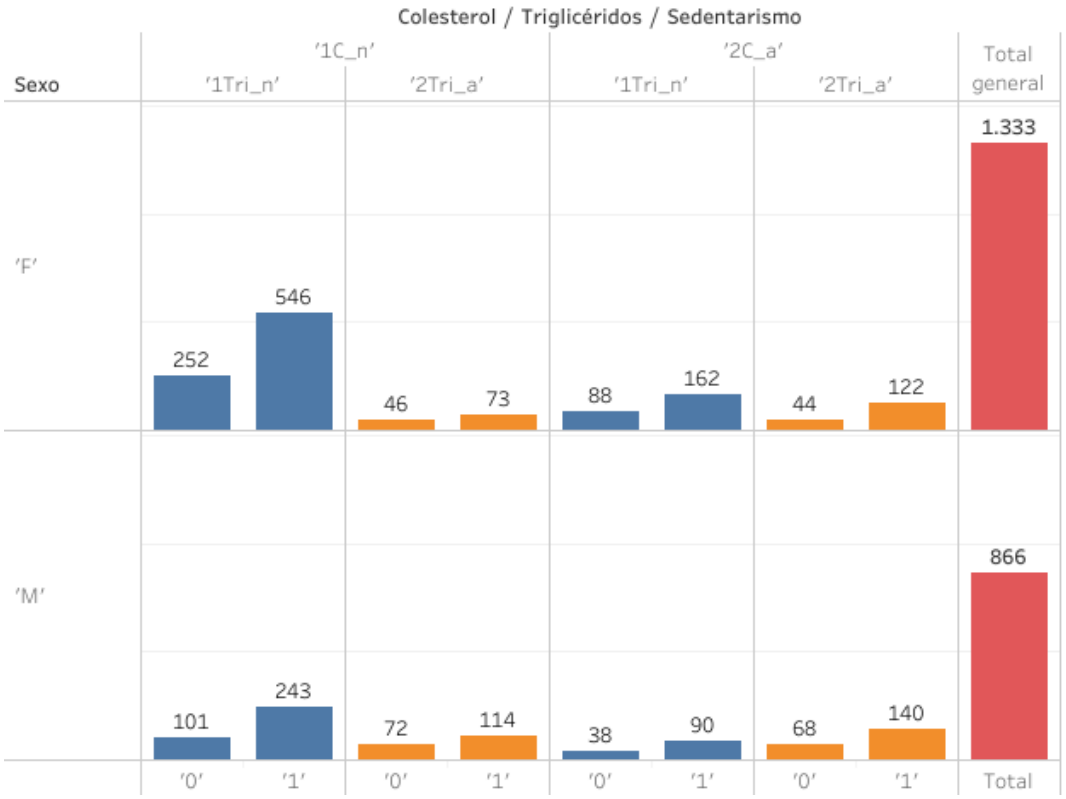


Figura 61. Distribución de la población del grupo Actividades Financieras de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo y Sexo.

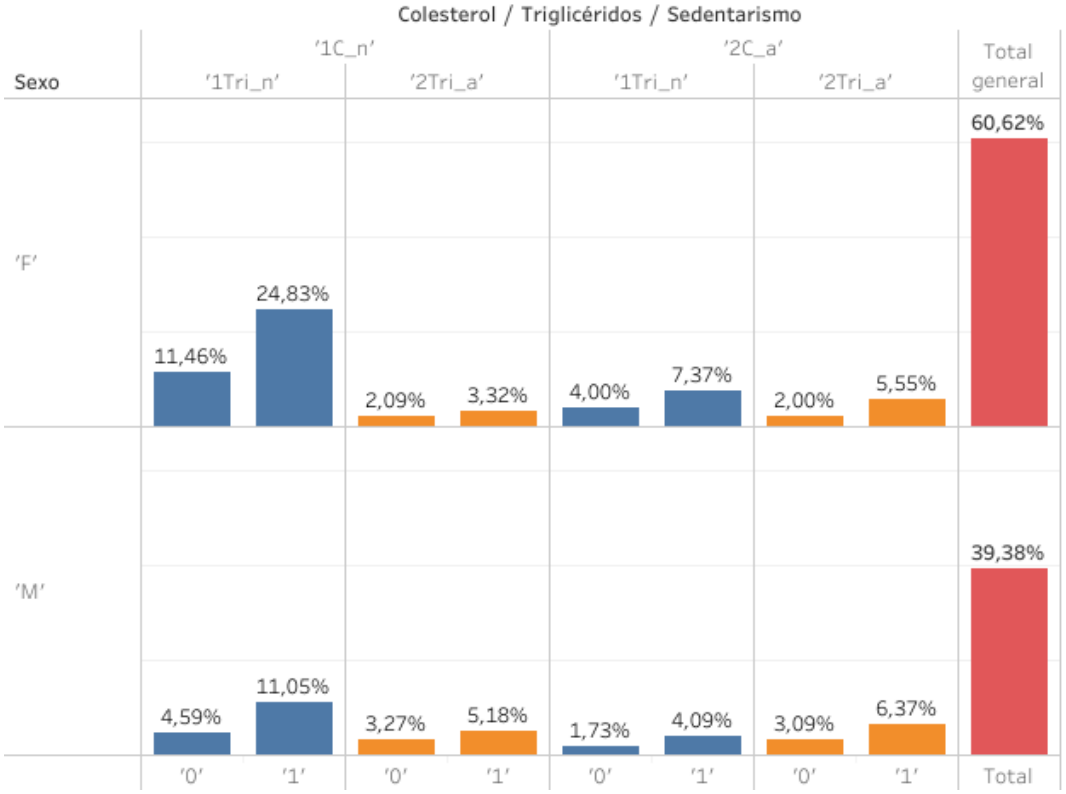


Figura 62. Distribución en porcentajes de la población del grupo Actividades Financieras de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo y Sexo.

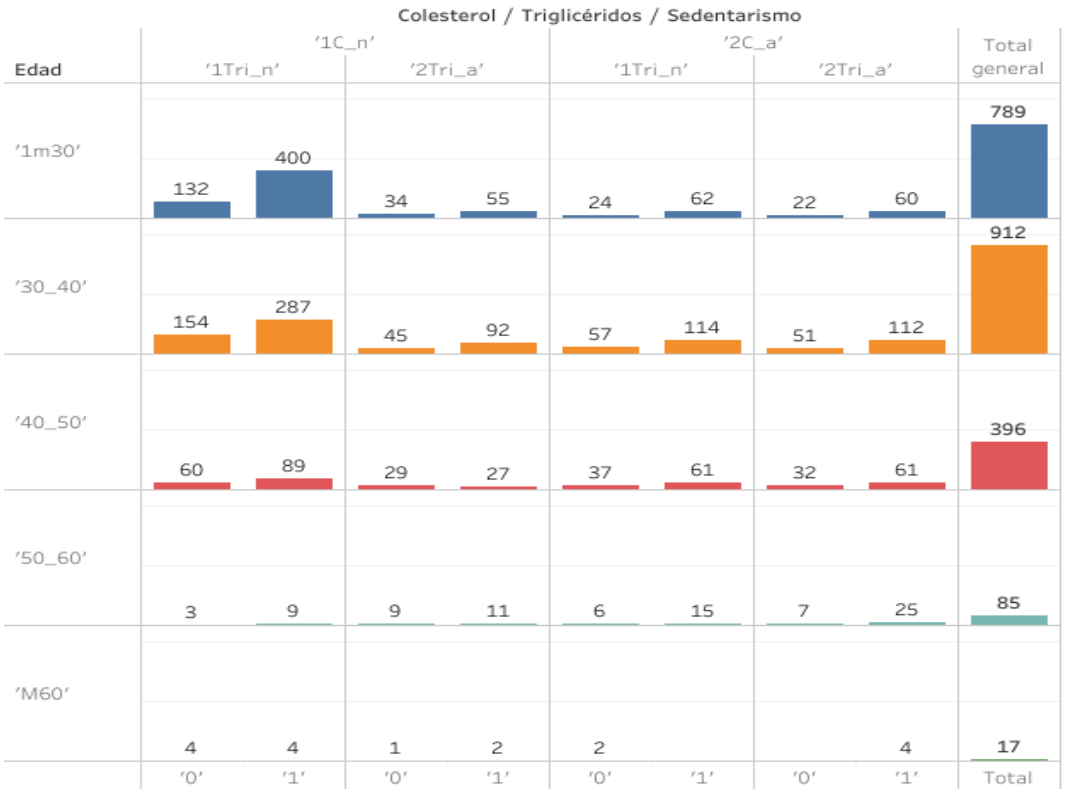


Figura 63. Distribución de la población del grupo Actividades Financieras de acuerdo a diversas categorías: Coleterol, Triglicéridos, Sedentarismo y Edad.

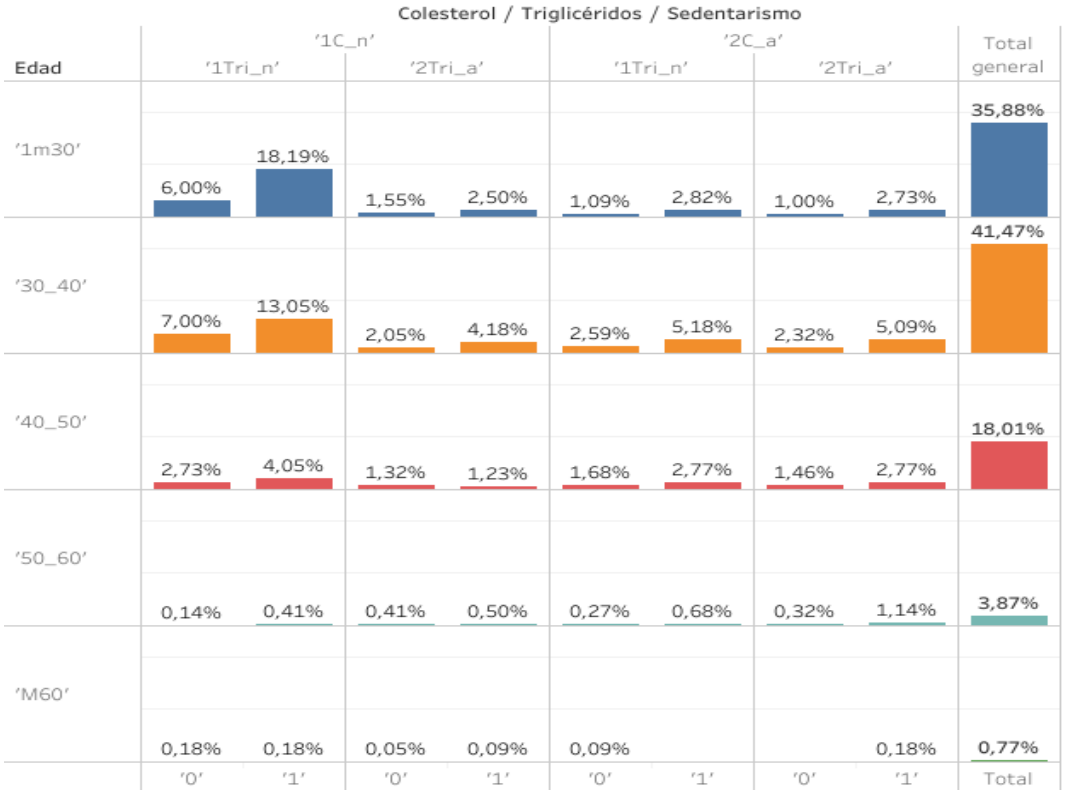


Figura 64. Distribución en porcentajes de la población del grupo Actividades Financieras de acuerdo a diversas categorías: Coleterol, Triglicéridos, Sedentarismo y Edad.

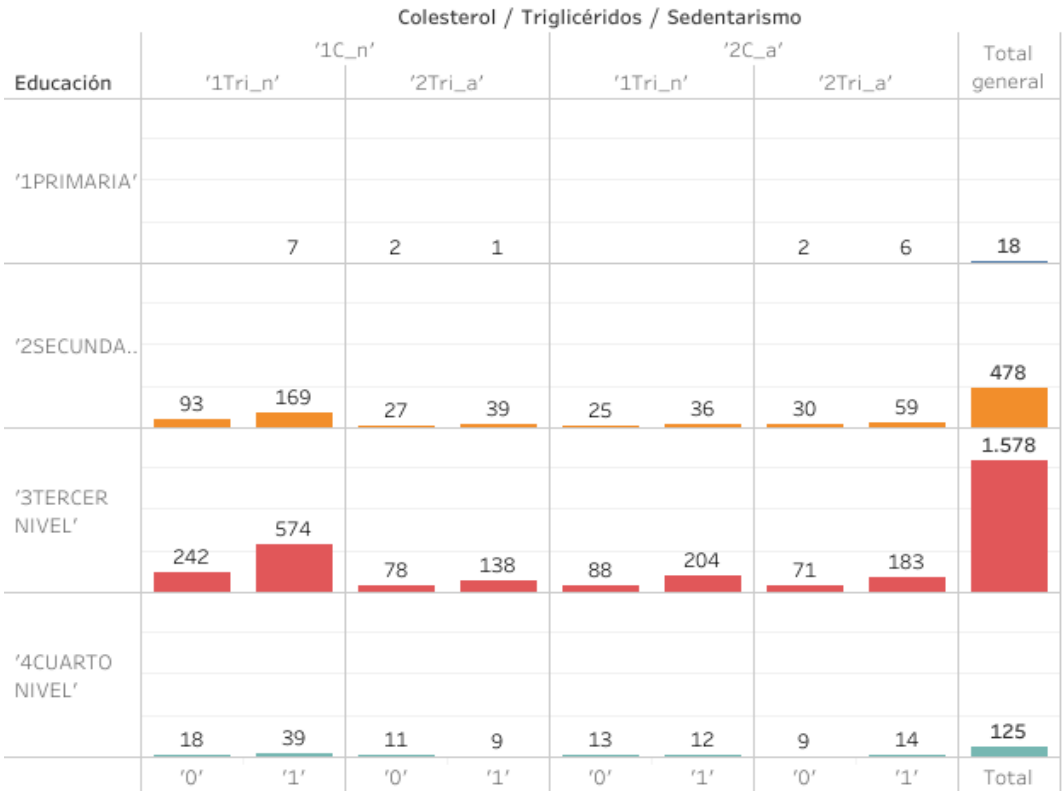


Figura 65. Distribución de la población del grupo Actividades Financieras de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo y Educación.

