UNIVERSIDAD CATÓLICA SANTO TORIBIO DE MOGROVEJO

FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

DESARROLLO DE UNA SOLUCIÓN DE MINERÍA DE DATOS PARA LA DETERMINACIÓN DE SEGMENTOS DE CLIENTES EN UNA EMPRESA DE CAPACITACIONES ONLINE CHICLAYO

TESIS PARA OPTAR EL TÍTULO DE

INGENIERO DE SISTEMAS Y COMPUTACIÓN

AUTORA

DANY YESENIA GELACIO MENDOZA

ASESOR

Ing. SEGUNDO JOSÉ CASTILLO ZUMARÁN

Chiclayo, 2019

DEDICATORIA

A Dios quien siempre me acompaña y todos los días me ofrece un nuevo comienzo después de cada tropiezo o dificultad que se me presenta.

A mis padres Miguel Gelacio y Alejandrina Mendoza, por ayudarme a ser una mejor persona inculcándome valores, brindándome sus consejos y sus palabras de aliento cuando más lo necesité en especial a mi madre, quien es mi inspiración para seguir adelante con este proyecto.

.

AGRADECIMIENTOS

A mi familia por haberme apoyado en el transcurso de esta etapa académica.

A mis amigos(as), por las recomendaciones, ayuda y consejos en esta etapa académica, diciendo “Tú puedes, ya acabas… te falta poco”.

A la profesora Vanessa Li Vega, las gracias infinitas por su apoyo incondicional, disposición de su tiempo y paciencia en resolver mis dudas sobre un nuevo conocimiento.

A mis asesores como especialistas y metodológicos; por su paciencia, apoyo y dedicación para ejecutar este proyecto, alentándome en el proceso de desarrollo de esta tesis, diciendo: “Tú puedes flaca”, “Vas a querer llevar tesis el próximo ciclo”, “Animo Gelacio, ya te falta poco”.

A la empresa, por brindarme el apoyo para ejecutar este proyecto en beneficio de su desarrollo.

RESUMEN

La minería de datos en el aspecto de segmentación permite descubrir ciertos aspectos relevantes de sus clientes, como su patrón de consumo, permitiendo a la empresa crear estrategias comerciales para el beneficio tanto del cliente como de la empresa misma. Se aplicó la metodología CRISP-DM para el proceso de minería de datos, como modelo empleado fue descriptivo para este tipo de solución, en la cual; se utilizaron técnicas de clustering de minería de datos, como algoritmos K-means, distancia, K-medoids. Las herramientas utilizadas fueron: Rstudio para efectuar el agrupamiento de datos, Power BI para mostrar los resultados de manera interactiva al usuario final y un dashboard para presentación de los resultados. Teniendo como resultados el patrón de clientes de consumo por su monto promedio de consumo, la cantidad de cursos y promedio de consumo, departamentos con mayor impacto, edad promedio de cada grupo(clúster). Y finalmente, como conclusiones se obtuvo que los datos almacenados en el periodo 2014-2018 permitió un análisis exploratorio, seleccionando variables cuantitativas y cualitativas creando así un nuevo modelo a evaluar por algoritmos de minería de datos, obteniéndose 3 grupos(clústeres) formados por homogeneidad.

PALABRAS CLAVE: segmentación, estrategia, patrón de cliente, CRISP-DM, minería de datos.

ABSTRACT

The mining of data in the aspect of segmentation allows discovering certain relevant aspects of customers such as its consumption pattern allowing the company to create commercial strategies for the benefit both the client and the company itself. The CRISP-DM methodology was applied for the data mining process, as the model used is descriptive for this type of solution, in which we used data mining clustering techniques, such as K-means, distance, K-medoids algorithms. The tools used are: Rstudio to perform the data grouping, Power BI to show the results interactively to the end user and a dashboard for presentation of the results. Obtaining as a result the pattern of consumer clients by its average amount of consumption, the number of courses and average consumption, departments with the greatest impact, average age of each group (cluster). And finally, the conclusions that we have obtained are that the data stored in the period 2014-2018 allowed an exploratory analysis, selecting quantitative and qualitative variables, thus creating a new model to be evaluated by data mining algorithms, obtaining 3 groups (clusters) formed by homogeneity.

KEYWORDS: segmentation, strategy, client pattern, CRISP-DM, data mining.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

I. INTRODUCCIÓN

Las empresas actualmente generan información valiosa que muchas veces no es aprovechada de manera oportuna, a pesar de ser la clave para todo tipo de negocio. La minería de datos según El Cronista*, […], es una alternativa que permite analizar los datos en base a patrones, los cuales permiten predecir comportamientos, hábitos de compras o, fuga de clientes y fraude, entre otras acciones […], es aquí donde entra en escena las soluciones de data mining o minería de datos. Por su parte, Daniel Yankelevich, CEO de la consultora Pragma nos indica que antes de empezar a minar es importante entender el valor para el negocio de esos datos y definir una estrategia para alinear el proyecto tecnológico. La ejecutiva de CMR Falabella cuenta con más de 10 años utilizando data mining permitiendo la automatización de sus procesos, en las cuales; identifica fraudes, riesgos y preferencia de los clientes. [1]*

El grupo viajero 2.0 de España; señala que *el problema actual de las organizaciones no es cómo obtener, ordenar y almacenar la información, sino cómo convertir en conocimiento tan numerosa y variada información. Y es precisamente el Data Mining (minería de datos), a través de sus herramientas inteligentes, el que extrae la información significativa de grandes bases de datos, detectando tendencias y correlaciones para permitir al usuario realizar predicciones que resuelven problemas del negocio, proporcionando ventajas competitivas.* [2]

En el Perú las empresas que están implementando aquellas tecnologías de análisis son más vinculadas al campo digital. El problema que presenta no es la implementación sino el desconocimiento en la utilización de dichas herramientas en el uso de la Data Mining que se genera. Luis Arellano director de Data Mining in the cloud en IBM menciona que el Data Mining; "todavía puede tener un impacto importante en el país. El único límite que existiría es la insuficiencia de datos para hacer un análisis profundo, lo que significa que solamente las empresas de cierto tamaño podrán obtener mejores resultados" [3]. El objetivo de la empresa es tener un área potente de innovación y establecer estrategias claras en las necesidades de los clientes aplicando correctamente las herramientas de minería de datos.

Según Ricardo Arce, en el diario La Gestión señala que: *Southern Perú maneja de manera muy precisa sus rubros tanto logísticos, los insumos que compran*

como la parte de mantenimiento (…) Una de las mayores palancas de valor dentro de una minera es el mantenimiento de sus equipos, anticipar cuando una máquina va a fallar y qué tipo de insumos pueden concentrarse para disminuir costos. Estas empresas para tener mejores tomas de decisiones están a la vanguardia de la tecnología que facilita y obtiene mayores ganancias. [4]

La empresa donde se realizó el proyecto se dedica a la prestación de servicios en capacitación online. La cual, se dedica a proporcionar cursos de capacitación en base a la demanda y actualización del mercado de manera virtual, y presencial a personas con estudios culminados, así mismo a estudiantes. Una de sus fortalezas son las sesiones en vivo y en tiempo real que ahí se brindan; en las cuales el estudiante puede hacer preguntas al exponente y de igual manera éstas son resueltas en su momento. Asimismo, cuentan con convenios con distintos colegios del Perú permitiéndoles el respaldo de sus certificados emitidos.

La presente tesis denominada "Desarrollo de una solución de minería de datos para la determinación de segmentos de clientes en una empresa de capacitaciones online Chiclayo", tuvo como finalidad la determinación de segmentos de clientes con el uso de minería de datos. Al inicio, la empresa, frente al lanzamiento de cierto curso (evento) enviaba correos, publicaciones en redes sociales y mensajes de Whatsapp masivamente sin contar con un grupo identificado, generando más de un 40% de correos que no eran recibidos o leídos y provocando molestias al cliente por SPAM. La entrevista realizada al gerente mostró una aceptación favorable entre el 50% de matrículas aceptadas por los medios de publicidad lanzada por la empresa, en algunas ocasiones no se llegaban a matricular la cantidad suficiente de clientes a un cierto curso lanzado, ocasionando pérdidas monetarias en muchas ocasiones. Ante esta realidad, fue importante formular la siguiente pregunta ¿Cuáles son las características de los segmentos de clientes en una empresa de capacitaciones online Chiclayo? Frente a esta pregunta y la necesidad de profundizar el problema, se realizó la investigación cuya población fue de 3463 clientes matriculados. Para ello, se determinó como objetivo general: Determinar las características de los segmentos de los clientes en la empresa de capacitaciones online Chiclayo, y como objetivos específicos 1. Realizar un análisis exploratorio a la base de datos; 2. Identificar los atributos que definan las características del cliente; 3. Identificar algoritmos para segmentar clientes de

acuerdo con sus características; 4. Analizar los clientes de acuerdo a la segmentación y 5. Sugerir patrones en base a la solución de minería de datos.

La presente tesis se dividió en nueve (9) capítulos: I) Introducción, II) Marco teórico, III) Metodología, IV) Resultados V) Discusión, VI) Conclusiones, VII) Recomendaciones, VIII) Lista de referencias y IX) Anexos.

En el capítulo I se habla sobre la actualidad a nivel mundial, nacional y la problemática de la empresa, el capítulo II hace referencia a los antecedentes y bases teóricas utilizadas, en el capítulo III se habla acerca de la metodología usada como CRISP-DM, en el capítulo IV se muestran los resultados obtenidos de la investigación. En el capítulo V se habla sobre la discusión de los resultados, en el capítulo VI se habla de las conclusiones como resultado de la investigación, en el capítulo VII se establecen las recomendaciones a la empresa, en el capítulo VIII se lista las referencias utilizadas y en el capítulo IX se presentan los anexos.

II. MARCO TEÓRICO

2.1. Antecedentes

Se han considerado para esta investigación los siguientes antecedentes:

2.1.1. Antecedentes internacionales

Riquelme [5], narra como problemática que la empresa Kaufmann S.A Vehículos no tiene identificado sus clientes activos, puesto que estos son importantes para el crecimiento de la empresa. Se aplicó como método jerárquico y técnica Kmeans, logrando obtener estrategias a partir de un análisis de clientes históricos. El valor agregado a esta investigación es la utilización de clustering para creación de estrategias de negocio. Finalmente, el autor propone estrategias de fidelización de clientes para la toma de decisiones en cuanto a sus clientes activos. Se tomó en consideración esta tesis ya que se trabaja con modelos descriptivos enfocados en clientes potenciales.

Orellana [6], narra la problemática presentada en Chilectra sobre el consumo asociado a sus clientes en energía, la cual no cuenta con un perfil de consumo asociados a los equipos, generando problemas en la medición de consumo de sus clientes. Se aplicó la metodología KDD, con técnicas de clustering, logrando identificar un análisis económico y los riesgos. El valor agregado a esta investigación es el análisis que se realiza históricamente en el consumo de energía asociados a los equipos de medición. Finalmente, el autor concluye con 5 segmentos de clientes asociados a sus consumos mostrando el nivel de crecimiento en el servicio de Chilectra.

Chamba [7], la autora presenta como problemática el área de marketing y ventas por la identificación de clientes potenciales y además porque no cuentan con estrategias para llegar a ellos en la empresa Master PC. Se aplicó como metodología CRISP-DM, logrando una segmentación de clientes con un modelo de RFM. El valor agregado a esta investigación es que mediante su recurrencia se puede identificar clientes potenciales creando estrategias de ventas de ciertos productos. Finalmente, la autora concluye que mediante los grupos de clientes encontrados la empresa puede

determinar promociones en base a sus similitudes de compra. La razón por la que se consideró este trabajo de investigación es por los grupos formados mediante el consumo.

2.1.2. Antecedentes nacionales

Grández [8], narra la problemática en una empresa que se dedica a la venta de suplementos nutricionales, que, al no contar con sus preferencias de consumo de los clientes, le resulta complicado realizar la tarea de ofrecer productos con ciertas promociones de acuerdo a su histórico de consumo. Se aplicó la metodología CRISP-DM, logrando obtener grupos de clientes mediante el algoritmo de asociación para la problemática de segmentación de clientes por consumo. El valor agregado de esta investigación es la aplicación de herramientas como Microsoft Sql Server Integration Services y el algoritmo de asociación. Finalmente, el autor concluye con los productos mayor demandados por los clientes. La razón por la que se consideró esta tesis es porque aplican algoritmos de asociación de consumo y producto.

Ramírez [9], narra como problemática la identificación de perfiles de clientes crediticios que realizaron transacciones en un determinado trimestre. Se aplicó el método segmentación K-means, logrando identificar perfiles de clientes de medio consumo con un pago medio. El valor agregado de esta investigación es el uso de la técnica k-means con una tasa del 94% de proximidad. Finalmente, el autor concluye con recomendaciones que permiten realizar evaluaciones en otros periodos continuamente para la identificación de nuevos grupos. La razón por la que se tomó en cuenta es por la consideración de variables cuantitativas y cualitativas.

Calderón y Vega [10], narra como problemática la insatisfacción de clientes en un supermercado por la visualización de productos, lo que demanda tiempo para que éstos encuentren un cierto producto. Se aplicó la metodología RUP, logrando identificar la ubicación de los productos. Finalmente, los autores concluyen con la implementación de una plataforma para la ubicación de los productos en base a las necesidades de los clientes. La razón por la

que se consideró está tesis es el análisis de acuerdo con las asociaciones encontradas por el consumo de los clientes.

2.1.3. Antecedentes locales

Cotrina y Gonzáles [11], narran como problemática el grado de satisfacción de los clientes en la empresa Supermercados El Super, el cual cuenta con un data amplia, sin embargo, desconocen las nuevas estrategias que se forman a partir de datos del cliente para la mejora de promociones de sus productos. Se aplicó la metodología CRISP-DM, logrando obtener un análisis de ventas para la generación de promociones. El valor agregado de esta investigación es la aplicación de reglas de asociación con el fin de encontrar relaciones dentro de las transacciones de ventas. Finalmente, los autores concluyeron con las herramientas de minería de datos y usando la técnica de asociación muestran información relevante a los ejecutivos para la creación de estrategias que este tome.

Jiménez [12], narra como problemática la dificultad de evidenciar los problemas delictivos, lo cuales; dificultan la demora en el proceso de analizar los perfiles de las personas denunciadas. Se aplicó la metodología CRISP-DM, logrando identificar patrones delictivos. El valor agregado a esta investigación es el sistema de alerta basada en minería de datos. Finalmente, la autora concluye con 12 clústeres identificados, logrando identificar los perfiles de personas con antecedentes delictivos, mejorando la eficiencia en la toma de decisiones. La razón por la que se consideró esta tesis es por el uso de variables mixtas para determinar perfiles.

Cubas [13], narra la problemática sobre el proceso de gestión académica del instituto ISAG, para la captación de nuevos alumnos. Se aplicó la metodología CRISP-DM, logrando obtener estrategias de captación de alumnos de acuerdo a su perfil histórico. Finalmente, el autor concluye que mediante su información histórica del alumno y aplicando herramientas de minería de datos se logra obtener una información de perfiles. La razón, por la que se consideró esta tesis es por la aplicación de clústeres con variables mixtas.

2.2. Bases teórico científicas

2.2.1. Segmentación de mercado

Es una herramienta de mercadotecnia que permite realizar un análisis del mercado de forma efectiva, en la división de grupos heterogéneos con características homogéneas. [14]

La segmentación de mercado, según León "es el proceso de dividir un mercado en subconjuntos de consumidores con necesidades o características comunes". [15]

El autor Arellano define la segmentación de mercado, como el proceso de mercado con el fin de identificar grupos de consumidores con características comunes según sus necesidades. Además, que la segmentación es un proceso de permanente cambio. Lo que permite un aprovechamiento de la empresa y a la vez la satisfacción de los clientes. [16]

"Los segmentos se crean en función de las características de los consumidores y no de función de los productos que los satisfacen". [16]

2.2.1.1 Variables utilizadas para la segmentación

Las variables más comunes utilizadas para la segmentación en el mercado son:

2.2.1.1.1 Segmentación demográfica

Son una de las variables mayor utilizadas por lo que brindan datos numéricos. Entre las variables demográficas tenemos las siguientes: [16]

- ✓ Sexo
- ✓ Edad
- ✓ Raza
- ✓ Lugar de residencia
- ✓ Características físicas

2.2.1.1.2 Segmentación Socioeconómica

Es una variable importante, ya que nos permite identificar el poder adquisitivo de nuestros futuros consumidores. [14]

Según el autor Rolando, los factores socioeconómicos más importantes son: [16]

- ✓ Nivel de ingreso
- ✓ Nivel de educación
- ✓ Profesión
- ✓ Clase social

2.2.1.1.3 Segmentación Psicográfica

Se hace referencia a las características psicológicas de los consumidores, de acuerdo a su personalidad, lo que permite segmentar a grupos de acuerdo a su sensibilidad de ciertos productos. Las más conocidas, según el autor Arellano son: [16]

- ✓ Nivel de extroversión
- ✓ Grado de innovación
- ✓ Características generales

2.2.1.1.4 Segmentación por tipo de uso

Corresponde a los consumidores que utilizan un determinado tipo de bienes. Las categorías son las siguientes: [16]

- ✓ Por la cantidad de uso
- ✓ Por el tipo de uso
- ✓ Por oportunidad de uso
- ✓ Por lealtad a la marca

2.2.1.1.5 Segmentación por estilos de vida

La segmentación por estilos de vida engloba ciertas características como demográfica, socioeconómico, psicográfica y por su tipo de uso. Lo que permite tener una segmentación multivariable, ya que sus datos son estadísticos. [16]

2.2.2. Minería de datos

Se define la minería de datos, como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. [14]
Según Sinnexus; “es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto”. [15]
Por lo tanto, minería de datos es proceso de información que utiliza el análisis matemático para deducir patrones y tendencias de datos. Que tiene como objetivo analizar los datos para extraer conocimiento.

2.2.3. Tipos de modelos

La minería de datos tiene como conocimiento los datos de la empresa que puede ser tanto de forma relacional, como patrones o reglas de datos desconocidos. Modelos que pueden ser de dos tipos: predictivos y descriptivos. [17]

2.2.3.1. Modelos Predictivos

[17]Estos modelos pretenden estimar valores futuros o desconocidos de variables objetivos para el estudio. Lo que comprende de variables objetivos o dependientes y variables independientes o predictivas de la base de datos.
Algunas técnicas predictivas son:

- Clasificación
- Regresión
- Predicción

2.2.3.2. Modelos Descriptivos

Identifican patrones que explican las propiedades de los datos examinados, no para predecir nuevos datos. Se basan en variables dependientes u objetivos (target). Algunas técnicas descriptivas son: [17]

- Clustering

- Asociaciones
- Dependencias

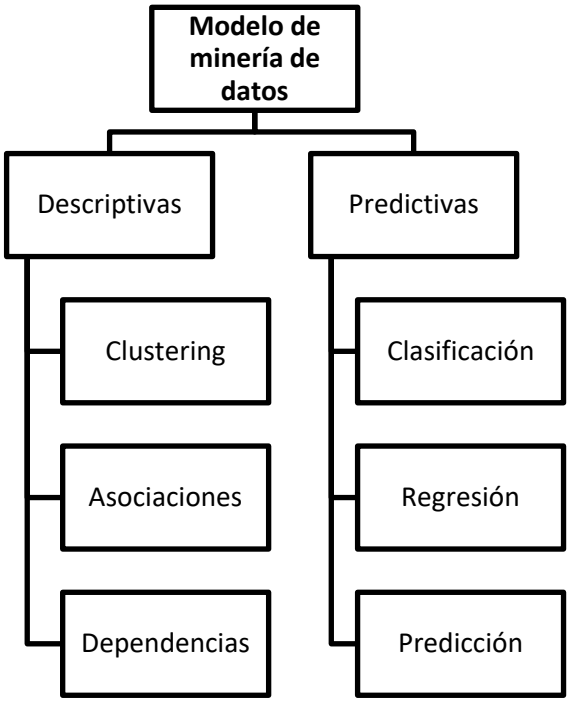


Fig. 1. Modelo de minería de datos

2.2.4. Clustering

Clustering también llamado agrupamiento o segmentación, es una de las técnicas del modelo descriptivo de minería de datos. Que comprende el proceso de división de datos en grupos con similitudes. Se representa la información obtenida mediante clúster. [17]

2.2.4.1. Clustering particional

Son un conjunto de técnicas cuyo objetivo es la división de datos en grupos y formando subgrupos con intersección vacía. La asignación de objetos se agrupa por centros llamado centroide, en la cual divide a estos grupos por homogeneidad. Los algoritmos más representados son k-means y k-medoids (PAM) [17]

2.2.4.1.1. K-means

Algoritmo de clasificación no supervisada K-Means (traducido al español es K-medias), frecuentemente utilizado por clustering; agrupa objetos en k grupos basándose en sus características. Este algoritmo resuelve el problema de optimización en la cual, hace uso del centroide

de cada clúster, se calcula con la media aritmética de los valores por sus atributos.

El algoritmo K-means consta de los siguientes pasos: [18]

- Inicialización; se escoge el número de grupos *k*, que se establecen como *k* centroides.
- Asignación de objetos a los centroides; cada objeto de los datos es asignado un centroide más cercano.
- Actualización de centroides; se actualiza la posición del centroide de cada grupo, obteniéndose una nueva ubicación del promedio de los objetos de cada grupo.

2.2.4.1.2. K-medoids

Algoritmo K-medoids o PAM es una técnica de partición de agrupación en clústeres, en la cual; agrupa datos de n objetos en k clústeres. Donde k, es el número de clúster asignado.

El algoritmo k-medoids a diferencia de k-means, este no se representa por un punto artificial sino por un objeto del propio clúster denominado medoid que minimiza las distancias entre objetos del clúster. [17]

2.2.4.3. Clustering Jerárquico

Clustering jerárquico tienen como objetivo agrupar nuevos objetos en grupos homogéneos posibles y más heterogéneos entre los grupos en la que se representa gráficamente un árbol llamado dendograma. Entre ellos encontramos: aglomerativos, se basan en formar grupos de manera ascendente; y disociativos, también llamados descendentes se conforma buscando grupos hasta llegar a un objeto construyéndose desde la raíz hasta las hojas. [19]

2.2.4.4. Clustering Basados en densidad

Este clustering basado en la densidad, se basa en la densidad de un objeto; definiéndose como número de vecinos alcanzables determinados por un radio. [17].

2.2.5. Lenguaje de programación Data Science

La ciencia de datos (Data Science), es un campo de big data que consiste en extraer conocimiento a partir de cantidades relevantes de datos para responder las preguntas formuladas. [20]

2.2.5.1 Python

Según Alvarez, define a Python como lenguaje interpretado, "…lenguaje de scripting independiente de plataforma y orientado a objetos, preparado para realizar cualquier tipo de programa, desde aplicaciones Windows a servidores de red o incluso, páginas web." [21]

2.2.5.2 Lenguaje R

R es un entorno integrado de programas para la manipulación de datos con un enfoque al análisis estadístico la cuales, contiene técnicas estadísticas permitiendo generar nuevos descubrimientos. [22]

TABLA I
Características de lenguaje en data science

Características	
Python	Lenguaje R
• Multiplataforma • Versátil • Librerías escasas • Facilidad de uso • Código abierto • < repositorios • Ágil	• Lenguaje robusto • Funciones integradas • Librerías en Cran • Visualización de datos • Código abierto • > repositorios

2.2.6. Metodologías de extracción de datos

Las metodologías orientadas a la extracción de datos son aplicadas sobre una gran cantidad de información y documentación acumulada durante varios años de experiencia. [23]

Existen varios tipos de estas metodologías, dentro de las cuales se describirán tres de las principales.

2.2.6.1. Metodología KDD (Knowledge Discovery in Databases)

Es un proceso interactivo e interactivo. KDD; es el proceso de usar métodos de minería de datos para extraer conocimiento utilizando una base de datos. Este se organiza en cinco fases: [24]

1. Integración y recopilación

Es la etapa de organizar el conjunto de datos, para descubrir las necesidades de la organización.

2. Selección, limpieza y transformación

La calidad del conocimiento de los datos minados. Es el proceso donde se selecciona y se prepara el subconjunto de datos que se va a minar.

3. Minería de datos

Fase más característica del KDD; tiene como objetivo nuevo conocimiento para el usuario.

4. Evaluación e interpretación

Son los datos de calidad de patrones descubiertos por algún algoritmo de minería de datos; a esto se le añade ciertas cualidades de precisión, comprensibles e interesantes.

5. Difusión y uso

Esta fase consiste en la difusión y distribución a posibles usuarios para su nuevo conocimiento que se da en la toma de decisiones.

2.2.6.2. Metodología SEMMA

SEMMA es el acrónimo a las cinco fases: (Sample, Explore, Modify, Model, Assess) La metodología es propuesta por SAS Institute Inc, y la define como: "… proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones de negocios desconocidos…". [25]

Fases de la metodología SEMMA:

Fase 1 Sample o Muestra

Entrada de datos. Ejemplo, partición de datos

Fase 2 Explore o Explorar

Exploración distribuida, múltiples particiones, intuición, asociación, selección de variables

Fase 3 Modify o Modificar

Transformación de variables, filtros a los datos fuera de rango, agrupamiento.