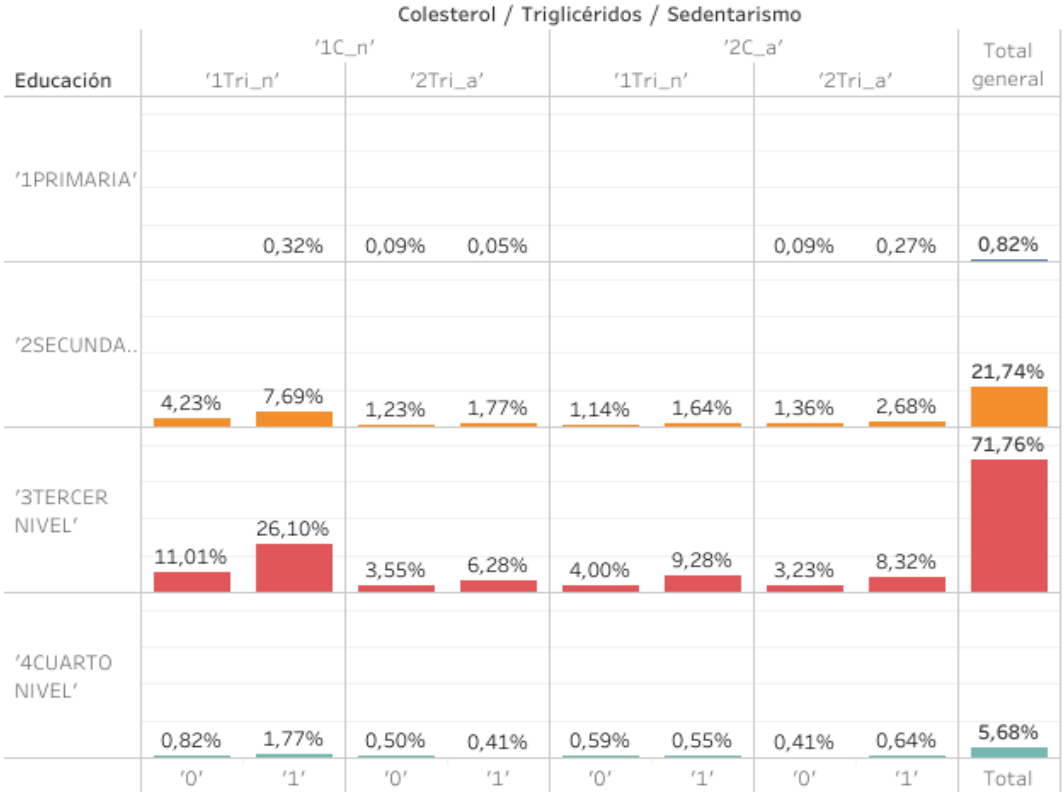


Figura 66. Distribución en porcentajes de la población del grupo Actividades Financieras de acuerdo a diversas categorías: Colesterol, Triglicéridos, Sedentarismo y Educación.

Anexo 5

Histogramas: población total y variables categóricas dicotómicas

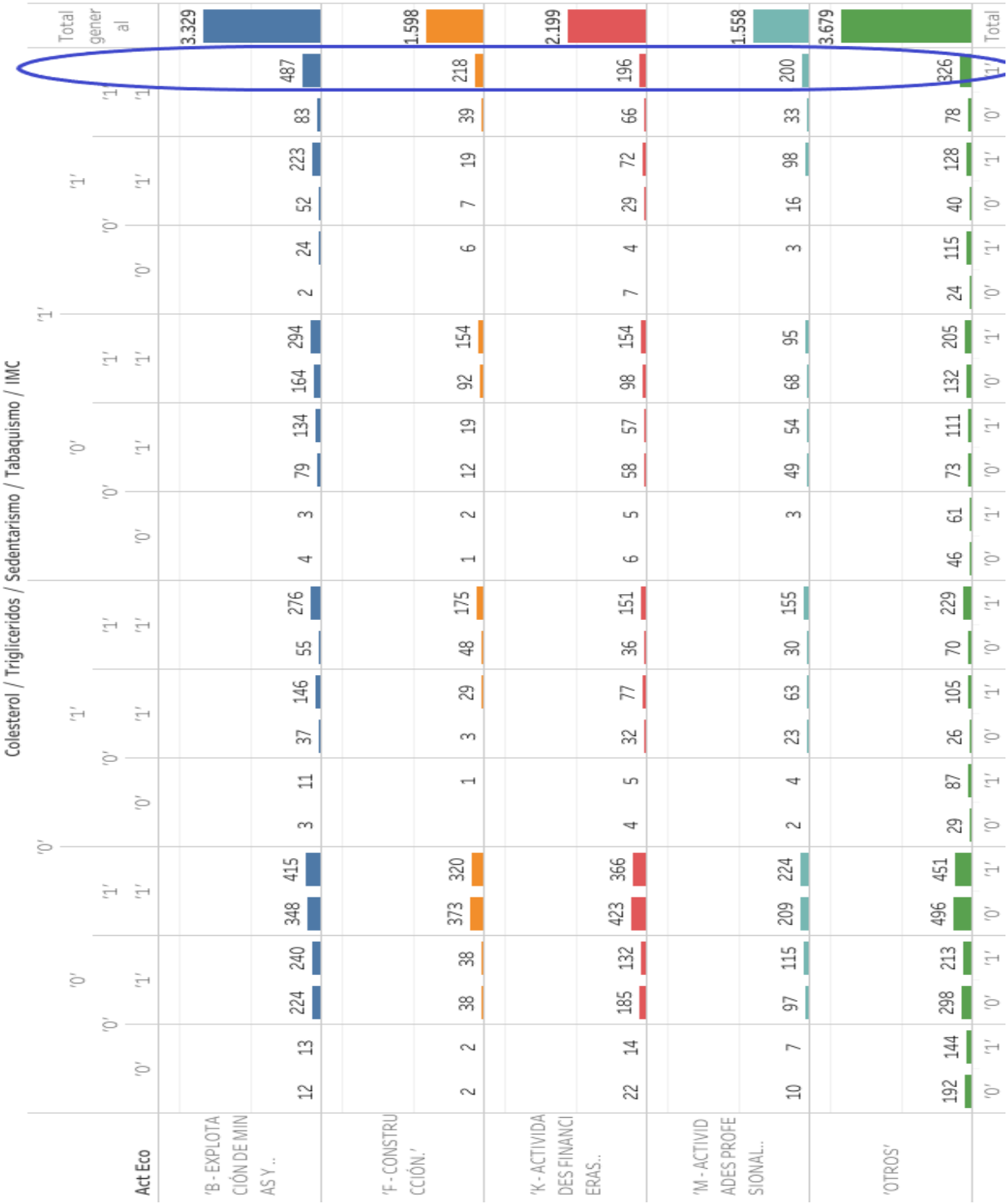


Figura 67. Distribución de la población total con variables dicotómicas: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC.

Figura 68. Distribución en porcentajes de la población total con variables dicotómicas: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC.

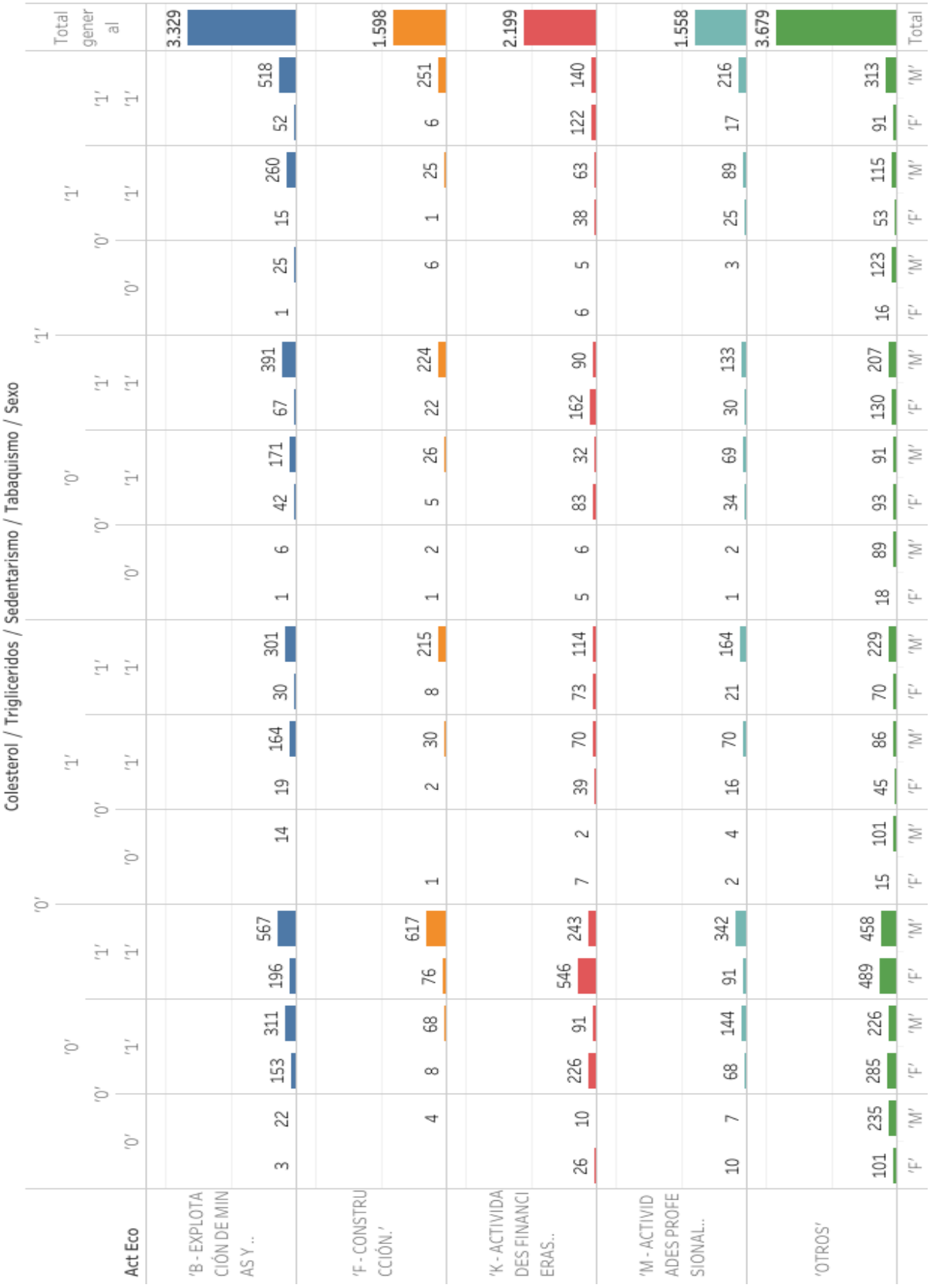


Figura 69. Distribución de la población total con variables dicotómicas: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Sexo.

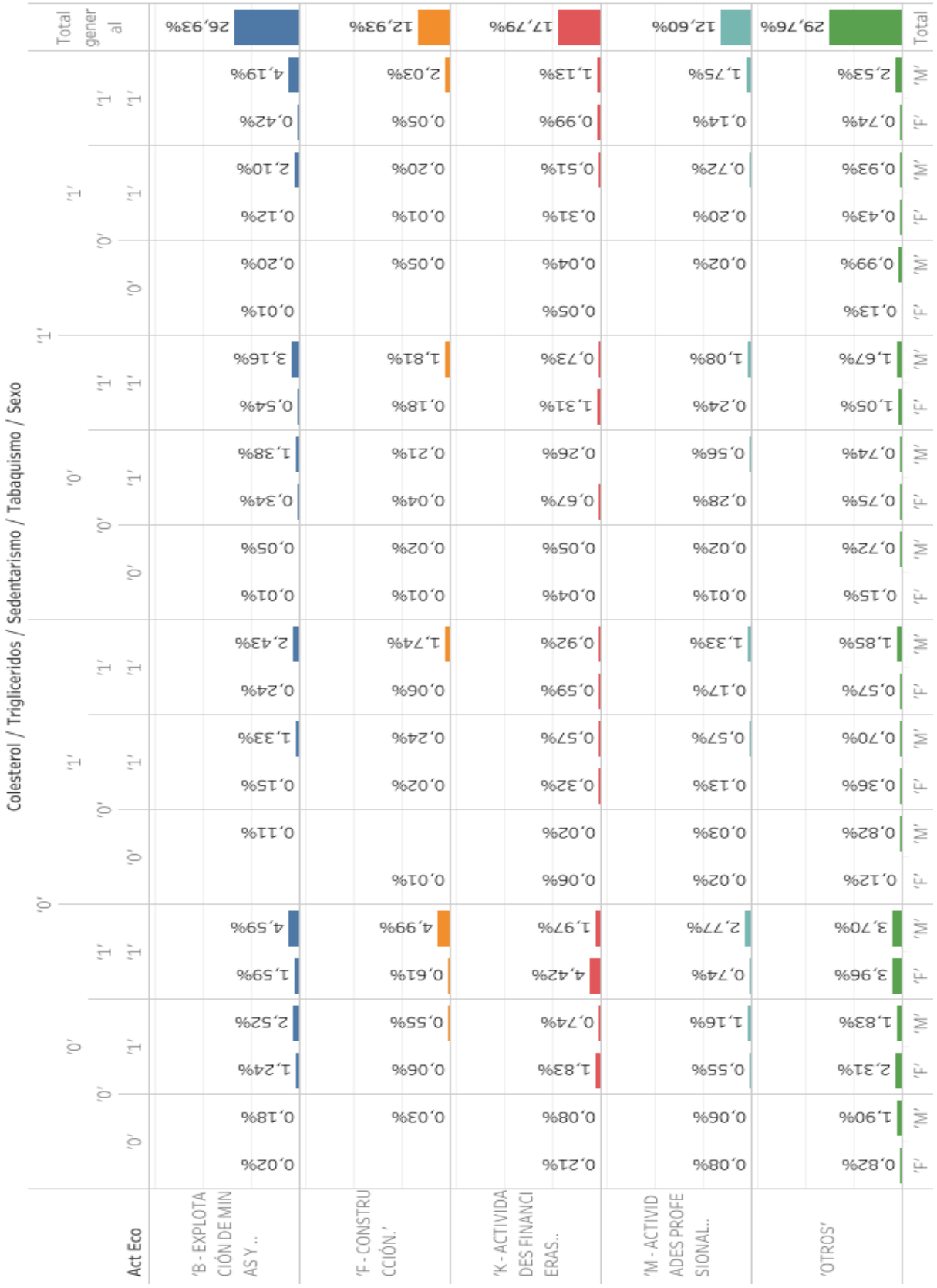


Figura 70. Distribución en porcentajes de la población total con variables dicotómicas: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Sexo.

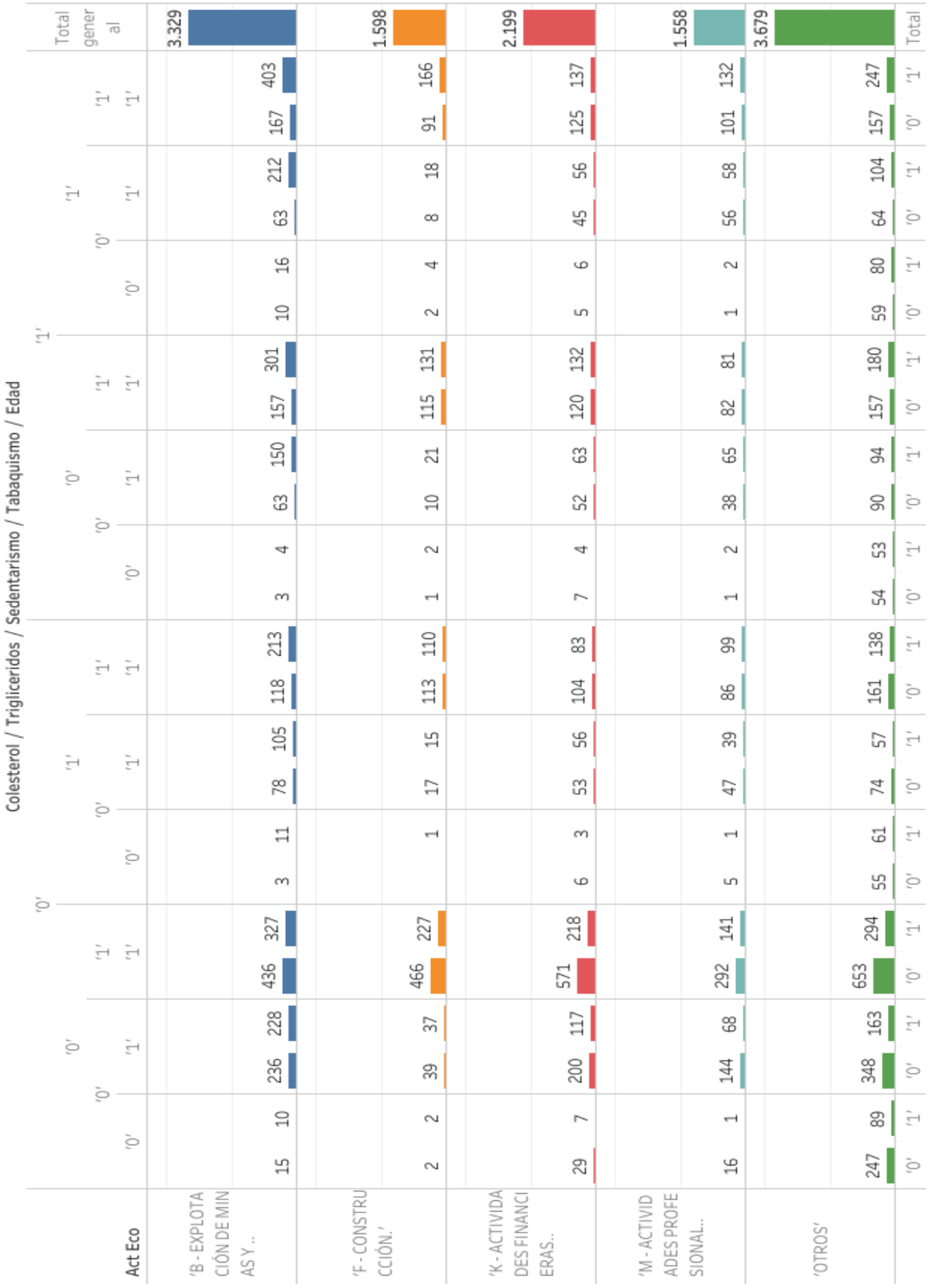


Figura 71. Distribución de la población total con variables dicotómicas: Actividad Económica, Coleterol, Triglicéridos, Sedentarismo, Tabaquismo y Edad.

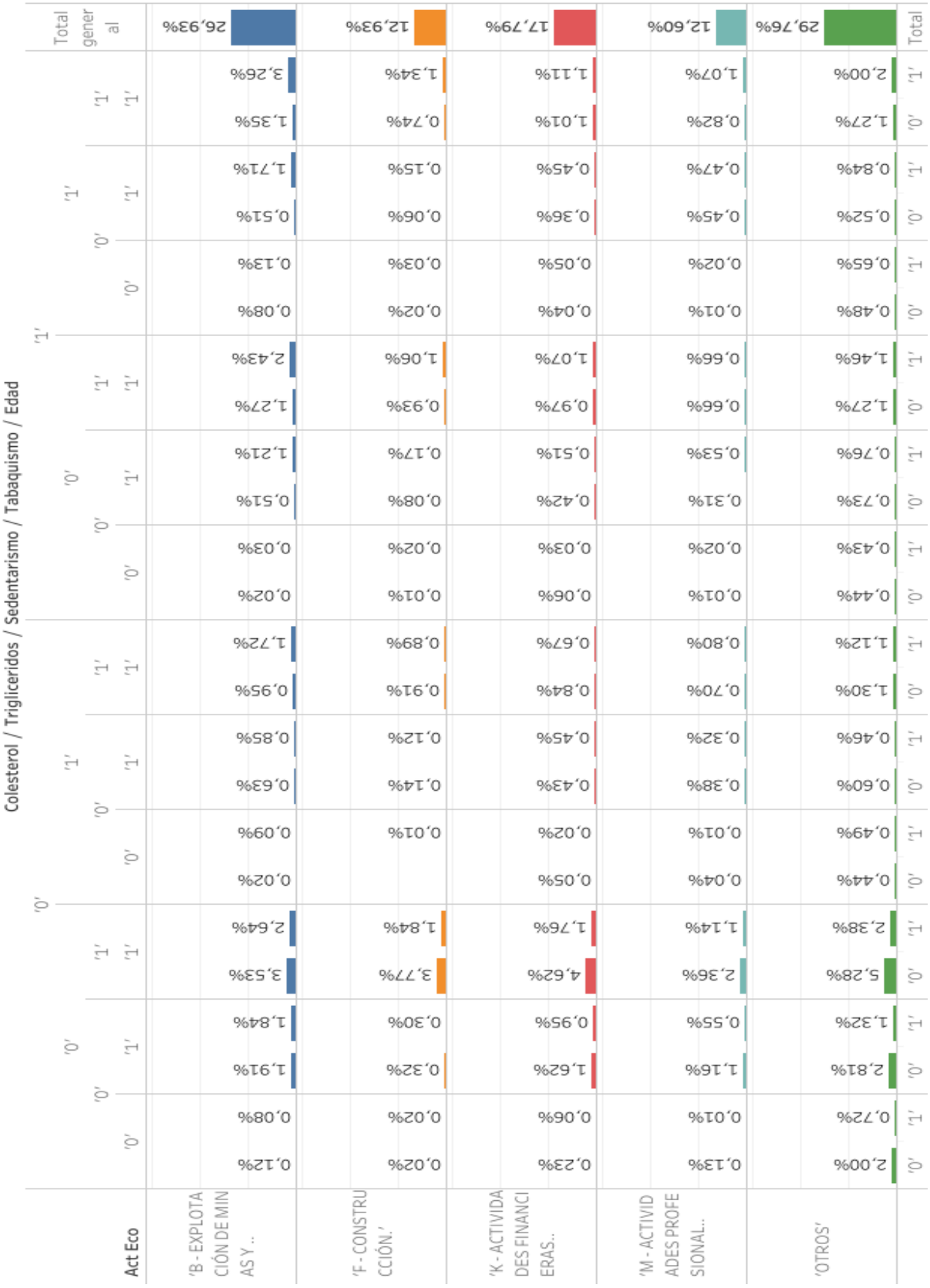


Figura 72. Distribución en porcentajes de la población total con variables dicotómicas: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Edad.

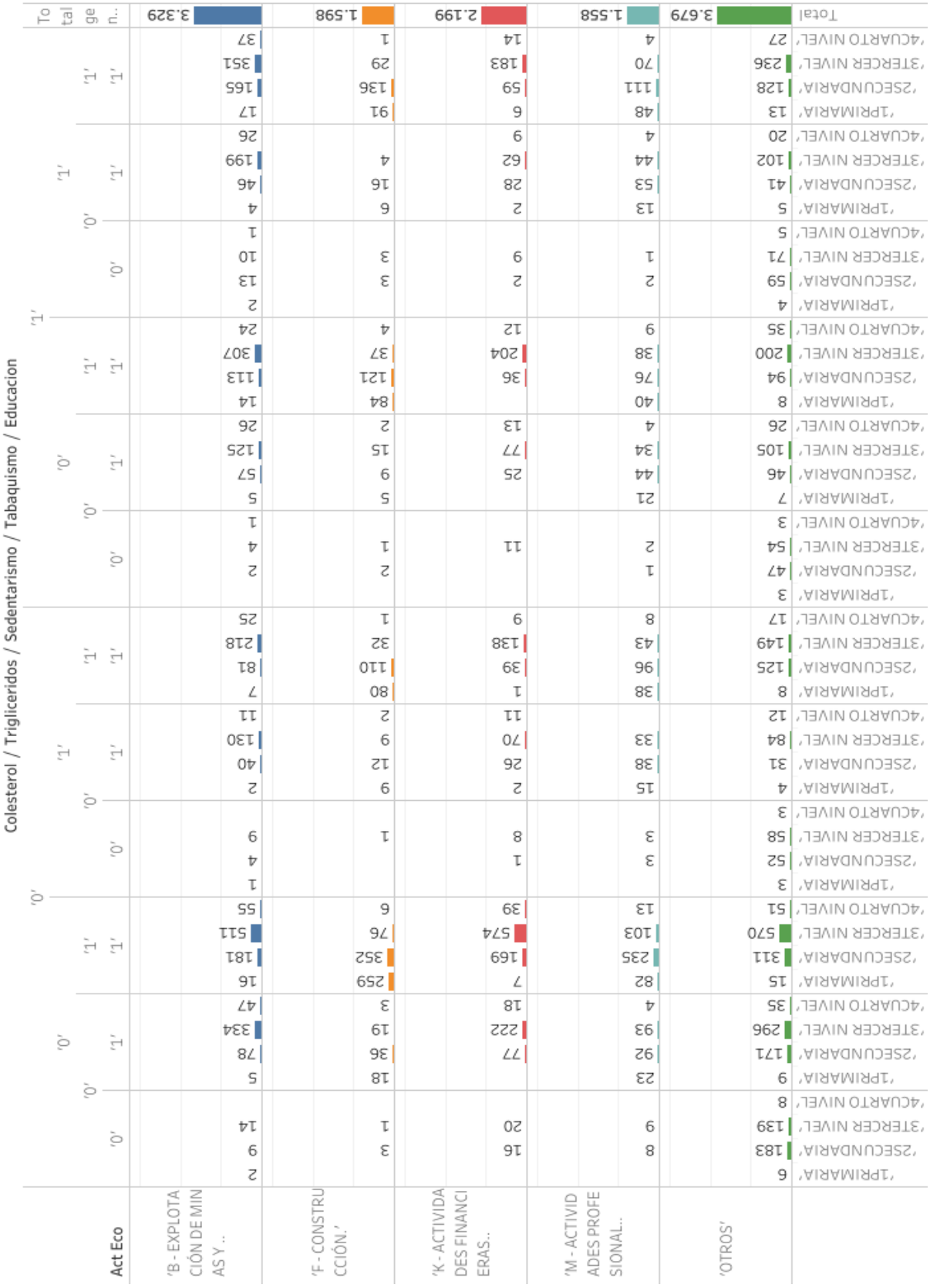


Figura 73. Distribución de la población total con variables dicotómicas: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo, y Educación.

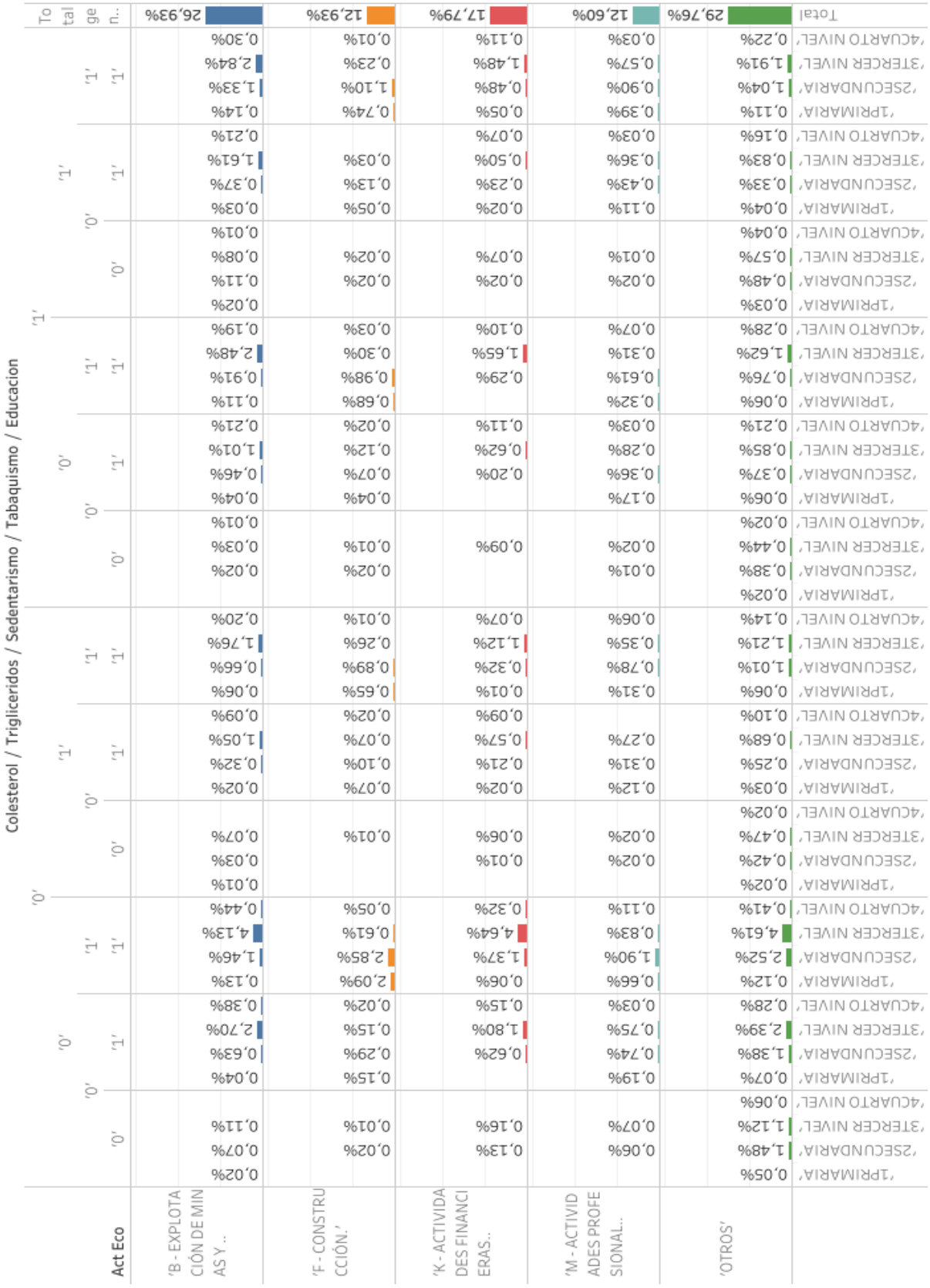


Figura 74. Distribución en porcentajes de la población total con variables dicotómicas: Actividad Económica, Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Educación.