Fase 4 Model o Modelado

Regresión, árboles, redes neuronales

Fase 5 Assess o Evaluar

Evaluación, medidas, reportes

2.2.5.3. Metodología CRISP-DM

La metodología CRISP-DM tiene como estructura un ciclo de vida de "un proyecto de Data Mining en seis fases, que interactúan entre ellas de forma interactiva durante el desarrollo del proyecto". [26]

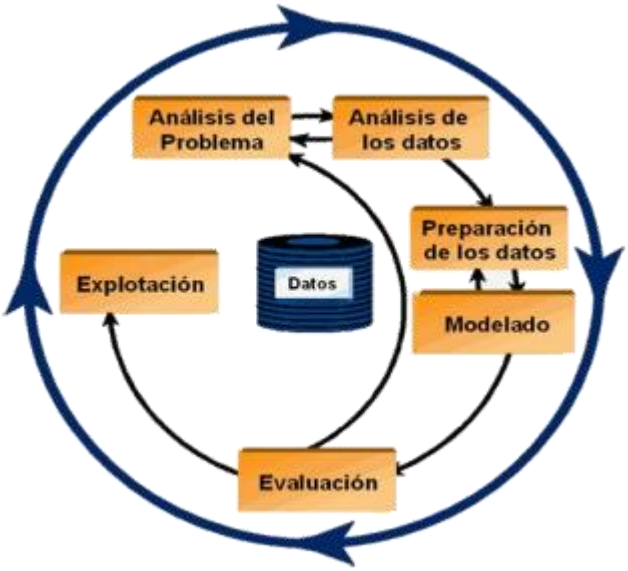


Fig. 2. Metodología CRISP-DM. [27]

Las fases de CRISP-DM:

Fase 1: Entendimiento del negocio o análisis del problema

Esta fase comprende el estudio de la problemática y establece objetivos del proyecto lo cual, se necesita para su desarrollo. [27]

Tareas:

1. Determinar los objetivos del negocio.

2. Valoración de la situación (realidad problemática de la empresa).

3. Determinación de los objetivos

Fase 2: Entendimiento de los datos o análisis de los datos

Esta fase se basa en la comprensión de la primera base del negocio donde se realiza la identificación de problemas de calidad de estos. [27]

Tareas:

1. Recopilar los datos iniciales.

2. Descripción de los datos.

3. Analizar los datos

4. Verificar la calidad de datos.

Fase 3: Preparación de los datos

En esta fase se va a construir los datos finales a partir de los datos iniciales. [27]

Tareas:

1. Selección de los datos

2. Limpieza de los datos: Se busca elevar la calidad de los datos al nivel requerido por las técnicas de BI.

3. Construcción de datos

4. Integración de datos

5. Aplicación de formatos a los datos.

Fase 4: Modelado

En esta fase se usan varias técnicas de modelamiento que son seleccionadas y aplicadas, buscando los valores óptimos de acuerdo a técnicas seleccionadas. [27]

Tareas:

1. Seleccionar la técnica de modelamiento o algoritmo

2. Construcción del modelo de pruebas

3. Implementación del modelo

4. Evaluación del modelo construido

Fase 5: Evaluación

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema en función a los criterios establecidos. [27]

Tareas:

1. Evaluación de los resultados

2. Revisión del proceso

3. Determinar los próximos pasos

Fase 6 Distribución

La distribución consiste en utilizar el conocimiento obtenido de la fase evaluación [27]

Tareas:

1. Planificación de distribución.

2. Planificación del control y del mantenimiento.

3. Creación de un informe final.

4. Revisión del Proyecto

TABLA II

COMPARACIÓN DE METODOLOGÍAS DE MINERÍA DE DATOS

Fases	KDD	SEMMA	CRISP-DM
Comprensión del negocio	Integración y recopilación	No cuenta con compresión del negocio.	Comprensión del negocio
Comprensión de minería de datos	Selección, limpieza y transformación	Muestreo	Comprensión de datos
		Exploración	
		Manipulación	Preparación de datos
	Minería de datos	Modelado	Modelamiento
	Evaluación e interpretación	Valoración	Evaluación
	Difusión y uso	No cuenta con difusión	Despliegue

III. METODOLOGÍA

La metodología usada es la investigación científica, de proceso cuantitativo. [28]

3.1. Tipo y nivel de investigación

3.1.1. Tipo de investigación

En esta investigación se realizó de tipo exploratorio, ya que busca crear un nuevo conocimiento para la empresa de capacitaciones online.

3.1.2. Nivel de investigación

Nivel de investigación cuasi experimental y exploratorio.

3.2. Diseño de investigación

De acuerdo a la investigación que se desarrolló y según los estudios el diseño es pre-test y post-test de grupo único. [28]

G: O1 X O2

Donde:

O1 = Clientes de la empresa

O2 = Clientes de la empresa después de la solución

3.3. Población, muestra y muestreo

3.3.1. Población

La población objeto del estudio estuvo constituida por el registro de clientes matriculados con un total de 3463 en los periodos 2014-2018 en la empresa capacitaciones online.

3.3.2. Muestra

La muestra se ha obtenido con la cantidad de clientes matriculados con un total de 2482. No se hace uso de la formula para obtención de muestra de población finita debido a que no es conveniente para minería de datos. Esto a razón que se necesita un volumen de datos mayor lo que genera mayor precisión a la solución aplicada. [17]

3.3.3. Muestreo

La técnica de muestreo que se aplicó ha sido por conveniencia.

3.4. Criterios de selección

La población se delimita a los clientes matriculados que se comprenden en el periodo 2014-2018 y que son los tipos de participantes que se evaluarán.

3.5. Operacionalización de variables

Las variables que se han utilizado como elementos básicos en el desarrollo de la hipótesis están identificadas de la siguiente manera:

3.5.1. Variables

Las variables que se han utilizado como elementos básicos en el desarrollo de la hipótesis están identificadas de la siguiente manera:

3.5.1.1. Variable independiente

Desarrollo de una solución de minería de datos.

3.5.1.2. Variable dependiente

Clientes de una empresa de capacitaciones online.

3.5.2. Indicadores (Operacionalización de variables)

TABLA III
INDICADORES

Objetivo específico	Indicador(es)	Definición conceptual	Unidad de medida	Instrumento	Definición operacional
Realizar un análisis exploratorio a la base de datos.	Número de datos disponibles.	Número de datos respecto a los atributos a analizar.	Número	Consultas SQL query	La cantidad de datos de cada atributo que dispone.
Identificar los atributos que definan las características del cliente.	Número de atributos identificados.	Número de atributos identificados para el modelado.	Número de atributos cuantitativas y cualitativas	Por segmentación de clientes	Número de atributos identificados disponibles.
Identificar algoritmos para segmentar clientes de acuerdo con sus características.	Número de algoritmos de clustering.	Número de algoritmos utilizadas para la segmentación.	Número	Algoritmos de clustering a evaluar	El número de algoritmos de clustering a evaluar
Analizar los clientes de acuerdo con la segmentación.	Número de grupos formados por la solución.	Número de grupos hallados a partir de los algoritmos de segmentación.	Número	Resultado de las técnicas	Resultados obtenidos por las técnicas de minería de datos.
Sugerir patrones en base a la solución de minería de datos	Número de características de los segmentos de clientes.	Recomendaciones sugeridas a partir de la solución.	Número	Estrategias de marketing	Número de características para el modelo de la solución.

3.6. Técnicas e instrumentos de recolección de datos

A continuación, en la siguiente tabla se muestra las técnicas e instrumentos que fueron útiles para la recolección de datos.

TABLA IV
TÉCNICAS E INSTRUMENTOS DE RECOLECCIÓN DE DATOS

Técnicas	Instrumentos	Elementos de la población	Propósito
Entrevista	Cuestionario de preguntas	Gerente General	Conocer los clientes existentes y el proceso
Observación	Ficha de Observación	Base de datos y registro de clientes	Proceso de matrícula de clientes a cursos (eventos)

3.7. Procedimientos

3.7.1. Metodología de desarrollo

La metodología utilizada para llevar a cabo la investigación de minería de datos fue CRISP-DM (Cross-Industry Standard Process for Data Mining). [26]

A continuación, se mencionan las actividades que se realizaron en cada una de las iteraciones de la metodología a seguir, en este caso CRISP-DM:

1. **Iteración #1: Comprensión del negocio**

   En esta iteración se desarrollaron las siguientes actividades:

   - ✓ Problemática.
   - ✓ Objetivos del negocio.
   - ✓ Describir la solución actual.
   - ✓ Inventario de recursos.
   - ✓ Identificación de origen de datos y almacén de conocimientos.
   - ✓ Riesgos y contingencias.
   - ✓ Datos.
   - ✓ Terminología.
   - ✓ Análisis de costos / beneficios.
   - ✓ Objetivos de minería de datos.
   - ✓ Plan de proyecto (cronograma).

2. **Iteración #2: Comprensión de datos**

   En esta iteración se desarrollaron las siguientes actividades:

   - ✓ Recopilación de datos iniciales.
   - ✓ Descripción de datos.
   - ✓ Tipo de valores.
   - ✓ Exploración de datos.
   - ✓ Verificación de calidad de datos.

3. **Iteración #3: Preparación de los datos**

   En esta iteración se desarrollaron las siguientes actividades:

   - ✓ Selección de datos.
   - ✓ Limpieza de datos.
   - ✓ Construcción de nuevos datos.
   - ✓ Integración de datos.
   - ✓ Formato de datos.

4. **Iteración #4: Modelado**

   En esta iteración se desarrollaron las siguientes actividades:

   - ✓ Selección de técnica de modelado.
   - ✓ Generación de un diseño de comprobación.
   - ✓ Generación de los modelos.
   - ✓ Evaluación del modelo.

5. **Iteración #5: Evaluación**

   En esta iteración se desarrollaron las siguientes actividades:

   - ✓ Evaluación de los resultados.
   - ✓ Proceso de revisión.
   - ✓ Determinación de los pasos siguientes.

6. **Iteración #6: Distribución**

   En esta iteración se desarrollaron las siguientes actividades:

   - ✓ Planificación de distribución.
   - ✓ Planificación del control y del mantenimiento.
   - ✓ Creación de un informe final.
   - ✓ Revisión final del proyecto.

3.7.2. Producto acreditable

1. Interfaces

Se construyeron las interfaces para mostrar los resultados haciendo uso estilos, html y js las mismas que se presentan en el ***ítem 4.1.6. Iteración #6: Distribución, en el capítulo IV. Resultados.***

2. Arquitectura

Se diseñó una arquitectura idónea para el funcionamiento de la solución para la segmentación de clientes, el cual se utiliza cliente – servidor.

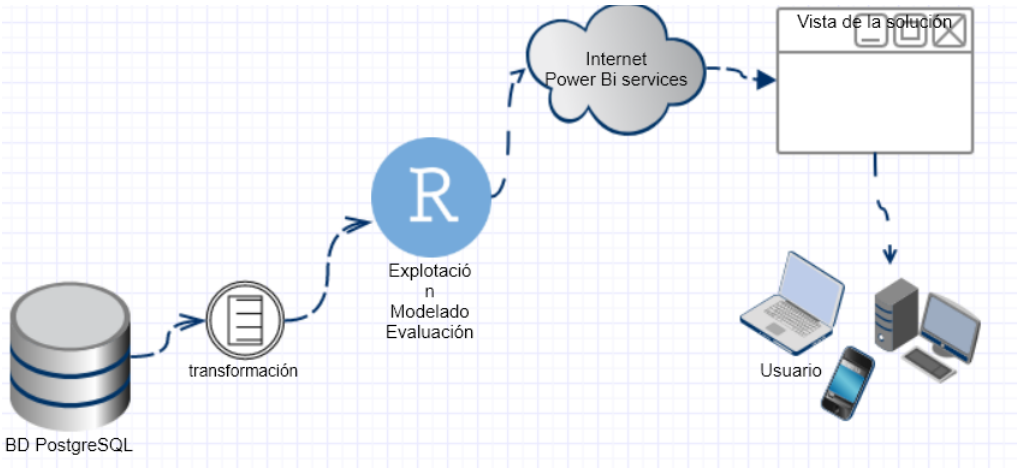


Fig. 3. DISEÑO DE ARQUITECTURA

3. Infraestructura tecnológica

Se definen las características de cada uno de sus componentes en base a recursos detallándose a continuación:

TABLA V.
CARACTERÍSTICAS DE RECURSOS

Recursos	Características
Laptop ó máquina de escritorio	Características mínimas: • Ram 2 Gb • Memoria 120 GB • Procesador Intel 3 • Disco duro 120 GB • Sistema operativo Windows 7 mínimo
Tablet	Requisitos mínimos: • Wifi • Pantalla 7 pulgadas • Resolución 1024 x 768 píxeles • RAM 512 Mb
Móvil	Requisitos mínimos: • RAM 512 MB • Wifi • Pantalla 4.5 pulgadas
Internet	Velocidad 2 Mb para visualización de la solución de minería de datos Proveedor: Telefonía Movistar Ancho de banda: 1 Mb mínimo
Base de datos PostgreSQL	Base de datos alojada en un servidor. Versión 9.3 mínima Arquitectura 64 ó 32 bits
Rstudio	IDE para lenguaje R código abierto. Versión: 1.2.1335
Power BI Pro service	Servicio de análisis, con conexión a internet
Interfaz web	Visualización de la solución de minería de datos

3.7.3. Manual de usuario

Se elaboró un manual de usuario con la finalidad de ayudar a los usuarios en el uso de la solución de minería de datos que se implementó, la cual se muestra en el *Anexo N.º 04*.

3.8. Plan de procesamiento y análisis de datos

Para el procesamiento y análisis de datos se realizará seis fases de la metodología CRISP-DM.

1. Comprensión del negocio: Se analiza la empresa, las problemáticas y requerimientos del proyecto a realizar.
2. Comprensión de los datos: Se identifican las características de los datos a analizar y recolectar.
3. Preparación de los datos: En esta fase se seleccionan los datos y se hace la limpieza para el modelamiento.
4. Modelado: En la fase del modelado con la debida preparación de los datos se procede a aplicar herramientas de minería de datos. En la cual se utilizará algoritmos de clustering.

Algoritmo de clústeres k-means, distancia; para determinar segmentos de clientes y para crear un modelo con características de homogeneidad entre clientes.

5. Evaluación: Se realizan las evaluaciones de acuerdo a los objetivos planteados en esta presente investigación.
6. Implementación: Se instalará en la empresa para la respectiva evaluación cumpliendo los segmentos de clientes arrogados por la solución de minería de datos.

3.9. Matriz de consistencia

TABLA VI
MATRIZ DE CONSISTENCIA

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES
FORMULACIÓN DEL PROBLEMA	OBJETIVO GENERAL	HIPÓTESIS	VARIABLES DE ESTUDIO
¿Cuáles son las características de los segmentos de clientes de una empresa de capacitaciones online Chiclayo?	Determinar el segmento de clientes en la empresa de capacitaciones online Chiclayo.	Desarrollo de una solución de minería de datos para la determinación de segmentos de clientes en la empresa de capacitaciones online Chiclayo.	VARIABLE INDEPENDIENTE Desarrollo de una solución de minería de datos VARIABLE DEPENDIENTE Clientes de una empresa de capacitaciones online.
OBJETIVOS ESPECÍFICOS	DESCRIPCIÓN DEL LOGRO DE LOS OBJETIVOS ESPECÍFICOS		INDICADORES
Realizar un análisis exploratorio a la base de datos.	Se obtiene un panorama sobre el negocio y la detección de datos disponibles.		Número de datos disponibles
Identificar los atributos que definan las características del cliente.	Los atributos sean identificados en base a la caracterización del cliente.		Número de atributos identificados.
Identificar algoritmos para segmentar clientes de acuerdo con sus características.	Identificación de algoritmos necesarias para la ejecución.		Número de algortimos de clustering.
Analizar los clientes de acuerdo con sus características.	Clientes con características comunes en grupos		Características formadas por la solución
Sugerir patrones en base a la solución de minería de datos.	Recomendaciones a partir de la solución		Número de grupos formados por la solución.

3.10. Consideraciones éticas

A continuación, se listan los aspectos que se han considerado para la protección y bienestar de los participantes de esta investigación, en este caso los clientes de la empresa de capacitaciones online, así como de la seguridad (resguardo) de los datos:

- ✓ Aplicación de técnicas de recolección de datos a los entrevistados. Se aplicó con debida programación con los dueños de la empresa, sin afectar los tiempos.
- ✓ Seguridad de la información sobre los datos expuestos, no se expone información completa de los clientes en esta investigación.
- ✓ Resguardo de los datos y secreto de la información, solo se proporciona acceso a los datos a personas involucradas en este caso a los dueños de la empresa capacitaciones online pactada en la firma de aceptación del tesista.
- ✓ La objetividad, se basa imparcialmente en la recolección de datos, no alterando la información a fines propios.

IV. RESULTADOS

Los resultados se presentan describiendo cada iteración de la metodología CRISP-DM.

4.1. En base a la metodología utilizada

La metodología de desarrollo es CRISP-DM, se detalla cada iteración ejecutada.

4.1.1. Iteración #1: Comprensión del negocio

- **Problemática**

  En esta parte de la fase de comprensión del negocio, se especifica la problemática de la empresa, en la cual; se ve reflejada por el envío masivo de publicidad por diversos canales como: correos, whatsapp, Facebook y llamadas sin tener en cuenta un grupo de clientes identificado acorde con sus características, ocasionando muchas veces perdidas monetaria y fastidio al cliente. En la cual, no presenta grupos con determinadas características de clientes que posee la empresa en lanzamiento de un determinado curso (evento).

- **Objetivo del negocio**
  - ✓ Mejorar sus ventas en base a sus eventos (cursos) lanzados.
  - ✓ Fomentar la recurrencia de sus clientes a matricularse en distintos eventos lanzados.

- **Descripción sobre la solución actual**

  Mediante técnicas de minería de datos de segmentación se pretende hallar segmentos de clientes acorde a sus características.

- **Inventario de recursos**

a. Recursos Hardware
   - Laptop personal con 8 GB de RAM, Core i7 y con un sistema operativo Windows 10 Pro.
   - Impresora, para imprimir información referida al proyecto, asimismo los informes y avances.

b. Recursos Software

- PgAdmin; se utilizó un gestor de base de datos libre, para la revisión de datos.
- MySQL; se utilizó para contrastar información del internet con la base de datos del sistema de escritorio.
- RStudio: IDE (Entorno de Desarrollo Integrado) empleado para el manejo del lenguaje R, siendo utilizado como herramienta para la minería de datos.
- Power BI: herramienta para visualización del resultado de la solución de minería de datos.

c. Recursos Humanos

- Personal de la empresa de capacitaciones online
- Asesores metodológicos
- Asesores especialistas
- Tesista

d. Servicios

- Energía eléctrica, utilizada para la realización del proyecto e informe.
- Internet, para la realización de investigaciones y descargas de manuales.

- **Identificación de origen de datos y almacén de conocimientos**
  - Una base de datos en un servidor en el internet, siendo MYSQL, para contrastar datos.
  - Una base de datos, PostgreSql versión 9.3 en una red local, siendo utilizada por una aplicación de escritorio.
  - Es necesario consultar base de datos externas, como RENIEC, ESSALUD y SPP, para la limpieza de datos.
- **Riesgos y contingencias**

  El análisis de riesgos en el desarrollo de la presente tesis se efectuó con la finalidad de identificar los puntos siguientes: Programación, financieros, datos y resultados los cuales son

afectados durante su desarrollo, las mismas se detallan en el Anexo N° 02.

- **Terminología**
  - ✓ Clúster: Centroides, son puntos que existen inicialmente en una tabla de datos. [17]
  - ✓ RFM: Modelo de recencia, frecuencia y valor monetario para cada cliente, donde determina el comportamiento o evolución de compra de clientes.
  - ✓ Base de datos: Según RAE; "Conjunto de datos organizados de tal modo que permita obtener con rapidez diversos tipos de información". [29]
- **Análisis de costes**

  En esta tarea de compresión de negocio, se estima un análisis de costo de la presente tesis, la cual, se detalla en el Anexo N° 04.
- **Objetivos de minería de datos**

  Como objetivo general tenemos:
  - ✓ Determinar segmentos de clientes en la empresa de capacitaciones online Chiclayo.
- **Plan del proyecto.**

  Se estima un tiempo determinado para la presente tesis en el periodo 2018 – Agosto hasta 2019- Julio. Se muestra en detalle Anexo N° 05.

4.1.2. Iteración #2: Comprensión de los datos

- **Recopilación de datos iniciales**

  Registros de la base de datos PgAdmin, utilizada por un sistema de escritorio, en donde se registran las matrículas.

  • Clientes; son aquellos que se han preinscrito en algún evento (curso) por interés de llevarlo, por otro lado, son aquellos clientes que se matriculan en algún evento en base a su interés profesional.

  • Eventos; considerados como cursos lanzados por la empresa, con fines de capacitación de acuerdo con las carreras profesionales.

- Matrícula, registros de los clientes de acuerdo con eventos, comprendidos desde el año 2014 – 2018.

- **Descripción de los datos**

  Cantidad de datos: comprendidos entre el periodo de 2014-2018.

  Clientes: se cuenta con 16960 registros de personas, entre ellos son: participantes, expositores, usuarios de la empresa y decanos de los distintos colegios a nivel nacional.

cantidad_participantes integer
16960

Fig. 4. CANTIDAD DE PARTICIPANTES

Matriculados, son aquellos clientes que llevaron un evento a más, con un total de 3643.

cantidad_matriculas integer
3643

Fig. 5. CANTIDAD DE MATRICULAS PERIODO 2014-2018

Eventos; son de tipo cursos especializados y diplomados, lanzados periódicamente sea mensual o trimestral; lo cual es un total de 191 lanzados desde 2014-2019 marzo.

cantidad_eventos integer
191

Fig. 6. CANTIDAD DE EVENTOS, PERÍODO 2014-2019

idevento integer	tipo character varying (60)	fecha_liq timestamp with time zone
2	Curso de Especialización	2015-05-14 00:00:00-05
3	Curso de Especialización	2015-03-29 00:00:00-05
4	Curso de Especialización	2015-03-22 00:00:00-05
5	Diplomado	2015-03-22 00:00:00-05
192	2019-04-24 00:00:00	2019-06-12 00:00:00

Fig. 7. FECHA DE INICIO DE EVENTOS(CURSOS)

- **Tipo de valores**

  Numéricos: Algunas columnas de datos se representan de manera numérica como: sexo representados en 0 (Masculino) y 1 (Femenino).

idsexo integer	count bigint
0	10959
1	3422

Fig. 8. CANTIDAD DE CLIENTES SEGÚN SU SEXO

Estado civil, representados de manera numérica, siendo 0 (soltero/a); 1(casado/a); 2(separado/a); 3(viudo/a); 4(conviviente).

idestadocivil integer	count bigint
2	2
4	6
0	6857
1	267
3	3

Fig. 9. CANTIDAD DE CLIENTES SEGÚN SU ESTADO CIVIL

Contrastado con la aplicación de escritorio

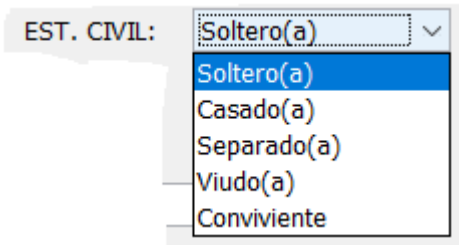


Fig. 10. ESTADO CIVIL SEGÚN EL SISTEMA

- **Exploración de datos**

  En la exploración de datos, se revisó las tablas que se utilizó para la minería de datos, descubriendo los tipos de datos que se cuenta. Las siguientes tablas que contienen los datos para el proyecto son:

  - Persona
  - Tipo de persona
  - Profesión
  - Grado académico

- Matrícula
- Cuota
- Evento
- Ubigeo

En la fig.11 y 12 se visualizó la relación entre las tablas que se utilizaron, verificando los campos de cada tabla los cuales contienen los datos necesarios para la solución de minería de datos.