Anexo 6

Histogramas: Explotación de Minas y variables categóricas dicotómicas

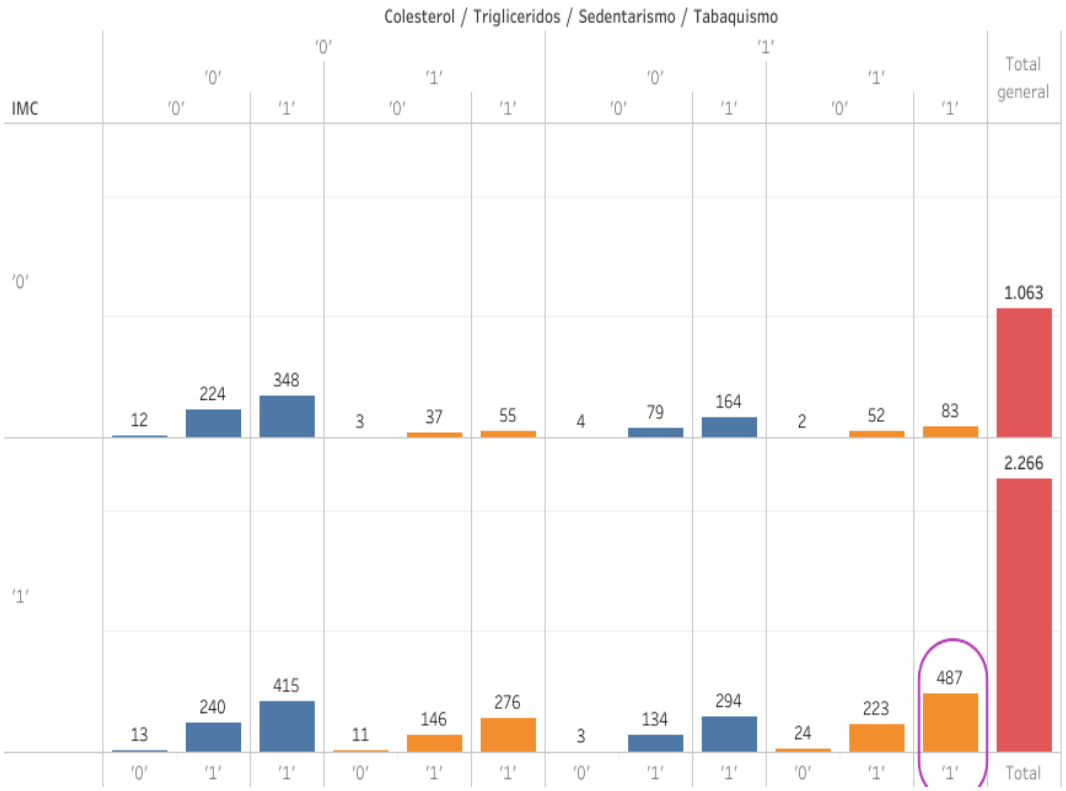


Figura 75. Distribución del grupo Explotación de Minas con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC.

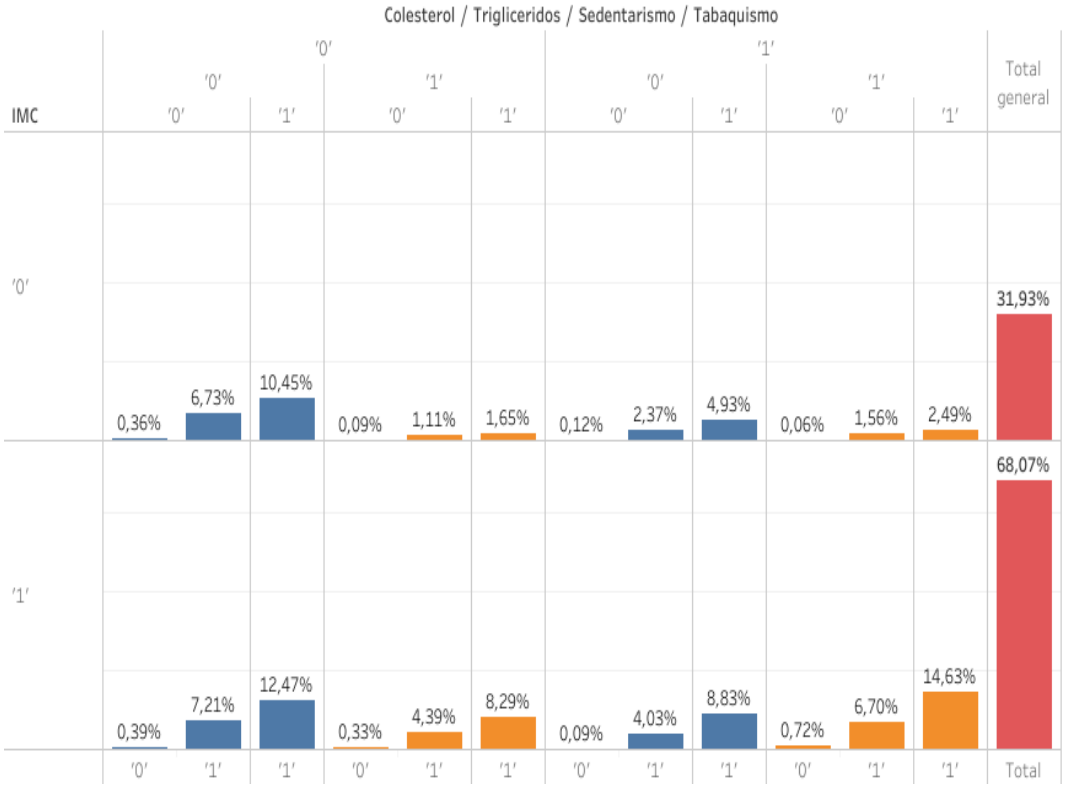


Figura 76. Distribución en porcentajes del grupo Explotación de Minas con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC.

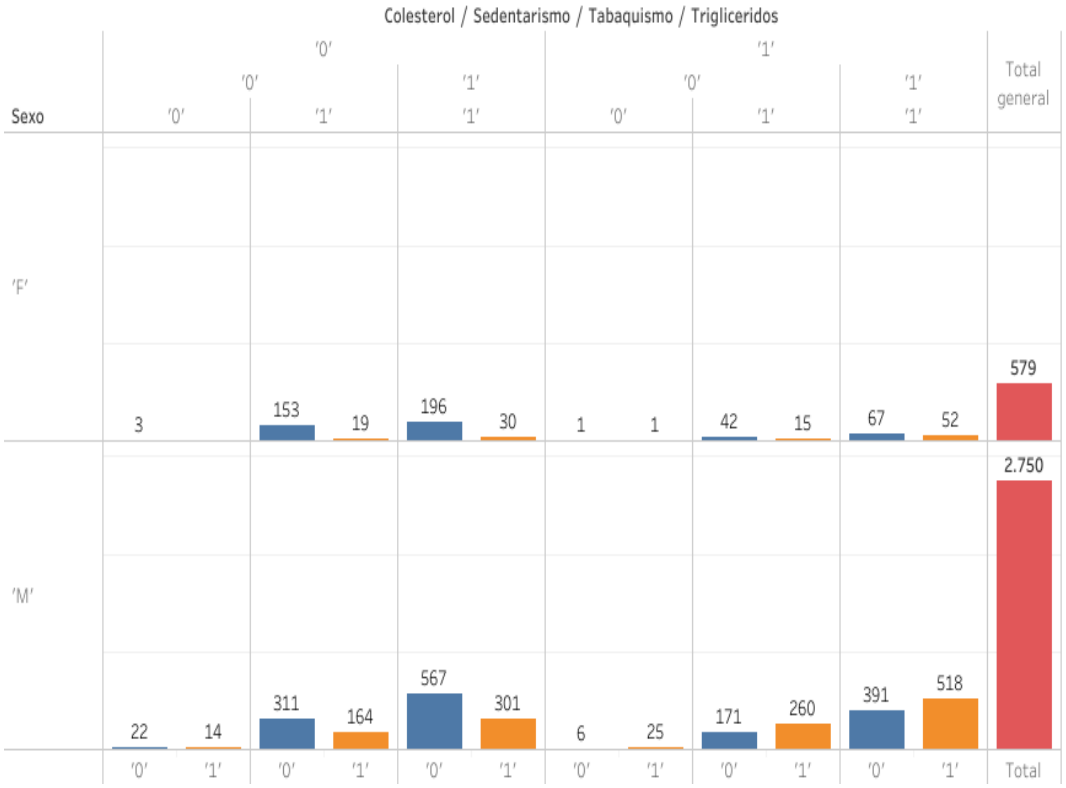


Figura 77. Distribución del grupo Explotación de Minas con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Sexo.

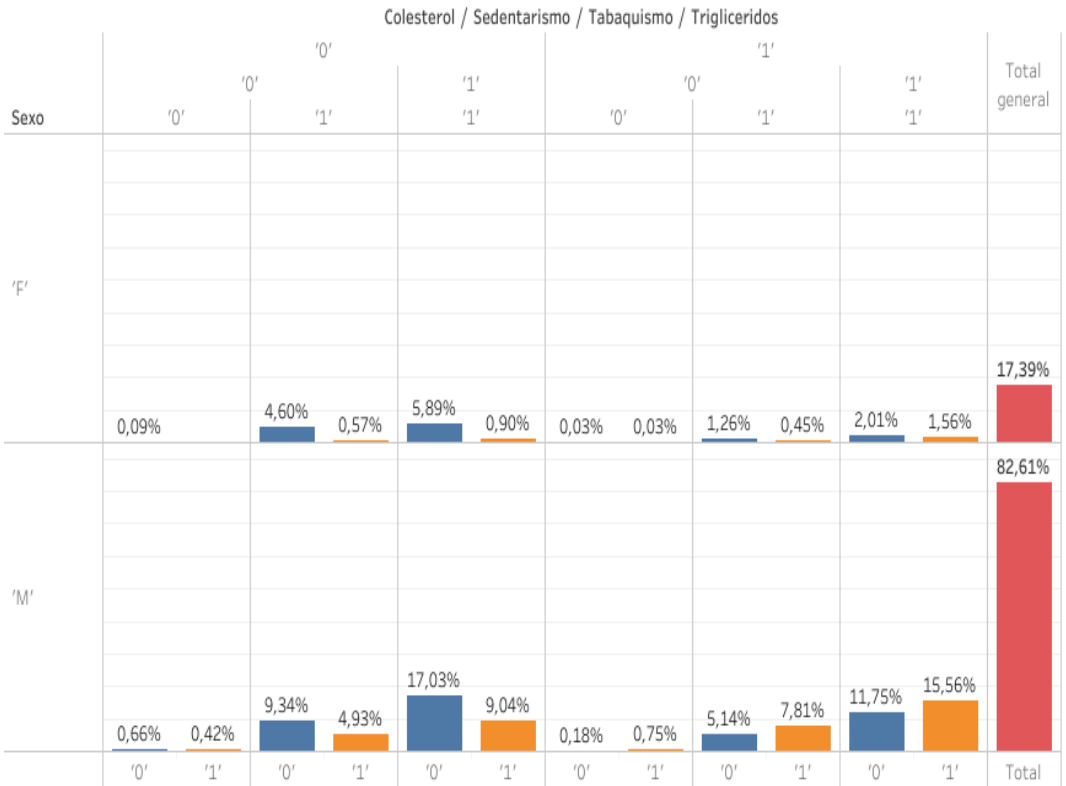


Figura 78. Distribución en porcentajes del grupo Explotación de Minas con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Sexo.

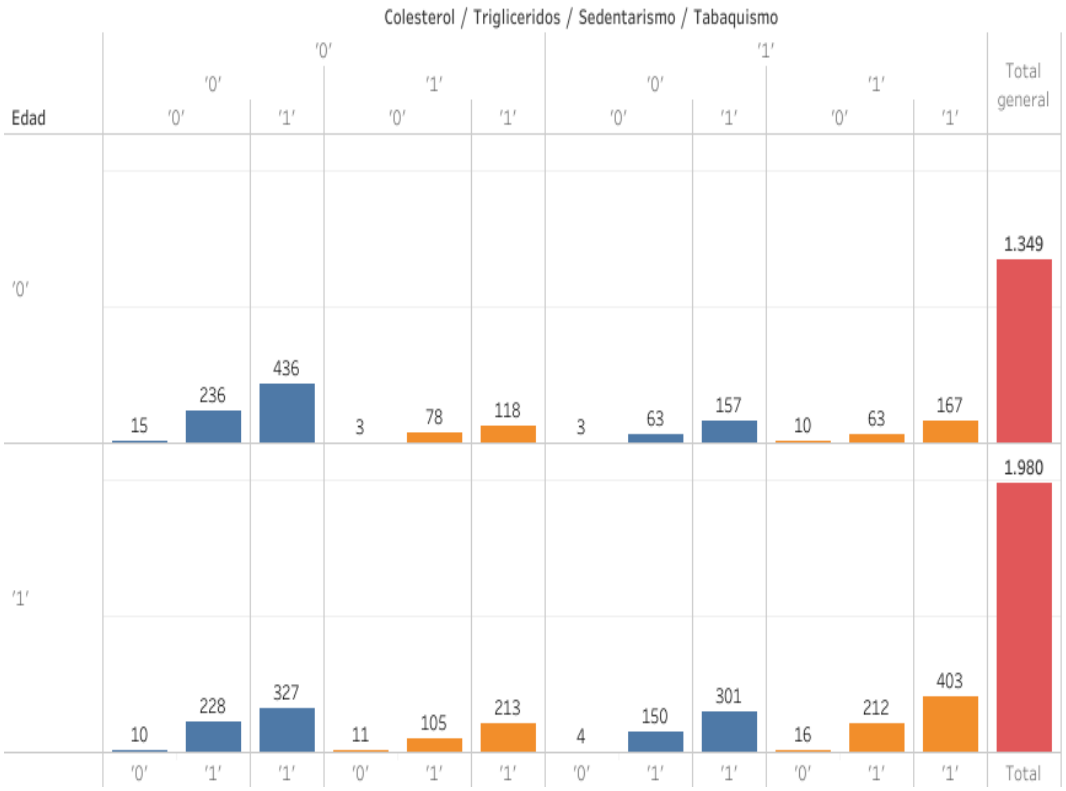


Figura 79. Distribución del grupo Explotación de Minas con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Edad.