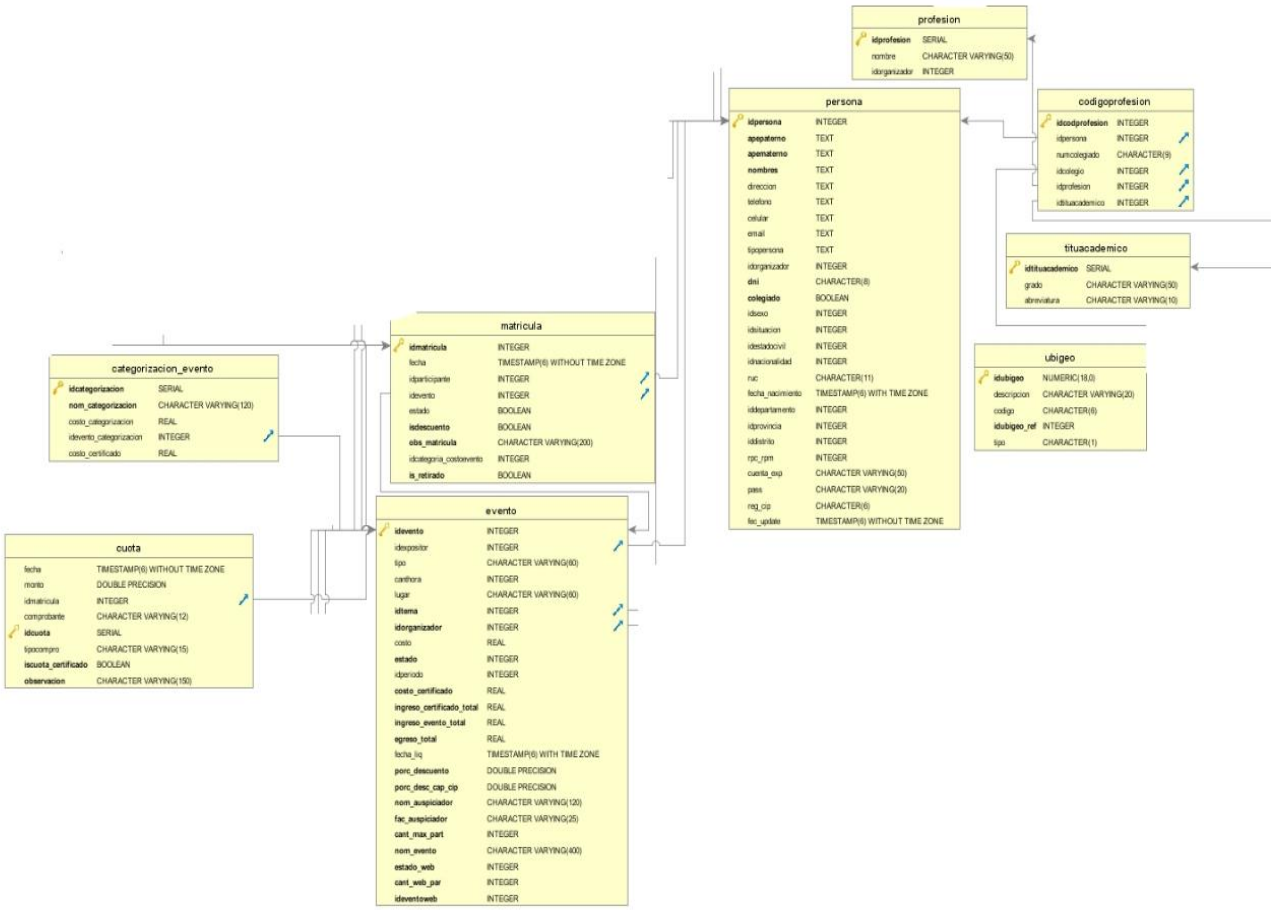


Fig. 11. TABLAS DE LA BASE DE DATOS POSTGRESQL A UTILIZAR

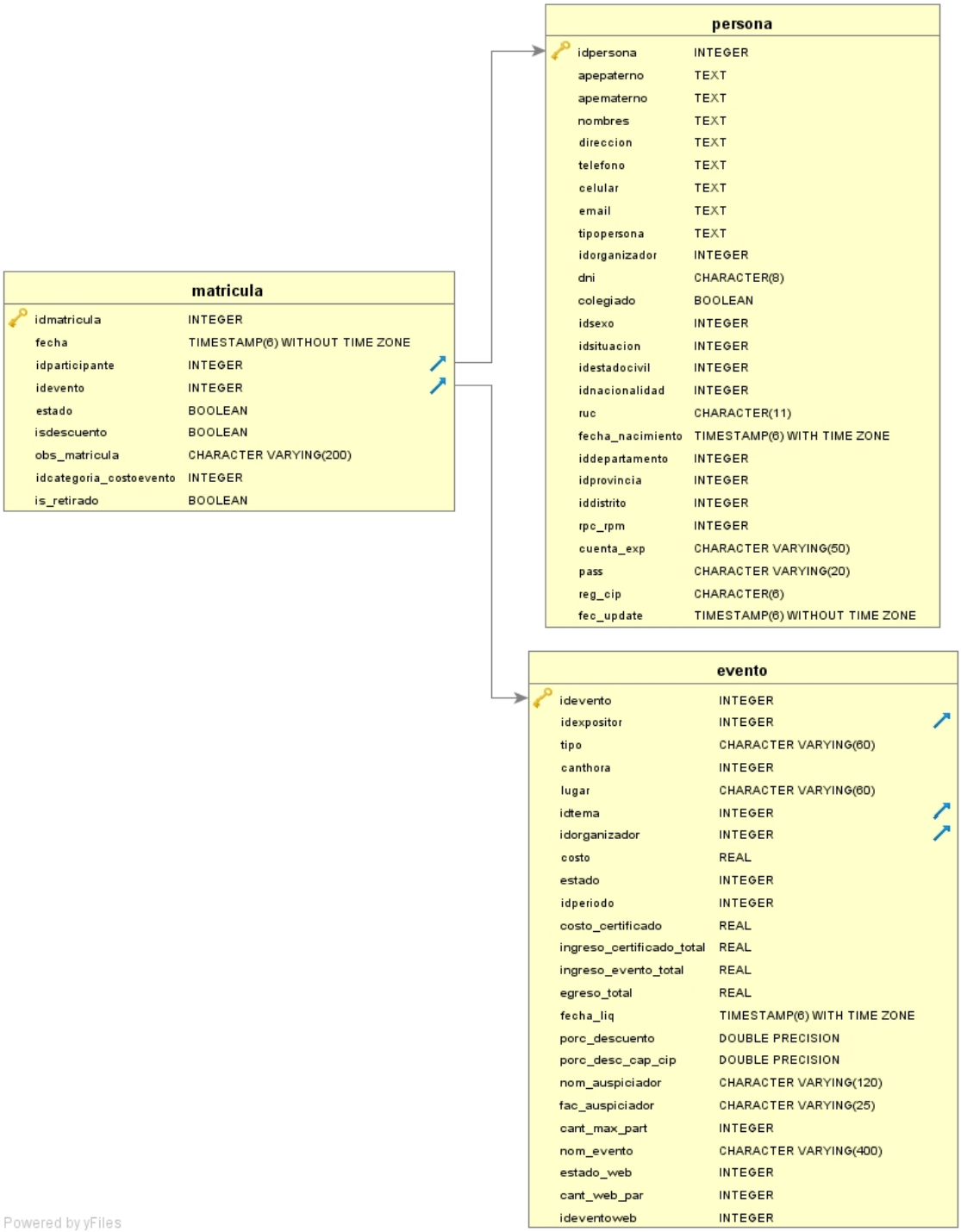


Fig. 12. CAMPOS DE LAS TABLAS VISUALIZADAS MEDIANTE HERRAMIENTA DBVISUALIZER

- **Verificación de calidad de datos**

  En la verificación de calidad de datos se pueden encontrar datos perdidos, errores de datos cometidos al ingresar al sistema de registro de clientes, incoherencias en algunos campos de la tabla de registro de personas.

  Se puede apreciar datos perdidos y errores de datos en los siguientes campos de la tabla persona.

- En el campo sexo, se puede apreciar que existe 7097 clientes sin registrar su sexo y un error de dato de -1.

	idsexo integer	count bigint
1	[null]	7097
2	0	8136
3	1	1507
4	-1	1

Fig. 13. VALORES ATÍPICOS DEL CAMPO SEXO

- En el campo estado civil, se aprecia datos perdidos null de 9843 clientes, de los cuales 6740 son solteros (0), 147 son casados (1), 2 son separados (2), 3 son viudos (3), 6 son conviviente (4).

idestadocivil integer	count bigint
2	2
4	6
[null]	9843
0	6740
1	147
3	3

Fig. 14. VALORES ATÍPICOS DEL CAMPO ESTADO CIVIL

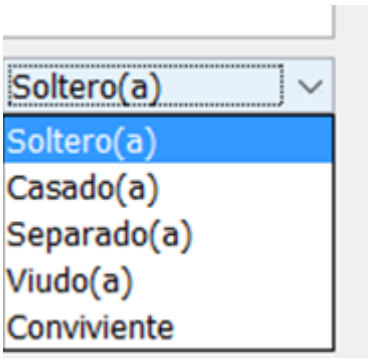


Fig. 15. ESTADO CIVIL ACORDE CON EL SISTEMA DE ESCRITORIO

- En el campo fecha de nacimiento, lo cual es importante para la determinación de la edad, 7123 campos son de tipo ‘null’.

fecha_nacimiento timestamp with time zone	count bigint
[null]	7123

Fig. 16. VALORES ATÍPICOS EN EL CAMPO FECHA DE NACIMIENTO

- En el campo tipo de persona se demuestran que son: Expositor, Participante, Miembro. En los cuales 2 se encuentran nulos, 4634 están llenado de espacios.

tipopersona text	count bigint
EXPOSITOR	1
[null]	2
EXPOSITOR	653
	4634
PARTICIPANTE	3512
MIEMBRO	7939

Fig. 17. VALORES ATÍPICOS EL CAMPO TIPO DE PERSONA

- En el campo colegiado de la tabla persona, se establece ‘true’ los que son colegiados que son 9802 y ‘false’ los que son público en general que es un total de 6939.

colegiado boolean	count bigint
false	6939
true	9802

Fig. 18.CLIENTES SEGÚN EL CAMPO COLEGIADO

- En los campos departamento, provincia y distrito se encuentran como registro ‘null’ con un total 14417 registros.

```
4  select count(*) from persona where iddepartamento is null;
5  select count(*) from persona where idprovincia is null;
6  select count(*) from persona where iddistrito is null;
```

Data Output | Explain | Messages | Query History

	count bigint
1	14417

Fig. 19. VALORES NULOS EN EL CAMPO UBIGEO

- En el campo tipo de evento, haciendo un cruce con una tabla dependiente se puede observar los tipos de eventos registrados.

tipo character varying (60)	count bigint
Curso Presencial	6
Curso de Especialidad	1
Diplomado	8
Panel Forum	1
Conferencia	1
Curso	123
Curso de Especialización	7

Fig. 20. TIPO DE EVENTO POR NORMALIZAR

4.1.3. Iteración #3: Preparación de los datos

- **Selección de datos**

  En esta fase se trabajó la selección de datos. En la cual, se preparó los datos para el modelado de la segmentación.

  Los campos que se tomaron en cuenta para la realización del proyecto de investigación se tuvieron en cuenta por los siguientes factores:

  - Variables de cuantitativo y cualitativos
  - Calidad de datos
  - Importancia en el modelado de estudio

TABLA VII
TIPOS DE DATOS Y DESCRIPCIÓN

CAMPO	DESCRIPCIÓN	TIPO DE DATO
Colegiado	Descripción si el cliente si tiene un colegiado	Bolean
Profesión	Que profesión tiene el cliente si es docente, ingeniero, enfermera o doctor	Cadena
Grado	Es el grado que el cliente pertenece.	Cadena
Estado civil	El estado civil del cliente	Cadena
Sexo	Sexo del cliente	Bolean
Edad	Edad del cliente	Numérico
Departamento	Ubicación del cliente a que departamento pertenece	Cadena
Monto promedio de consumo	Es el promedio del monto de los distintos cursos que el cliente se ha matriculado	Numérico
Número de cursos	Es la cantidad de cursos que ha cliente se ha matriculado	Numérico

- **Limpieza de datos**

  Los campos seleccionados para el modelo se tuvieron que limpiar ya que presentan datos sucios e incompletos. Para este caso se encontró errores de datos, metadatos ausentes; siendo el paso de selección y usando las validaciones de DNI con la RENIEC para actualizar estos datos.

Data Output | Explain | Messages | Query History

celular text	email text	tipopersona text	idorganizador integer	dni character (8)	colegiado boolean	idsexo integer	idsituacion integer	idestadocivil integer	idnacionalidad integer	ruc character (11)	fecha_nacimiento timestamp with time zone	iddepartamento integer
			[null]	99999993	false	[null]	[null]	[null]	[null]	[null]	[null]	[null]
			[null]	01360014	false	0	[null]	0	[null]	[null]	[null]	[null]
		MIEMBRO	5	83908390	true	[null]	[null]	[null]	[null]	[null]	1957-11-11 00:00:00-05	[null]
			[null]	10181018	false	[null]	[null]	[null]	[null]	[null]	[null]	[null]
			[null]	16727267	false	0	[null]	[null]	[null]	[null]	1975-01-02 00:00:00-05	[null]
979028...	eduar...		[null]	46593244	false	0	[null]	[null]	[null]	[null]	1989-09-08 00:00:00-05	[null]
#96969...	xhear_...	PARTICIPANTE	[null]	46679125	false	0	[null]	0	[null]	[null]	1990-11-28 00:00:00-05	[null]
950049...		PARTICIPANTE	[null]	42914494	false	0	[null]	0	[null]	[null]	1985-03-14 00:00:00-05	[null]
458818	victor_...	MIEMBRO	5	17442808	true	0	[null]	[null]	[null]	[null]	1972-03-06 00:00:00-05	[null]
			[null]	00831119	false	0	[null]	[null]	[null]	[null]	1975-12-03 00:00:00-05	[null]
	evelin_...		[null]	70891343	false	1	[null]	0	[null]	[null]	1995-09-02 00:00:00-05	[null]
979401...	carlom...	MIEMBRO	6	16735364	true	0	[null]	[null]	[null]	[null]	1970-12-16 00:00:00-05	[null]
957987...	psicolo...		[null]	16735413	false	1	[null]	0	[null]	[null]	1975-07-16 00:00:00-05	[null]
	nadia_...		[null]	72654362	false	1	[null]	0	[null]	[null]	1992-04-05 00:00:00-05	[null]
			[null]	42404059	false	1	[null]	[null]	[null]	[null]	1982-09-04 00:00:00-05	[null]
979006...	zacaria...		[null]	46675293	false	0	[null]	0	[null]	[null]	1990-11-03 00:00:00-05	[null]

Fig. 21. DATOS CON CAMPOS NULOS Y ATÍPICOS

Data Output | Explain | Messages | Query History

idsexo integer	idsituacion integer	idestadocivil integer	idnacionalidad integer	ruc character (11)	fecha_nacimiento timestamp with time zone	iddepartamento integer	idprovincia integer	iddistrito integer
[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]
0	[null]	0	[null]	[null]	[null]	[null]	[null]	[null]
[null]	[null]	[null]	[null]	[null]	1957-11-11 00:00:00-05	[null]	[null]	[null]
[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]
0	[null]	[null]	[null]	[null]	1975-01-02 00:00:00-05	[null]	[null]	[null]
0	[null]	[null]	[null]	[null]	1989-09-08 00:00:00-05	[null]	[null]	[null]
0	[null]	0	[null]	[null]	1990-11-28 00:00:00-05	[null]	[null]	[null]
0	[null]	0	[null]	[null]	1985-03-14 00:00:00-05	[null]	[null]	[null]
0	[null]	[null]	[null]	[null]	1972-03-06 00:00:00-05	[null]	[null]	[null]
0	[null]	[null]	[null]	[null]	1975-12-03 00:00:00-05	[null]	[null]	[null]
1	[null]	0	[null]	[null]	1995-09-02 00:00:00-05	[null]	[null]	[null]
0	[null]	[null]	[null]	[null]	1970-12-16 00:00:00-05	[null]	[null]	[null]
1	[null]	0	[null]	[null]	1975-07-16 00:00:00-05	[null]	[null]	[null]
1	[null]	0	[null]	[null]	1992-04-05 00:00:00-05	[null]	[null]	[null]
1	[null]	[null]	[null]	[null]	1982-09-04 00:00:00-05	[null]	[null]	[null]
0	[null]	0	[null]	[null]	1990-11-03 00:00:00-05	[null]	[null]	[null]

Fig. 22. DATOS CON CAMPOS NULOS II

Para este proceso de limpieza de datos se tomaron en cuenta los clientes matriculados en distintos eventos, en los cuales son 2482 clientes matriculados en distintos eventos. Asimismo, se tuvo que realizar un algoritmo para actualizar datos faltantes en cuanto al sexo, fecha de nacimiento, ubigeo, estado civil y contrastación de datos correctos.