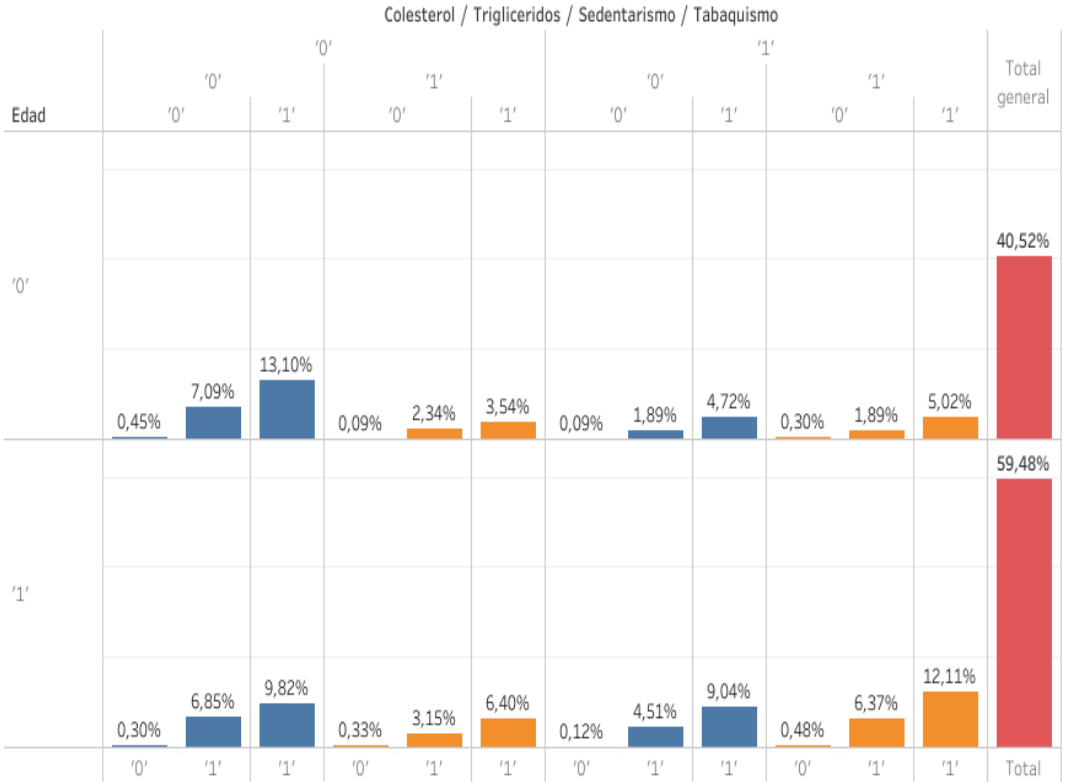


Figura 80. Distribución en porcentajes del grupo Explotación de Minas con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Edad.

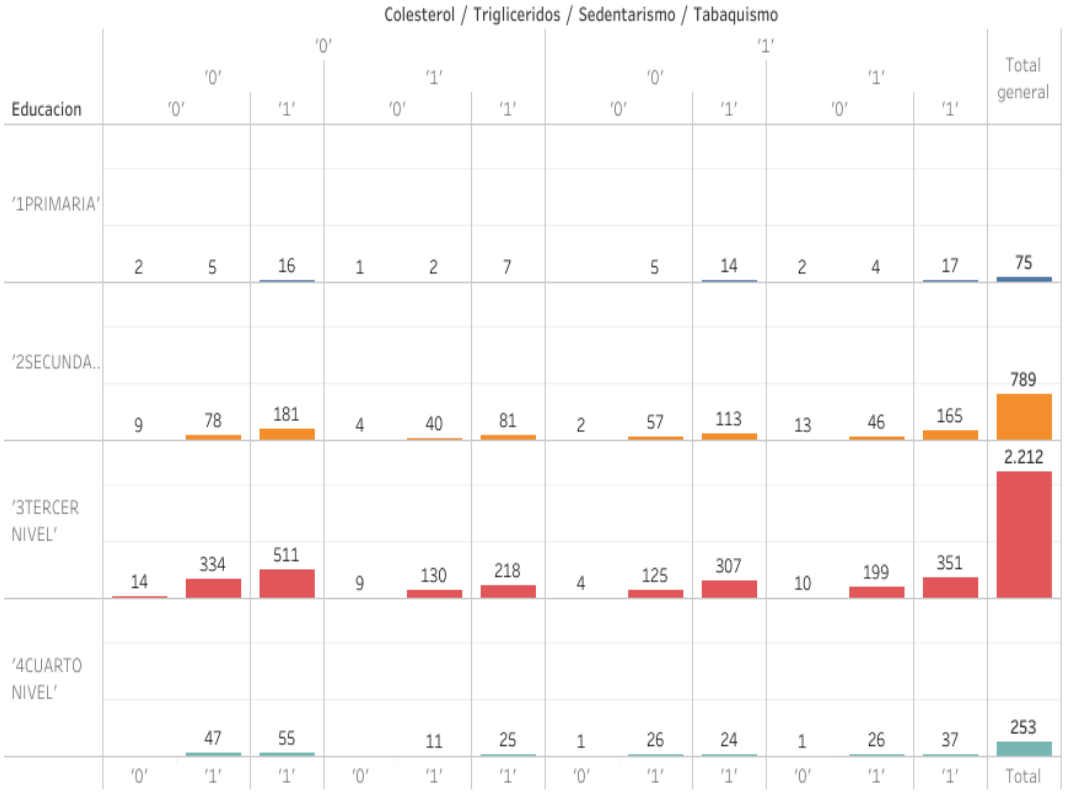


Figura 81. Distribución del grupo Explotación de Minas con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo y Educación.

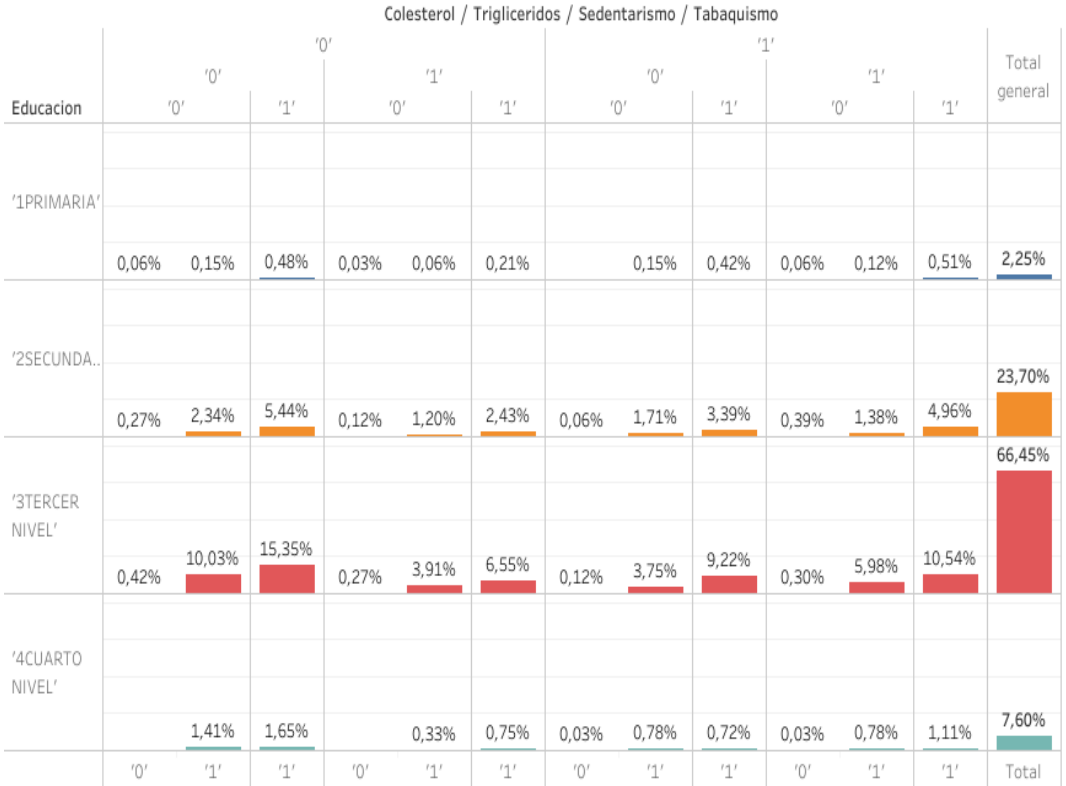


Figura 82. Distribución en porcentajes del grupo Explotación de Minas con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Educación.

Anexo 7

Histogramas: Actividades Finacieras y variables categóricas dicotómicas

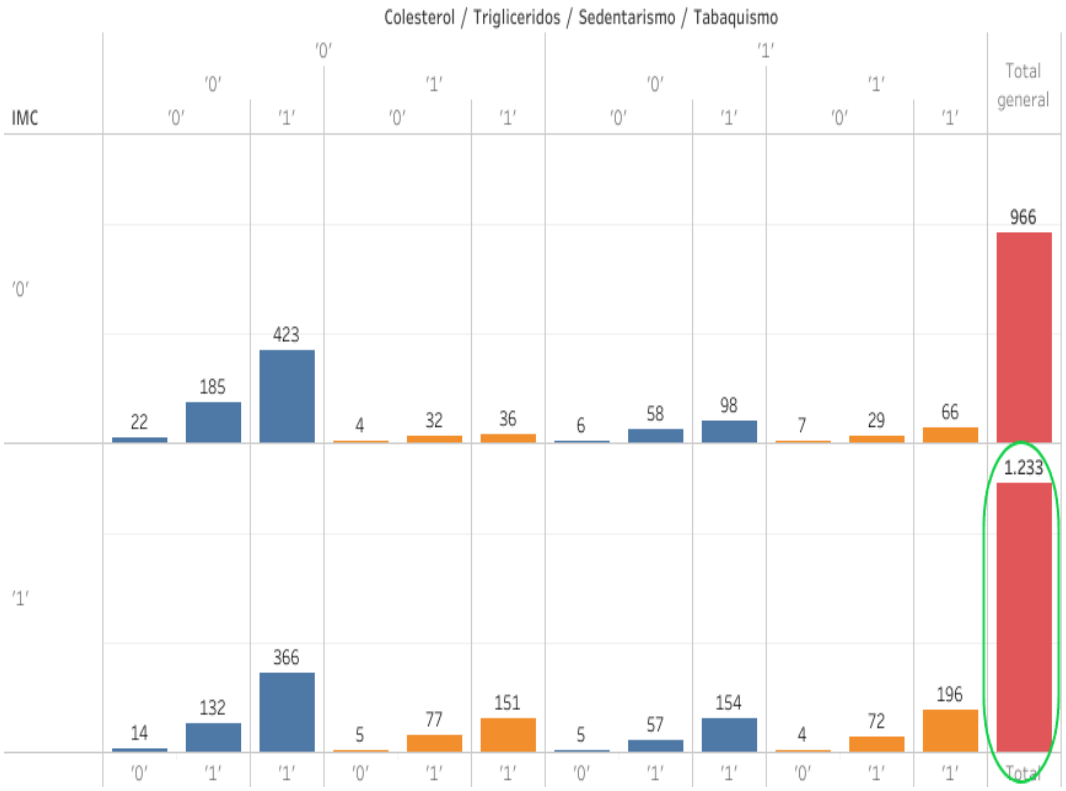


Figura 83. Distribución del grupo Actividades Financieras con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC.

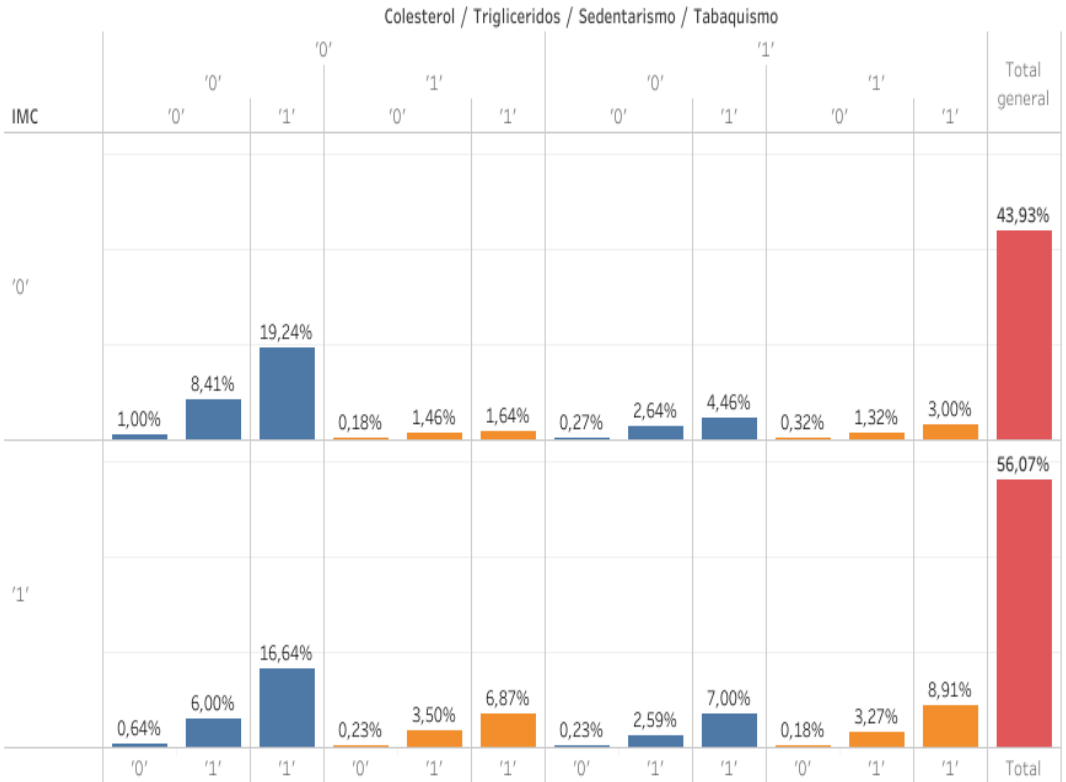


Figura 84. Distribución en porcentajes del grupo Actividades Financieras con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo e IMC.