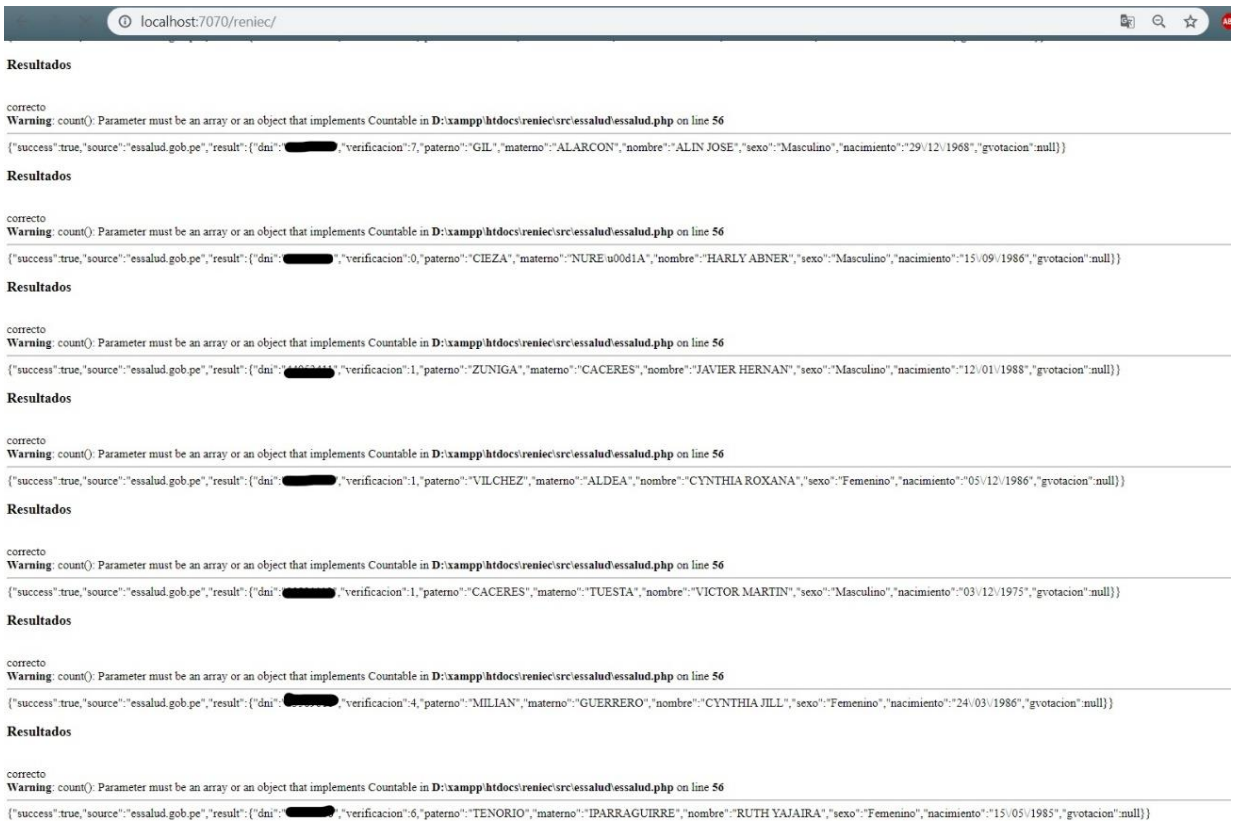


Fig. 23. LIMPIEZA DE DATOS CON FUENTES DE DATOS EXTERNAS RENIEC

- **Construcción de nuevos datos**

  En esta etapa se construye nuevos datos, seguida de la limpieza de datos. Se hace selección de los campos a implementar el nuevo modelo, separando los campos requeridos. En este caso se utilizó clientes con monto de consumo > 0; puesto que ellos generan ingresos a la empresa.

```
#conexión a la base de datos
con = dbConnect(PostgreSQL(),user="postgres",password="atto",dbname="bdTESIS")
## consulta de la base de datos, datos seleccionados
consulta=dbGetQuery(con,"select DISTINCT(p.idpersona),p.dni, (case when p.colegiado=true then 'No' else 'Si' end) as colegiado,
pf.nombre as profesion, ta.grado as Grado,
p.idestadocivil as estadoCivil, (case p.idsexo when '0' then 'M' else 'F' end) as sexo,
date_part('year', age( p.fecha_nacimiento))::integer as edad, ub.descripcion as departamento,
(select avg (ct.monto) from cuota ct inner join matricula mat on mat.idmatricula=ct.idmatricula
where mat.idparticipante = p.idpersona and ct.monto !=0 AND ct.monto is not null)::int  as montopromconsumo,
(select count(*) from matricula mat where mat.idparticipante=p.idpersona)::integer as cantCursos
from matricula mat inner join evento e on mat.idevento=e.idevento
inner join cuota ct on mat.idmatricula=ct.idmatricula
inner join persona p on mat.idparticipante=p.idpersona
inner join codigoprofesion cp on p.idpersona=cp.idpersona
inner join tituacademico  ta on cp.idtituacademico=ta.idtituacademico
inner join profesion pf on cp.idprofesion=pf.idprofesion
inner join ubigeo ub on p.iddepartamento=ub.idubigeo
where p.fecha_nacimiento is not null and ct.monto is not null")

#Limpiando datos nulos
datosRed <- na.omit(consulta)
```

Fig. 24. CONEXIÓN A LA BASE DE DATOS Y OMITIENDO VALORES NULOS

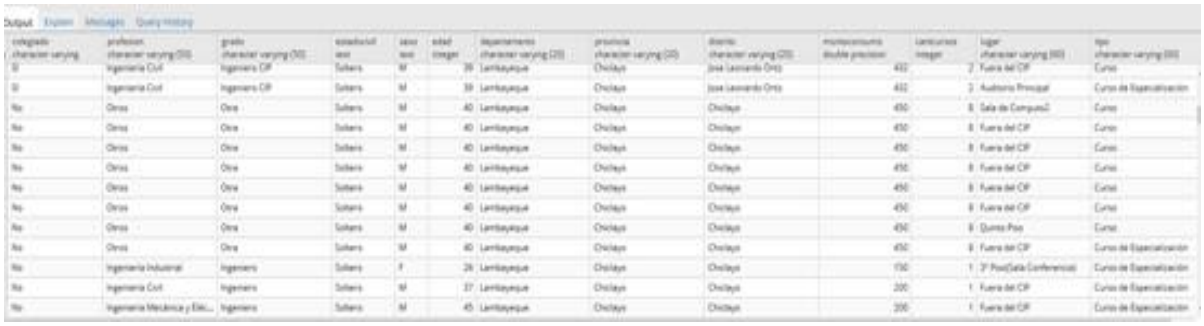


Fig. 25. DATOS NUEVOS

- **Integración de datos**

  En esta etapa se integra datos como los eventos llevados por un cliente, lo cual es mostrado como la cantidad de cursos llevados, este se identifica con el ID evento, el ID matricula asociado con el ID del cliente en el cual se identifica correctamente sin la duplicidad de datos.

idpersona	dni	colegiado	profesion	grado	estadocivil	sexo	edad	departamento	montopromconsumo	cantcursos
16604	45234911	Si	IngenierÃa Ambiental	Otra	0	F	32	Lima	765	1
15324	46473959	Si	IngenierÃa Civil	Ingeniero	0	M	29	Apurimac	750	1
15325	46014876	Si	Arquitectura	Arquitecto	0	M	31	Arequipa	750	1
15329	40560215	Si	Arquitectura	Arquitecto	0	F	39	Lima	750	1
15332	15732193	No	IngenierÃa Civil	Ingeniero CIP	1	M	59	Lima	750	1
16827	18109455	Si	Arquitectura	Arquitecto	0	M	49	Piura	700	1
15215	24487832	Si	IngenierÃa Civil	Otra	0	M	44	Cusco	675	2
16683	43572704	Si	IngenierÃa Civil	Magister	0	M	33	Cusco	630	1
16780	40280600	Si	IngenierÃa Civil	Magister	0	M	41	Cusco	630	2
15204	42632863	Si	Agroindustrial	Ingeniero	0	M	35	Apurimac	600	1
15211	07515921	Si	Otros	Otra	1	F	51	Lima	600	1
15221	04431287	Si	IngenierÃa Civil	Ingeniero	0	F	46	Tacna	600	1
15356	46245606	Si	Industrial	Ingeniero	0	M	29	Lima	600	1
16128	45223487	Si	Agroindustrial	Ingeniero	0	M	31	Apurimac	565	2
15334	40477962	Si	IngenierÃa Civil	Ingeniero	0	M	41	Apurimac	548	5
16256	16698726	Si	Arquitectura	Arquitecto	1	M	51	Piura	510	2
15323	45766927	Si	Arquitectura	Arquitecto	0	F	30	Cusco	500	2

ng 1 to 21 of 2,263 entries, 11 total columns

Fig. 26. VERIFICACIÓN DE INTEGRACIÓN DE DATOS

Se realizó la integración de datos de los clientes matriculados, IdPersona, colegiado, profesión, grado académico, estado civil, sexo, ubigueo, lugar. Además, se integró un conjunto de datos calculados como: el promedio de monto consumo, edad (calculada por la fecha de nacimiento) y la cantidad de cursos llevados por cada participante.

```
> summary(consulta)
   idpersona       colegiado        profesion         grado         estadocivil          sexo             edad        iddepartamento
 Min.   :   90   Min.   :0.0000   Min.   : 1.00   Min.   : 1.000   Min.   :0.00000   Min.   :0.0000   Min.   :19.00   Min.   :   1.0
 1st Qu.:15268   1st Qu.:1.0000   1st Qu.: 8.00   1st Qu.: 2.000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:30.00   1st Qu.: 367.0
 Median :15878   Median :1.0000   Median :17.00   Median : 3.000   Median :0.00000   Median :0.0000   Median :37.00   Median :1118.0
 Mean   :15486   Mean   :0.9458   Mean   :30.84   Mean   : 7.315   Mean   :0.04638   Mean   :0.2215   Mean   :39.13   Mean   : 944.8
 3rd Qu.:16549   3rd Qu.:1.0000   3rd Qu.:61.00   3rd Qu.:11.000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:46.00   3rd Qu.:1389.0
 Max.   :17181   Max.   :1.0000   Max.   :99.00   Max.   :47.000   Max.   :1.00000   Max.   :1.0000   Max.   :73.00   Max.   :2034.0
  montoconsumo      cantcursos       lugar
 Min.   :  25.0   Min.   : 1.0   Min.   :1
 1st Qu.: 130.0   1st Qu.: 1.0   1st Qu.:1
 Median : 220.0   Median : 1.0   Median :1
 Mean   : 305.4   Mean   : 1.4   Mean   :1
 3rd Qu.: 320.0   3rd Qu.: 1.0   3rd Qu.:1
 Max.   :5488.0   Max.   :17.0   Max.   :1
```

Fig. 27. CONSULTA DE DATOS EN RSTUDIO

- **Formato de datos**

  Los algoritmos usados en la fase de modelado son sensibles a datos atípicos; lo cual se busca la frecuencia de inscripción de los participantes, reduciendo datos engañosos. Así mismo, se utilizó para el modelo a evaluar variables mixtas (variables cuantitativas, variables cualitativas).