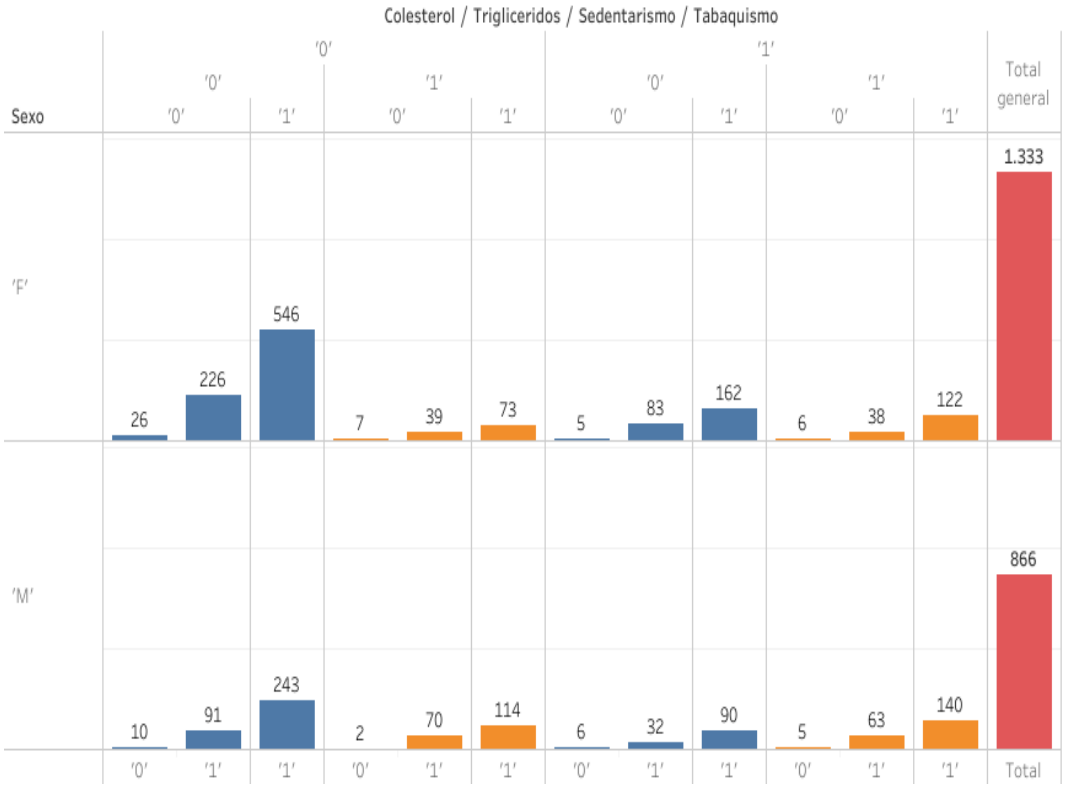


Figura 85. Distribución del grupo Actividades Financieras con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Sexo.

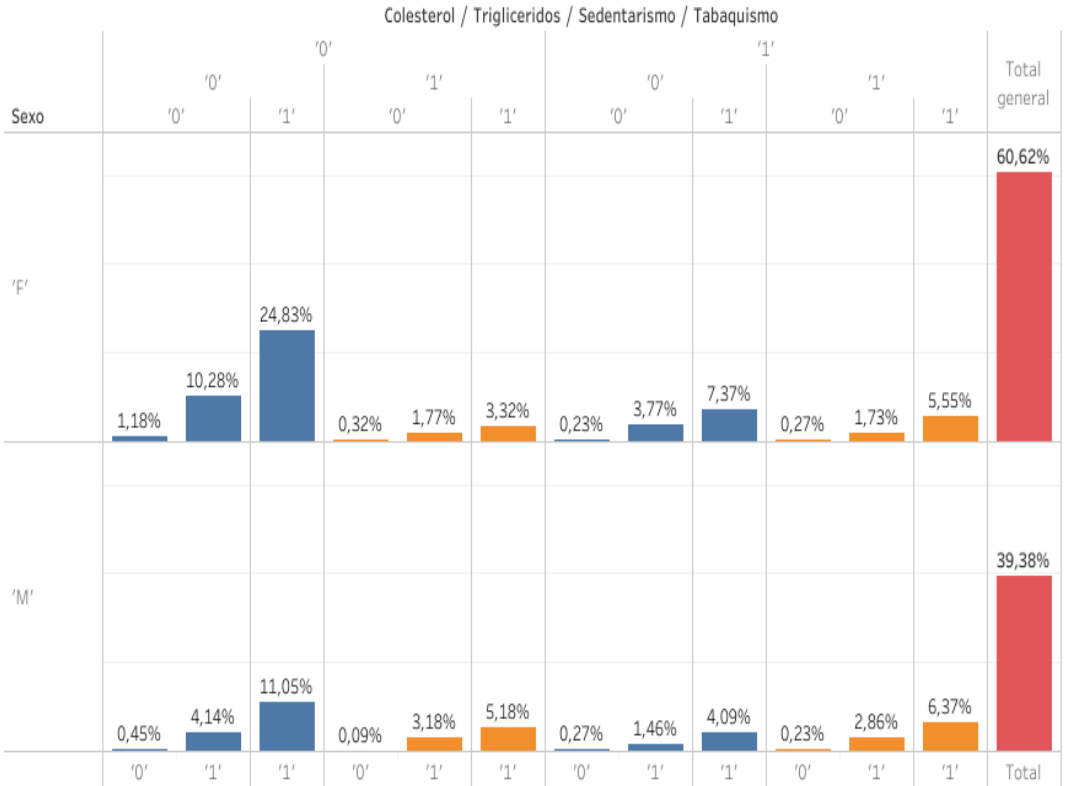


Figura 86. Distribución en porcentajes del grupo Actividades Financieras con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Sexo.

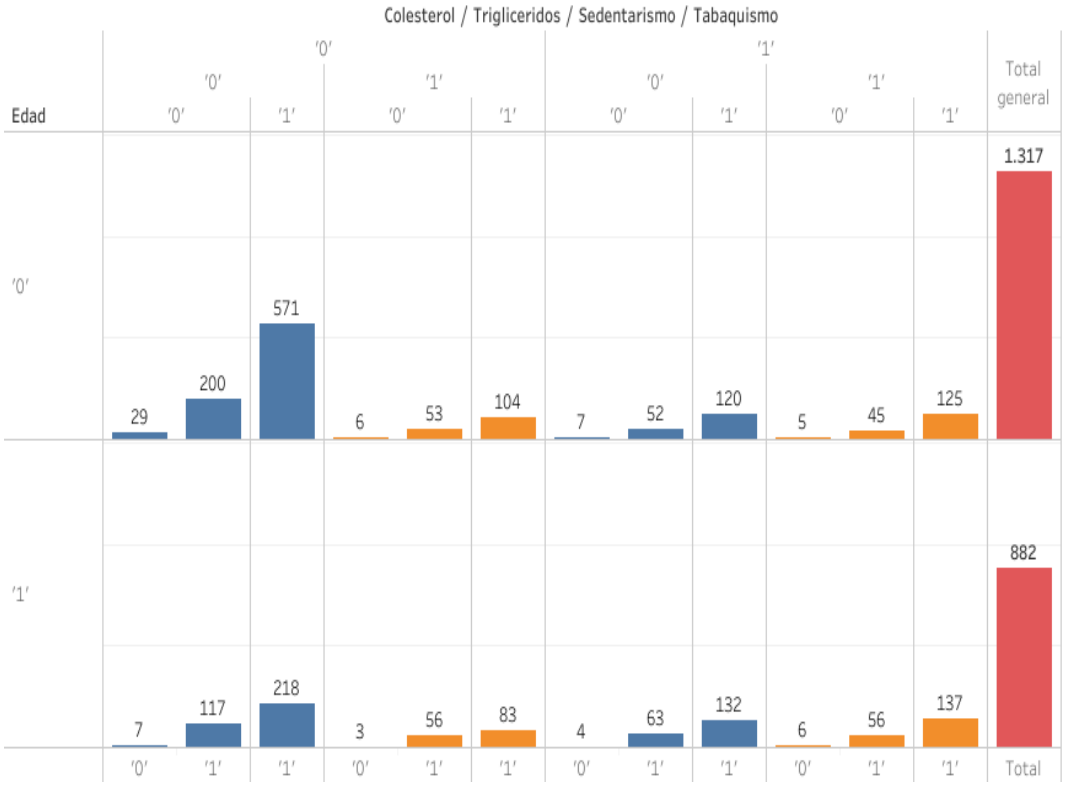


Figura 87. Distribución del grupo Actividades Financieras con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Edad.

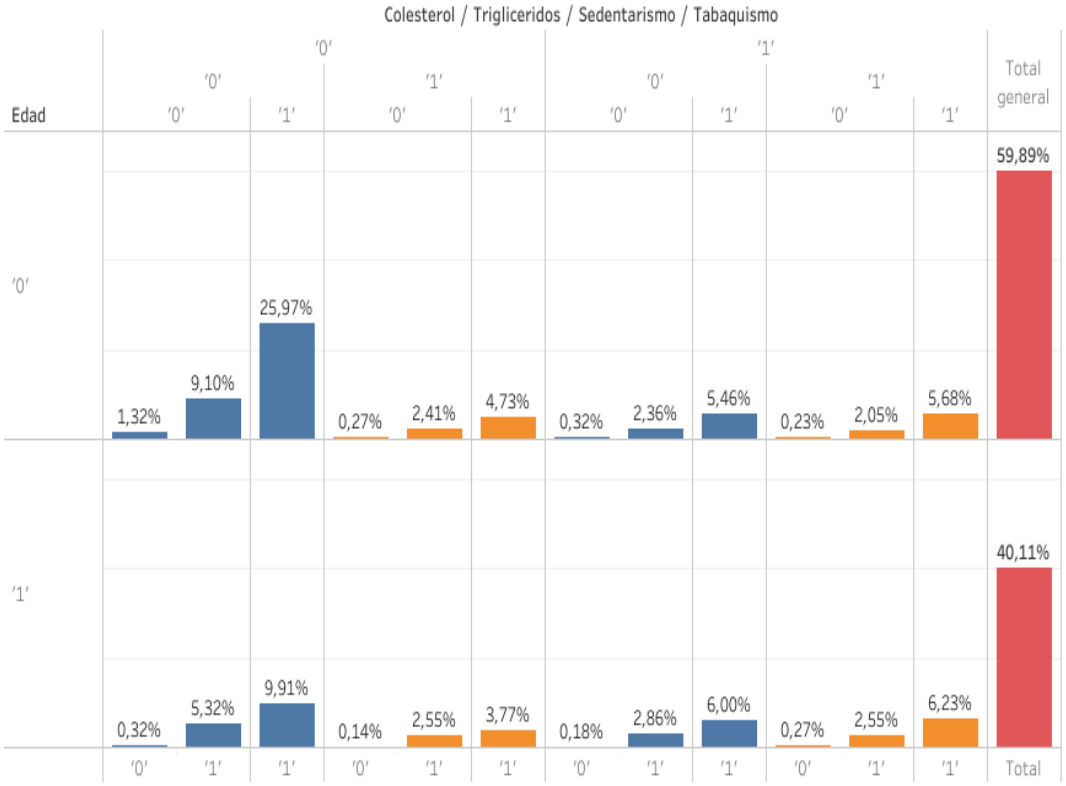


Figura 88. Distribución en porcentajes del grupo Actividades Financieras con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Edad.

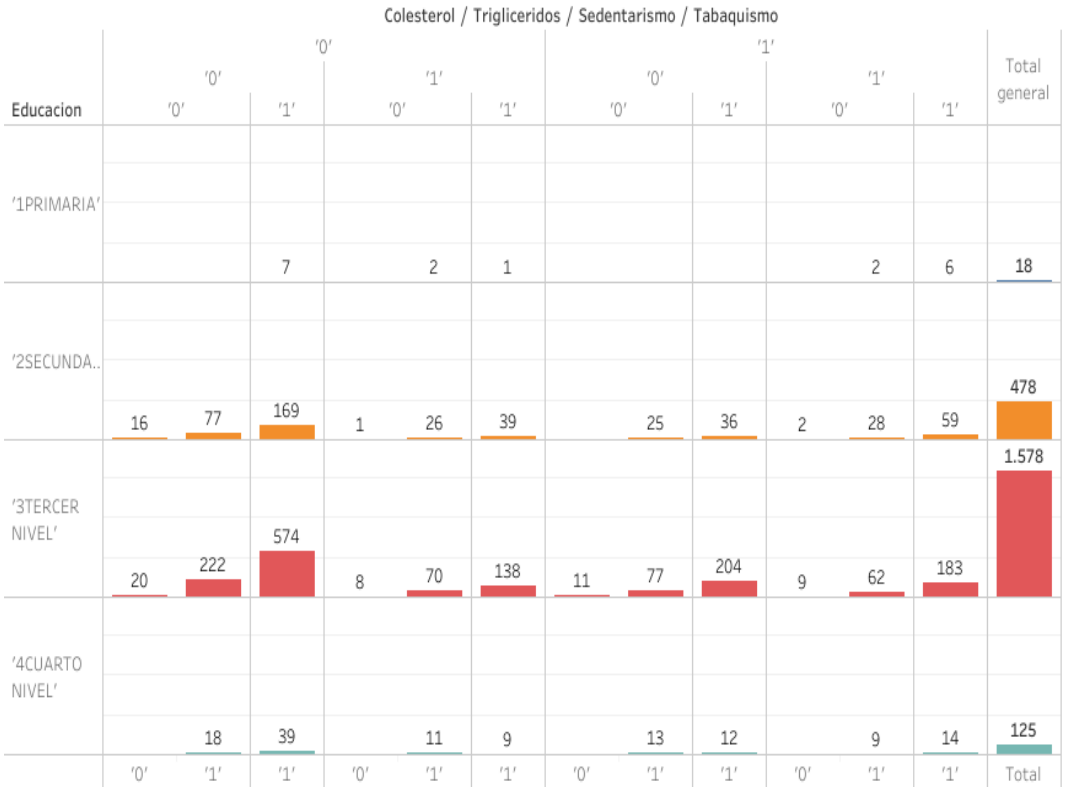


Figura 89. Distribución del grupo Actividades Financieras con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Educación.

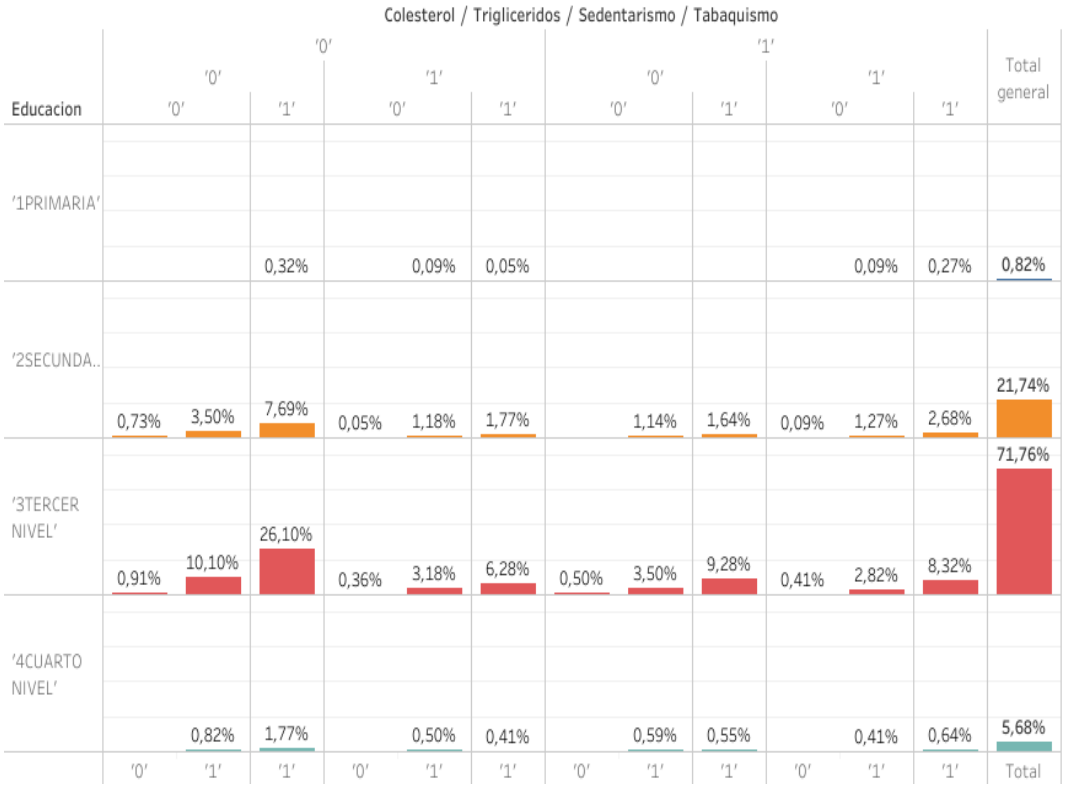


Figura 90. Distribución en porcentajes del grupo Actividades Financieras con variables dicotómicas: Colesterol, Triglicéridos, Sedentarismo, Tabaquismo y Educación.

Anexo 8

Pseudo-códigos de programas utilizados

Algorithm 1 Eliminación de observaciones que no contengan datos en alguna de las variables escogidas de la data.

1: $load\ Data\ 03$
2: $A \leftarrow Data\ 03(:, [1, 5, 18 : 20, 32 : 34])$
3: $A \leftarrow table2array(A)$
4: $ind \leftarrow all(\ isnan(A), 2)$
5: $B \leftarrow Data\ 03(ind, [1, 2, 5, 8 : 13, 18 : 20, 32 : 35,])$

Algorithm 2 Estadístico descriptivo de Índice de Masa Corporal.

1: $A \leftarrow summary(B)$
2: $ElemenGrpImc \leftarrow A.imcCat.Counts$
3: $porc \leftarrow round(ElemenGrpImc * 1000/12363)/10$
4: $C \leftarrow double(B.imc)$
5: $a \leftarrow mean(C)$
6: $b \leftarrow median(C)$
7: $d \leftarrow var(C)$
8: $[media, varianza, mediana, grupos] \leftarrow grpstats(B.imc, B.imcCat, ...$
9: $['mean','var','median','gname'])$
10: $ResultadosImc \leftarrow dataset(grupos, ElemenGrpImc, media, varianza, mediana, porc)$
11: $histogram(B.imc)$

Algorithm 3 Estadístico descriptivo de Edad.

1: $A \leftarrow summary(B)$
2: $ElemenGrpImc \leftarrow A.EdadCat.Counts$
3: $porc \leftarrow round(ElemenGrpEdad * 1000/12363)/10$
4: $C \leftarrow double(B.imc)$
5: $a \leftarrow mean(C)$
6: $b \leftarrow median(C)$
7: $d \leftarrow var(C)$
8: $[media, varianza, mediana, grupos] \leftarrow grpstats(B.Edad, B.EdadCat, ...$
9: $['mean','var','median','gname'])$
10: $ResultadosEdad \leftarrow dataset(grupos, ElemenGrpEdad, media, varianza, mediana, porc)$
11: $histogram(B.Edad)$

Algorithm 4 Estadístico descriptivo de Glucosa.

$A \leftarrow summary(B)$
$ElemenGrpGlu \leftarrow A.GLUCOSACat.Counts$
$porc \leftarrow round(ElemenGrpGlu * 1000/12363)/10$
$C \leftarrow double(B.GLUCOSA)$
$a \leftarrow mean(C)$
$b \leftarrow median(C)$
$d \leftarrow var(C)$
$[media, varianza, mediana, grupos] \leftarrow grpstats(B.GLUCOSA, ...$
$B.GLUCOSACat, ['mean','var','median','gname'])$
$ResultadosGlu \leftarrow dataset(grupos, ElemenGrpGlu, media, varianza, mediana, porc)$
$histogram(B.GLUCOSA)$

Algorithm 5 Estadístico descriptivo de Colesterol.

$A \leftarrow summary(B)$
$ElemenGrpCol \leftarrow A.COLESTEROLCat.Counts$
$porc \leftarrow round(ElemenGrpCol * 1000/12363)/10$
$C \leftarrow double(B.COLESTEROL)$
$a \leftarrow mean(C)$
$b \leftarrow median(C)$
$d \leftarrow var(C)$
$[media, varianza, mediana, grupos] \leftarrow grpstats(B.COLESTEROL, ...$
$B.COLESTEROLCat, ['mean','var','median','gname'])$
$ResultadosCol \leftarrow dataset(grupos, ElemenGrpCol, media, varianza, mediana, porc)$
$histogram(B.COLESTEROL)$

Algorithm 6 Estadístico descriptivo de Triglicéridos.

$A \leftarrow summary(B)$
$ElemenGrpTri \leftarrow A.TRIGLICERIDOSCat.Counts$
$porc \leftarrow round(ElemenGrpTri * 1000/12363)/10$
$C \leftarrow double(B.TRIGLICERIDOS)$
$a \leftarrow mean(C)$
$b \leftarrow median(C)$
$d \leftarrow var(C)$
$[media, varianza, mediana, grupos] \leftarrow grpstats(B.TRIGLICERIDOS, ...$
$B.TRIGLICERIDOSCat, ['mean','var','median','gname'])$
$ResultadosTri \leftarrow dataset(grupos, ElemenGrpTri, media, varianza, mediana, porc)$
$histogram(B.TRIGLICERIDOS)$

Algorithm 7 Estadístico descriptivo de Presión Sistólica.

1: $A \leftarrow summary(B)$
2: $ElemenGrpPre1 \leftarrow A.presion\ arterialCat.Counts$
3: $porc \leftarrow round(ElemenGrpPre1 * 1000/12363)/10$
4: $C \leftarrow double(B.presion\ arterial)$
5: $a \leftarrow mean(C)$
6: $b \leftarrow median(C)$
7: $d \leftarrow var(C)$
8: $[media, varianza, mediana, grupos] \leftarrow grpstats(B.presion\ arterial, ...$
9: $B.presion\ arterialCat, ['mean','var','median','gname'])$
10: $ResultadosPre1 \leftarrow dataset(grupos, ElemenGrpPre1, media, varianza, mediana, porc)$
11: $histogram(B.presion\ arterial)$

Algorithm 8 Estadístico descriptivo de Presión Distólica.

1: $A \leftarrow summary(B)$
2: $ElemenGrpPre2 \leftarrow A.presion\ arterial2Cat.Counts$
3: $porc \leftarrow round(ElemenGrpPre2 * 1000/12363)/10$
4: $C \leftarrow double(B.presion\ arterial2)$
5: $a \leftarrow mean(C)$
6: $b \leftarrow median(C)$
7: $d \leftarrow var(C)$
8: $[media, varianza, mediana, grupos] \leftarrow grpstats(B.presion\ arterial2, ...$
9: $B.presion\ arterial2Cat, ['mean','var','median','gname'])$
10: $ResultadosPre2 \leftarrow dataset(grupos, ElemenGrpPre2, media, varianza, mediana, porc)$
11: $histogram(B.presion\ arterial2)$

Algorithm 9 Determinación de valores mínimo, medio y máximo de las variables numéricas de la matriz de datos.

1: $load\ B$
2: $quantile(B.Edad, [0, .5, 1])$
3: $quantile(B.imc, [0, .5, 1])$
4: $quantile(B.GLUCOSABASAL, [0, .5, 1])$
5: $quantile(B.COLESTEROLTOTAL, [0, .5, 1])$
6: $quantile(B.TRIGLICERIDOS, [0, .5, 1])$
7: $quantile(B.presion\ arterial, [0, .5, 1])$
8: $quantile(B.presion\ arterial2, [0, .5, 1])$

Algorithm 10 Creación de variables categóricas dicotómicas.

1: $load\ B$
2: $B.EdadCat \leftarrow ordinal(B.Edad, ['0','1'], [16, 35, 77])$
3: $B.imcCat \leftarrow ordinal(B.imc, ['0','1'], [14.8, 25, 53.39])$
4: $B.GLUCOSACat \leftarrow ordinal(B.GLUCOSA, ['0','1'], [47, 100, 545])$
5: $B.COLESTEROLCat \leftarrow ordinal(B.COLESTEROL, ['0','1'], [4, 200, 930])$
6: $B.TRIGLICERIDOSCat \leftarrow ordinal(B.TRIGLICERIDOS, ['0','1'], [5, 150, 2500])$
7: $[5, 150, 2500])$
8: $B.presion\ arterialCat \leftarrow ordinal(B.presion\ arterial, ['0','1'], [30, 140, 195])$
9: $B.presion\ arterial2Cat \leftarrow ordinal(B.presion\ arterial2, ['0','1'], [39, 90, 123])$

Algorithm 11 Creación de variables valoradas $0-1$, para la determinación de factores de riesgo.

```
load B
for i ← 1 : 12363 do
    if B.imcCat(i) ==' 0' then
        B.imcV(i) ← 0
    else
        B.imcV(i) ← 1
    end if
    if B.habitoS(i) ==' 0' then
        B.habSV(i) ← 0
    else
        B.habSV(i) ← 1
    end if
    if B.habitoT(i) ==' 0' then
        B.habTV(i) ← 0
    else
        B.habTV(i) ← 1
    end if
    if B.EdadCat(i) ==' 0' then
        B.EdadV(i) ← 0
    else
        B.EdadV(i) ← 1
    end if
    if B.GLUCOSACat(i) ==' 0' then
        B.GLUV(i) ← 0
    else
        B.GLUV(i) ← 1
    end if
end for
```

Algorithm 12 Creación de variables valoradas $0-1$, para la determinación de factores de riesgo (continuación).

```
load B
for i ← 1 : 12363 do
    if B.COLESTEROLCat(i) ==' 0' then
        B.COLV(i) ← 0
    else
        B.COLV(i) ← 1
    end if
    if B.TRIGLICERIDOSCat(i) ==' 0' then
        TRIV(i) ← 0
    else
        B.TRIV(i) ← 1
    end if
    if B.presion arterialCat(i) ==' 0' then
        B.PRE1V(i) ← 0
    else
        B.PRE1V(i) ← 1
    end if
    if B.presion arterial2Cat(i) ==' 0' then
        B.PRE2V(i) ← 0
    else
        B.PRE2V(i) ← 1
    end if
end for
C ← B(:, 23 : 31)
C ← table2array(C)
B.SUMA ← sum(C, 2)
```

Algorithm 13 División de la matriz total en submatrices de acuerdo al número de factores de riesgo.

$ind \leftarrow B.SUMA == 0$
$B0 \leftarrow B(ind, :)$
$ind \leftarrow B.SUMA == 1$
$B1 \leftarrow B(ind, :)$
$ind \leftarrow B.SUMA == 2$
$B2 \leftarrow B(ind, :)$
$ind \leftarrow B.SUMA == 3$
$B3 \leftarrow B(ind, :)$
$ind \leftarrow B.SUMA == 4$
$B4 \leftarrow B(ind, :)$
$ind \leftarrow B.SUMA == 5$
$B5 \leftarrow B(ind, :)$
$ind \leftarrow B.SUMA == 6$
$B6 \leftarrow B(ind, :)$
$indt \leftarrow B.SUMA == 7$
$B7 \leftarrow B(ind, :)$
$ind \leftarrow B.SUMA == 8$
$B8 \leftarrow B(ind, :)$
$ind \leftarrow B.SUMA == 9$
$B9 \leftarrow B(ind, :)$

Algorithm 14 Formación de submatrices para las actividades económicas con más del 10 % de observaciones.

$indM \leftarrow B.CIIU40\_1 ==' B\ EXPLOTACION\ DE\ MINAS'$
$BMinas \leftarrow B(indM, :)$
$indC \leftarrow B.CIIU40\_1 ==' F\ CONSTRUCCION.'$
$BCons \leftarrow B(indC, :)$
$indM \leftarrow B.CIIU40\_1 ==' K\ ACTIVIDADES\ FINANCIERAS'$
$BFinan \leftarrow B(indF, :)$
$indM \leftarrow B.CIIU40\_1 ==' M\ ACTIVIDADES\ PROFESIONALES'$
$BProf \leftarrow B(indP, :)$
$indM \leftarrow double(indM)$
$indC \leftarrow double(indC)$
$indF \leftarrow double(indF)$
$indP \leftarrow double(indP)$
$A \leftarrow [indM,\ indC,\ indF,\ indP]$
$A1 \leftarrow\ any(A, 2))$
$BXOtros \leftarrow B(A1, :)$
$BXOtros.CIIU40\_1(:, 1) \leftarrow' OTROS'$
$B \leftarrow [BMinas; BFinan; BCons; BProf; BXOtros])$
$B \leftarrow sortrows(B,' ID',' ascend')$

Algorithm 15 División de la matriz de la actividad económica Explotación de Minas en submatrices de acuerdo al número de factores de riesgo.