4.1.4. Iteración #4: Modelado

- **Selección de técnica de modelado**

  El modelo para evaluar comprende de variables mixtas, para poder determinar se utilizó como técnica de segmentación ya que procesa tanto variables cualitativas y mixtas, maximizando la similitud de variables.

  Considerando que existe algoritmos para la segmentación se consideró algoritmo k-means y distancias.

- **Generación de modelos**

  Para la selección de grupos(clúster) se aplicó la técnica de partición, hallando la matriz de distancias, el clúster más cercano, clúster más lejano, clúster promedio y clúster enlace centroide. Como pasos siguientes en generación de grupos son:

  - Selección k centroides aleatoriamente.
  - Creación de clústeres asignado a centroide más cercano.
  - Creación de clústeres asignado a centroide más lejano.
  - Creación de clústeres asignado a centroide promedio.
  - Evaluación el número de clúster.
  - Aplicación del algoritmo k-means.

- **Evaluación de los modelos**

  Para la evaluación de las variables se utilizó la gráfica de caja para proporcionar el tamaño de muestra y verificar los valores.

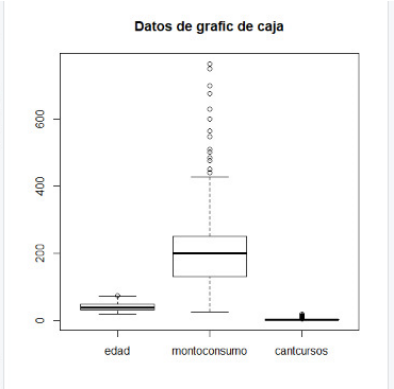


Fig. 28. GRÁFICA DE DATOS POR BOXPLOT

Se realizó la comparación de clústeres generados por los métodos de distancia. En la cual se obtuvo una distancia menor de 0.26 a partir de los métodos utilizados como: centroid, Ward.D2, single, average, complete.

```
#-------------------------------------------------------------------#
#                      Comparando clusters                          #
#-------------------------------------------------------------------#
library(dendextend)
# Creación dos dendrograms
dend1 <- as.dendrogram (Clus3)
dend2 <- as.dendrogram (Clus5)
# Create a list to hold dendrograms
dend_list <- dendlist(dend1, dend2)
tanglegram(dend1, dend2,
           highlight_distinct_edges = FALSE, # Turn-off dashed lines
           common_subtrees_color_lines = TRUE, # Turn-off line colors
           common_subtrees_color_branches = TRUE, # Color common branches
           main = paste("entanglement =", round(entanglement(dend_list), 2))
)
```

Fig. 29. CÓDIGO COMPARACIÓN DE MÉTODOS

Se comparo los dendrogramas con diferentes algoritmos, permitiendo visualizar los agrupamientos con un enlazamiento mínimo obteniendo el óptimo algoritmo.

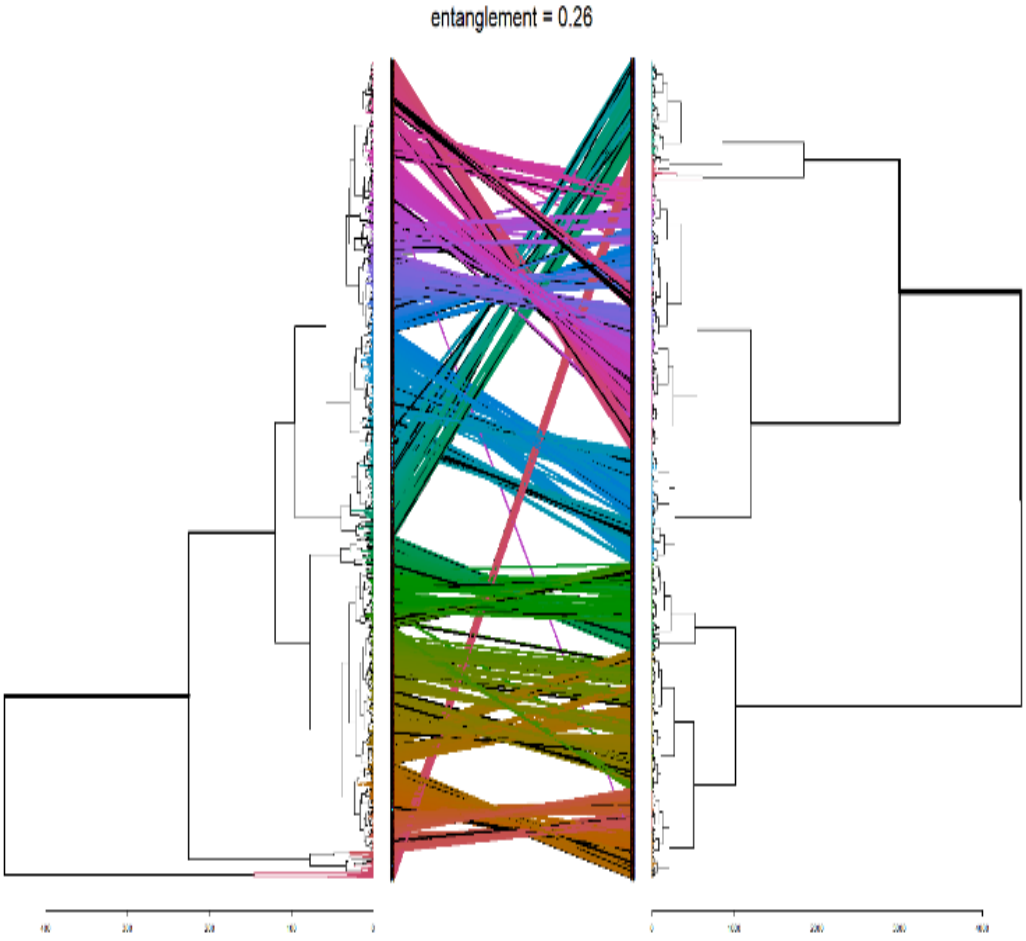


Fig. 30. COMPARACIÓN DE DENTOGRAMAS EN RSTUDIO

Se realizo una comparación de métodos como: Single, Complete, Average, Centroid y Ward, en la cual, agrupa puntos similares de un grupo y separar los diferentes atributos en otros grupos.

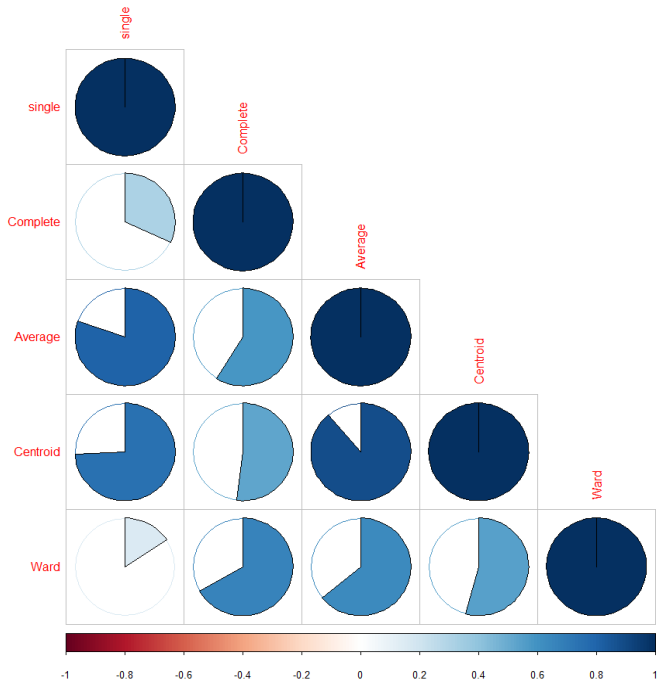


Fig. 31. COMPARATIVA DE METODOS PARA ELECCIÓN DEL CLÚSTER ÓTIMO

Para la elección del optimo clúster se realizó la medida de distancia euclidean y suma de error. De acuerdo con la función como número de clúster arrojado como optimo es 3.

```
#Hallando k
  # Crea vector "Errores", sin datos
  # Crea variable "K_Max" con la cant. maxima de k a analizar
Errores <-NULL
K_Max   <-10

#------------------------------------------------------------------------------
#HALLANDO CLUSTER PARA CONSULTA datosRed
for (i in 1:K_Max)
{
  Errores[i] <- sum(kmeans(DatosRed[-1], centers=i)$withinss)
}
# versi?n con suma de errores
wss <-(nrow(DatosRed)-1*sum(apply(DatosRed, 2, var)))

for (i  in 2:15) {
  wss[i] <- sum(kmeans(DatosRed,centers = i , nstart = 25)$withinss)
}

plot(1:15, wss, type = "b",
     xlab = "N?mero de clusters",
     ylab = "suma del error al cuadrado"
     )
```

Fig. 32.CÓDIGO DE SUMA DE ERROR PARA ENCONTRAR EL ÓPTIMO K.

```
> res <- NbClust(data = Datos1, diss = NULL, distance = "euclidean", min.nc = 2, max.nc = 15,method = "ward.D2", index = "all", alphaBeale = 0.1)
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 3 proposed 2 as the best number of clusters
* 6 proposed 3 as the best number of clusters
* 3 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 2 proposed 8 as the best number of clusters
* 3 proposed 11 as the best number of clusters
* 2 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
>
```

Fig. 33. CLÚSTER ÓPTIMO MEDIANTE DISTANCIA EUCLIDEAN RSTUDIO

Agrupando por 3 grupos (clústeres) por homogeneidad de sus características.

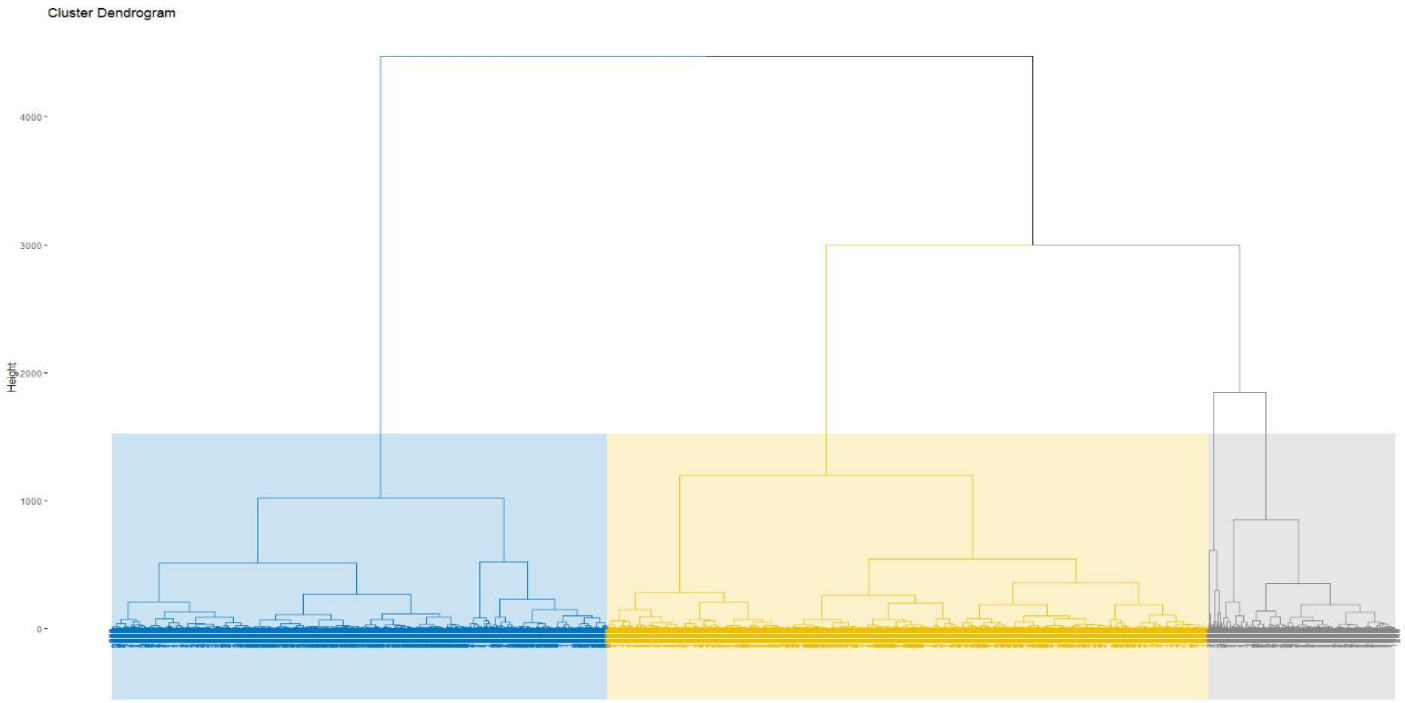


Fig. 34. FORMACIÓN DE GRUPOS POR HOMOGENIEDAD RSTUDIO

Grupos formados por la solución, siendo el primer grupo 46.86% del total con 1163 clientes, segundo grupo 38.6% con 958 clientes y tercer grupo 14.56% con 351 clientes en la empresa de capacitaciones online.