$ind \leftarrow BMinas.SUMA == 0$
$B0m \leftarrow BMinas(ind, :)$
$ind \leftarrow BMinas.SUMA == 1$
$B1m \leftarrow BMinas(ind, :)$
$ind \leftarrow BMinas.SUMA == 2$
$B2m \leftarrow BMinas(ind, :)$
$ind \leftarrow BMinas.SUMA == 3$
$B3m \leftarrow BMinas(ind, :)$
$ind \leftarrow BMinas.SUMA == 4$
$B4m \leftarrow BMinas(ind, :)$
$ind \leftarrow BMinas.SUMA == 5$
$B5m \leftarrow BMinas(ind, :)$
$ind \leftarrow BMinas.SUMA == 6$
$B6m \leftarrow BMinas(ind, :)$
$indt \leftarrow BMinas.SUMA == 7$
$B7m \leftarrow BMinas(ind, :)$
$ind \leftarrow BMinas.SUMA == 8$
$B8m \leftarrow BMinas(ind, :)$
$ind \leftarrow BMinas.SUMA == 9$
$B9m \leftarrow BMinas(ind, :)$

Algorithm 16 Coeficiente de similaridad de Sokal-Michener.

$functionS = sokal(X)$
$[n, p] \leftarrow size(X)$
$J \leftarrow ones(n, p)$
$a \leftarrow X * X'$
$d \leftarrow (J - X) * (J - X)'$
$S \leftarrow (a + d)/p$

Algorithm 17 Coeficiente de similaridad de Russell-Rao.

$functionRR = RussellRao(X)$
$[n, p] \leftarrow size(X)$
$a \leftarrow X * X'$
$RR \leftarrow a/p$

Algorithm 18 Coeficiente de similaridad de Rogers-Tanimoto.

$functionRT = RogersTanimoto(X)$
$[n, p] \leftarrow size(X)$
$J \leftarrow ones(n, p)$
$a \leftarrow X * X'$
$d \leftarrow (J - X) * (J - X)'$
$bc = p - a - d$
$RT \leftarrow (a + d)/(p + bc)$

Algorithm 19 Coordenadas Principales.

```
load D
function[X, vaps, percent, acum] = coorp(D)
[n, n] ← size(D)
H ← eye(n) − ones(n)/n
B ← −H ∗ D ∗ H/2
L ← eig(B)
m ← min(L)
epsilon ← 1.e − 6
if abs(m) < epsilon then
    D1 ← non2euclid(D)
    B ← −H ∗ D1 ∗ H/2
end if
[T, Lambda, V] ← svd(B)
vaps ← diag(Lambda)
j ← 1
while vaps(j) > epsilon do
    T1 ← T(:, 1 : j)
    X ← T1 ∗ sqrt(Lambda(1 : j, 1 : j))
    j ← min(j + 1, n)
end while
percent ← vaps/sum(vaps) ∗ 100
acum ← zeros(1, n)
for i ← 1 : n do
    acum(i) ← sum(percent(1 : i))
end for
plot(X(:, 1), X(:, 2))
```

Algorithm 20 Transformación de una matriz no euclidea a matriz euclidea.

```
functionD1 ← non2euclid(D)
[n, n] ← size(D)
H ← eye(n) − ones(n)/n
[T, Lambda] ← eig(−H ∗ D ∗ H/2)
m ← min(diag(Lambda))
D1 ← D − 2 ∗ m ∗ ones(n) + 2 ∗ m ∗ eye(n)
```

Algorithm 21 Creación de variables categóricas y valoradas de IMC.

```
load B
[n, p] ← size(B)
B.imcCat ← ordinal(B.imc, ['1Bp','2Nor','3Sp','4Ob1','5Ob2','6Ob ex'], ...
[14.8, 18.5, 25, 30, 35, 40, 54])
for i ← 1 : n do
    if B.imcCat(i) ==' 1Bp' then
        B.imcV(i) ← 1)
    else
        if B.imcCat(i)==’2Nor’ then
            B.imcV(i) ← 2
        else
            if B.imcCat(i)==’3Sp’ then
                B.imcV(i) ← 3
            else
                if B.imcCat(i)==’4Ob1’ then
                    B.imcV(i) ← 4
                else
                    if B.imcCat(i)==’5Ob2’ then
                        B.imcV(i) ← 5
                    else
                        B.imcV(i) ← 6
                    end if
                end if
            end if
        end if
    end if
end for
```

Algorithm 22 Creación de variables categóricas y valoradas de Edad.

```
load B
[n, p] ← size(B)
B.EdadCat ← ordinal(B.Edad, ['1men30','30a40','40a50','50a60','may60'], ...
[16, 30, 40, 50, 60, 77])
for i ← 1 : n do
    if B.EdadCat(i) == '1men30' then
        B.EdadV(i) ← 1)
    else
        if B.EdadCat(i)=='30a40' then
            B.EdadV(i) ← 2
        else
            if B.EdadCat(i)=='40a50' then
                B.EdadV(i) ← 3
            else
                if B.EdadCat(i)=='50a60' then
                    B.EdadV(i) ← 4
                else
                    B.EdadV(i) ← 5
                end if
            end if
        end if
    end if
end for
```

Algorithm 23 Creación de variables categóricas y valoradas de Glucosa.

```
load B
[n, p] ← size(B)
B.GLUCOSACat ← ordinal(B.GLUCOSA, ['1Gb','2Gn','3Ga'], ...
[47, 70, 100, 545])
for i ← 1 : n do
    if B.GLUCOSABASALCat(i) == '1Gb' then
        B.GluV(i) ← 1)
    else
        if B.GLUCOSABASALCat(i)=='2Gn' then
            B.GluV(i) ← 2
        else
            B.GluV(i) ← 3
        end if
    end if
end for
```

Algorithm 24 Creación de variables categóricas y valoradas de Colesterol.

```
load B
[n, p] ← size(B)
B.COLESTEROLCat ← ordinal(B.COLESTEROL, ['1CTn','2CTa'], ...
[4, 200, 890])
for i ← 1 : n do
    if B.COLESTEROLCat(i) ==' 1CTn' then
        B.ColV(i) ← 1)
    else
        B.ColV(i) ← 2
    end if
end for
```

Algorithm 25 Creación de variables categóricas y valoradas de Triglicéridos.

```
load B
[n, p] ← size(B)
B.TRIGLICERIDOSCat ← ordinal(B.TRIGLICERIDOS, ...
['1Tn','2Ta'], [5, 150, 2070])
for i ← 1 : n do
    if B.TRIGLICERIDOSCat(i) ==' 1Tn' then
        B.TriV(i) ← 1)
    else
        B.TriV(i) ← 2
    end if
end for
```

Algorithm 26 Creación de variables categóricas y valoradas de Presión Sistólica.

```
load B
[n, p] ← size(B)
B.presion arterialCat ← ordinal(B.presion arterial, ['1PSb','2PSn', ...
'3PSa]', [60, 100, 140, 183])
for i ← 1 : n do
    if B.presion_arteria Cat(i) ==' 1Gb' then
        B.Pre1V(i) ← 1)
    else
        if B.presion arterialCat(i)=='2Gn'  then
            B.Pre1V(i) ← 2
        else
            B.Pre1V(i) ← 3
        end if
    end if
end for
```

Algorithm 27 Creación de variables categóricas y valoradas de Presión Distólica.

```
load B
[n, p] ← size(B)
B.presion2Cat ← ordinal(B.presion2, ['1PDb','2PDn','3PDna','4PDa'], ...
[39, 60, 90, 95, 123])
for i ← 1 : n do
    if B.presion2Cat(i) ==' 1PDb' then
        B.Pre2V(i) ← 1)
    else
        if B.presion2Cat(i)==’2PDn’ then
            B.Pre2V(i) ← 2
        else
            if B.presion2Cat(i)==’3PDna’ then
                B.Pre2V(i) ← 3
            else
                B.Pre2V(i) ← 4
            end if
        end if
    end if
end for
```

Algorithm 28 Determinación del Índice de Salud.

```
BB ← B(:, 25 : 29)
B.SUMA ← sum(BB, 2)
B.ESTADOSALUD ← ordinal(B.SUMA, ['1bajo','2normal','3alto', ...
'4critico'], [7, 10, 11, 13, 16])
ind ← B.ESTADOSALUD ==' 1bajo
Bbajo ← B(ind, :)
ind ← B.ESTADOSALUD ==' 2normal'
Bnor ← B(ind, :)
ind ← B.ESTADOSALUD ==' 3alto'
Balto ← B(ind, :)
ind ← B.ESTADOSALUD ==' 4critico'
Bcritico ← B(ind, :)
```

Referencias

[Amabili, 2016] Amabili, A. (2016). La era del big data también llegó al sector salud. http://www.ey.com/Publication/vwLUAssets/ey-la-era-del-big-data-tambien-llego-al-serctor-salud/$FILE/ey-la-era-del-big-data-tambien-llego-al-serctor-salud.pdf.

[Baíllo and Grané, 2008] Baíllo, A. and Grané, A. (2008). *100 Problemas Resueltos de Estadística Multivariante (Implementados en Matlab)*. Publicaciones Delta, Madrid.

[Baroni and Buser, 1976] Baroni, C. and Buser, M. (1976). Similarity of binary data. *Systematic Zoology*, pages 251–259.

[Calle, 2011] Calle, J. (2011). Fundación para la diabetes. https://www.fundaciondiabetes.org/general/articulo/68/que-es-el-sindrome-metabolico.

[Cano, 2014] Cano, J. (2014). The v's of big data: velocity, volume, value, variety and veracity. *Rethink Maintenance*.

[Castaño, 2012] Castaño, E. (2012). *Introducción al Análisis de Datos Multivariados en Ciencias Sociales*. Universidad Nacional de Colombia, Medellín.

[Cuadras, 2014] Cuadras, C. (2014). *Nuevos Métodos de Análisis Multivariante*. CMC Editions, Barcelona.

[Curtis et al., 2008] Curtis, H., Barnes, S., Schek, A., and Massarini, A. (2008). *Curtis Biología*. Editorial Médica Panamericana, Madrid.

[Drake, 2009] Drake, F. (2009). El tutorial de python. http://docs.python.org.ar/tutorial/pdfs/TutorialPython2.pdf.

[Fuente, 2011] Fuente, S. (2011). *Análisis Conglomerados*. UAM, Madrid.

[Fuster, 2013] Fuster, V. (2013). Salud integral. *Fundación CEDE / ICLD (International Center for Leadership Development)*, pages 20–25.

[Garside and Cox, 2013] Garside, W. and Cox, B. (2013). *Big Data Storage*. John Wiley and Sons, Chichester.

[Gasteiz, 2002] Gasteiz, V. (2002). *Guía de Práctica Clínica sobre Hipertensión Arterial*. Servicio vasco de salud, Barcelona.

[Goette, 2014] Goette, P. (2014). Julia, un lenguaje del futuro. https://www.genbeta.com/desarrollo/julia-un-lenguaje-del-futuro.

[Gonzáles, 2017] Gonzáles, P. (2017). Cuando el Big Data mande en nuestra salud. https://www.efesalud.com/big-data-salud.

[Grané, 2008] Grané, A. (2008). Distancias estadísticas y escalado multidimensional. http://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/slides_Coorp_reducido.pdf.

[Hurwitz et al., 2013] Hurwitz, J., Nugent, A., Helper, F., and Kaufmaw, M. (2013). *Big Data*. John Wiley and Sons, Inc, Hoboken.

[Jain, 2016] Jain, A. (2016). The five Vs of big data. *Healthcare Data Analitics*.

[Jasim et al., 2015] Jasim, H., Hameed, A., Hadishaheed, S., and Haji, A. (2015). Big Data and five V's characteristics. *International Journals of advances in electronics and computer Science*, pages 16–20.

[Joyanes and Poyatos, 2013] Joyanes, L. and Poyatos, J. (2013). Big Data y el sector de la salud: el futuro de la sanidad. http://poyatosdiaz.com/index.php/big-data-y-el-sector-de-la-/salud-el-futuro-de-la-sanidad.

[Lam, 2016] Lam, J. (2016). Manual MDS, Aterosclerosis. https://www.msdmanuals.com/es-ec/professional/trastornos-cardiovasculares/arteriosclerosis/aterosclerosis.

[MathWorks, 2017] MathWorks (2017). *Statistics and Machine Learning Toolbox User's Guide*. The MathWorks, Inc, Natick Massachusetts.

[MedlinePlus, 2019a] MedlinePlus (2019a). Colesterol. https://medlineplus.gov/spanish/cholesterol.html.

[MedlinePlus, 2019b] MedlinePlus (2019b). Diabetes. https://medlineplus.gov/spanish/diabetes.html.

[MedlinePlus, 2019c] MedlinePlus (2019c). Triglicéridos. https://medlineplus.gov/spanish/triglycerides.html.

[Múnera and Escobar, 2007] Múnera, M. and Escobar, S. (2007). *Carta de Laboratorio Clínico*. Obras de la Congregación Mariana, Medellín.

[OMS, 2018] OMS (2018). Obesidad y sobrepeso. https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight.

[Pateiro, 2016] Pateiro, B. (2016). Introducción a lenguajes avanzados de computación: Matlab en la docencia en Química. http://mathgene.usc.es/matlab-profs-quimica/analisis_datos.pdf.

[Peña, 2002] Peña, D. (2002). *Análisis de Datos Multivariantes*. McGraw-Hill, Madrid.

[Peña, 2008] Peña, D. (2008). *Fundamentos de Estadística*. AlianzaEditorial, Madrid.

[Rotaeche del Campo, 2002] Rotaeche del Campo, R. (2002). Guía de práctica clínica sobre hipertensión arterial. http://www.sld.cu/galerias/pdf/servicios/hta/guia_practica_clinica_sobre_hta_vasca.pdf.

[Saiz, 1980] Saiz, F. (1980). Experiencias en el uso de criterios de similitud en el estudio de comunidades. *Archibo Biologico Medico*, pages 387–402.

[Trigo and López, 2012] Trigo, M. and López, J. (2012). Amor a primera vista tableau public. http://blogs.lanacion.com.ar/data/herramientas/amor-a-primera-vista/-tableau-public/.

[Urueña et al., 2012] Urueña, A., Ballestero, M., and Prieto, M. (2012). *Big Data en salud digital informe de resultados*. Fundación Vodafone España, Madrid.

[Zaforas, 2016] Zaforas, M. (2016). ¿Qué puede aportar el Big Data al mundo de la medicina? https://www.paradigmadigital.com/dev/puede-aportar-big-data-al-mundo/-la-medicina/.