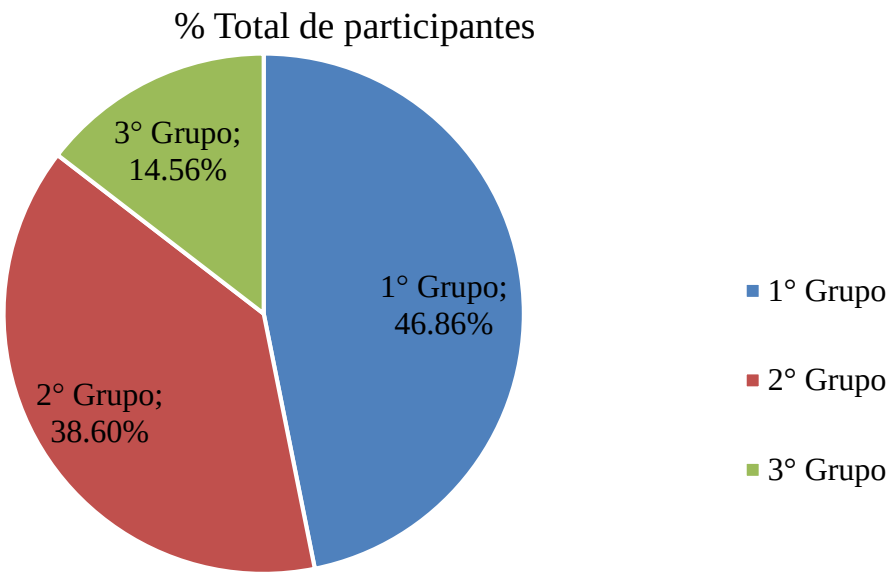
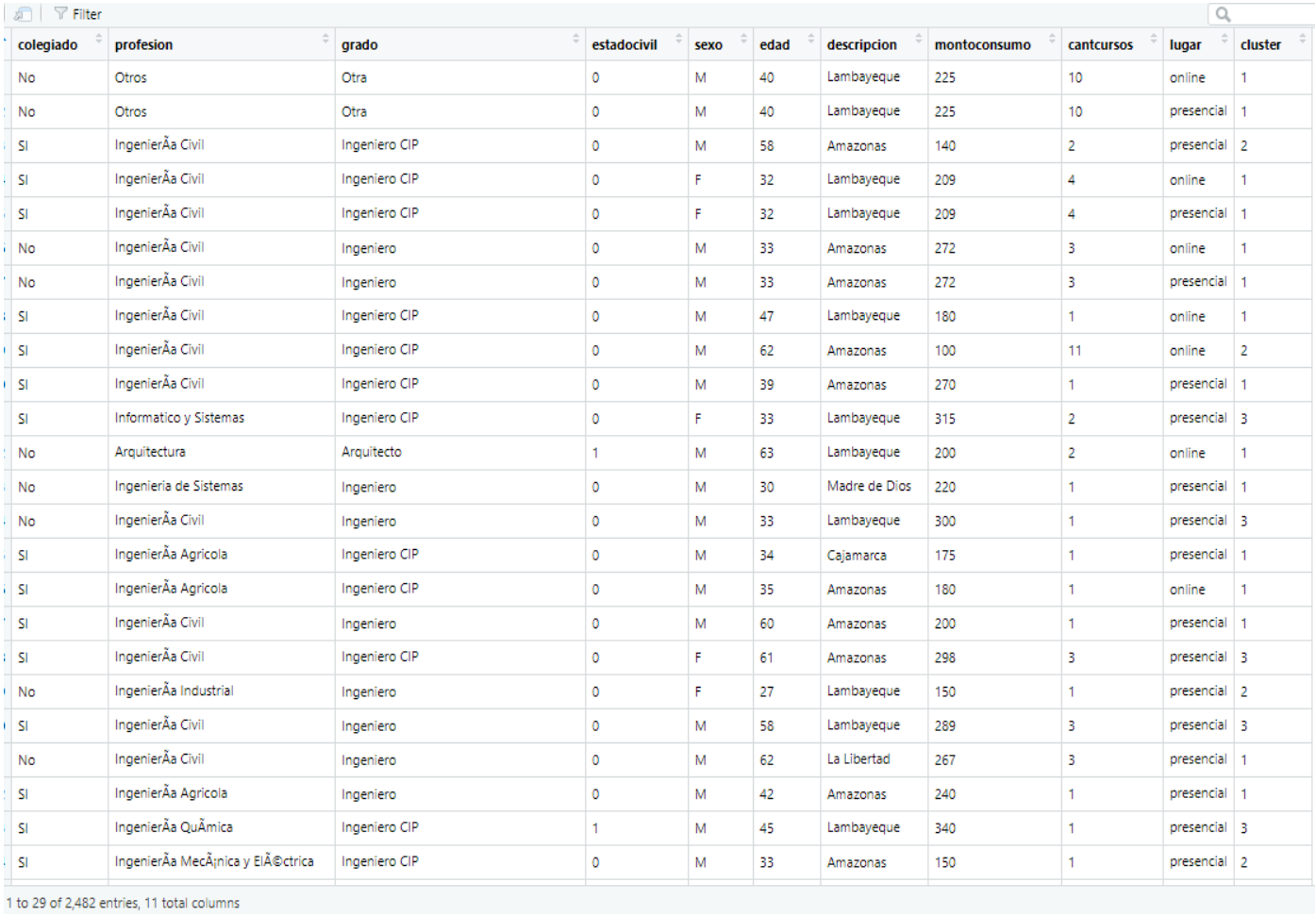


Fig. 35. GRÁFICA CIRCULAR SOBRE EL PORCENTAJE TOTAL QUE REPRESENTA CADA CLÚSTER.

TABLA VIII

Cantidad de clientes por clúster

Clúster	Cantidad de clientes
1	1163
2	958
3	361



Filter

colegiado	profesion	grado	estadocivil	sexo	edad	descripcion	montoconsumo	cantcursos	lugar	cluster
No	Otros	Otra	0	M	40	Lambayeque	225	10	online	1
No	Otros	Otra	0	M	40	Lambayeque	225	10	presencial	1
SI	IngenierÃa Civil	Ingeniero CIP	0	M	58	Amazonas	140	2	presencial	2
SI	IngenierÃa Civil	Ingeniero CIP	0	F	32	Lambayeque	209	4	online	1
SI	IngenierÃa Civil	Ingeniero CIP	0	F	32	Lambayeque	209	4	presencial	1
No	IngenierÃa Civil	Ingeniero	0	M	33	Amazonas	272	3	online	1
No	IngenierÃa Civil	Ingeniero	0	M	33	Amazonas	272	3	presencial	1
SI	IngenierÃa Civil	Ingeniero CIP	0	M	47	Lambayeque	180	1	online	1
SI	IngenierÃa Civil	Ingeniero CIP	0	M	62	Amazonas	100	11	online	2
SI	IngenierÃa Civil	Ingeniero CIP	0	M	39	Amazonas	270	1	presencial	1
SI	Informatico y Sistemas	Ingeniero CIP	0	F	33	Lambayeque	315	2	presencial	3
No	Arquitectura	Arquitecto	1	M	63	Lambayeque	200	2	online	1
No	Ingenieria de Sistemas	Ingeniero	0	M	30	Madre de Dios	220	1	presencial	1
No	IngenierÃa Civil	Ingeniero	0	M	33	Lambayeque	300	1	presencial	3
SI	IngenierÃa Agricola	Ingeniero CIP	0	M	34	Cajamarca	175	1	presencial	1
SI	IngenierÃa Agricola	Ingeniero CIP	0	M	35	Amazonas	180	1	online	1
SI	IngenierÃa Civil	Ingeniero	0	M	60	Amazonas	200	1	presencial	1
SI	IngenierÃa Civil	Ingeniero CIP	0	F	61	Amazonas	298	3	presencial	3
No	IngenierÃa Industrial	Ingeniero	0	F	27	Lambayeque	150	1	presencial	2
SI	IngenierÃa Civil	Ingeniero	0	M	58	Lambayeque	289	3	presencial	3
No	IngenierÃa Civil	Ingeniero	0	M	62	La Libertad	267	3	presencial	1
SI	IngenierÃa Agricola	Ingeniero	0	M	42	Amazonas	240	1	presencial	1
SI	IngenierÃa QuÃmica	Ingeniero CIP	1	M	45	Lambayeque	340	1	presencial	3
SI	IngenierÃa MecÃ¡nica y ElÃ©ctrica	Ingeniero CIP	0	M	33	Amazonas	150	1	presencial	2

1 to 29 of 2,482 entries, 11 total columns

Fig. 36. DATOS CON SUS RESPECTIVOS GRUPOS(CLÚSTER)

4.1.5. Iteración #5: Evaluación

- **Evaluación de los resultados**

  En esta fase se evalúa los resultados obtenidos a través del modelado lo cual, son 3 grupos(clústeres) mostrando información de acuerdo con su edad, cantidad de cursos llevados y el monto promedio de consumo.

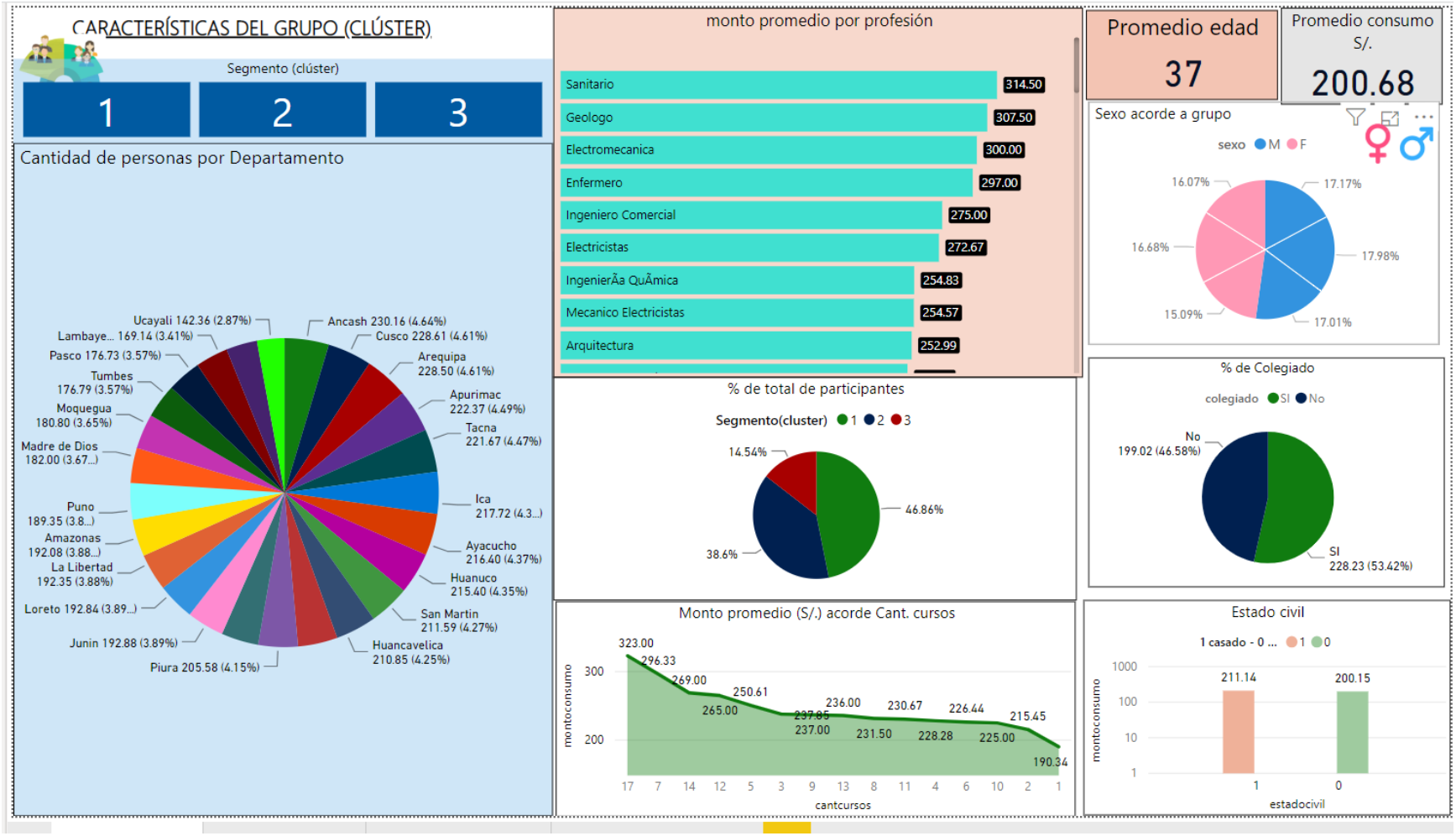


Fig. 37. INFORMACIÓN GENERAL DE CADA GRUPO(CLÚSTER)

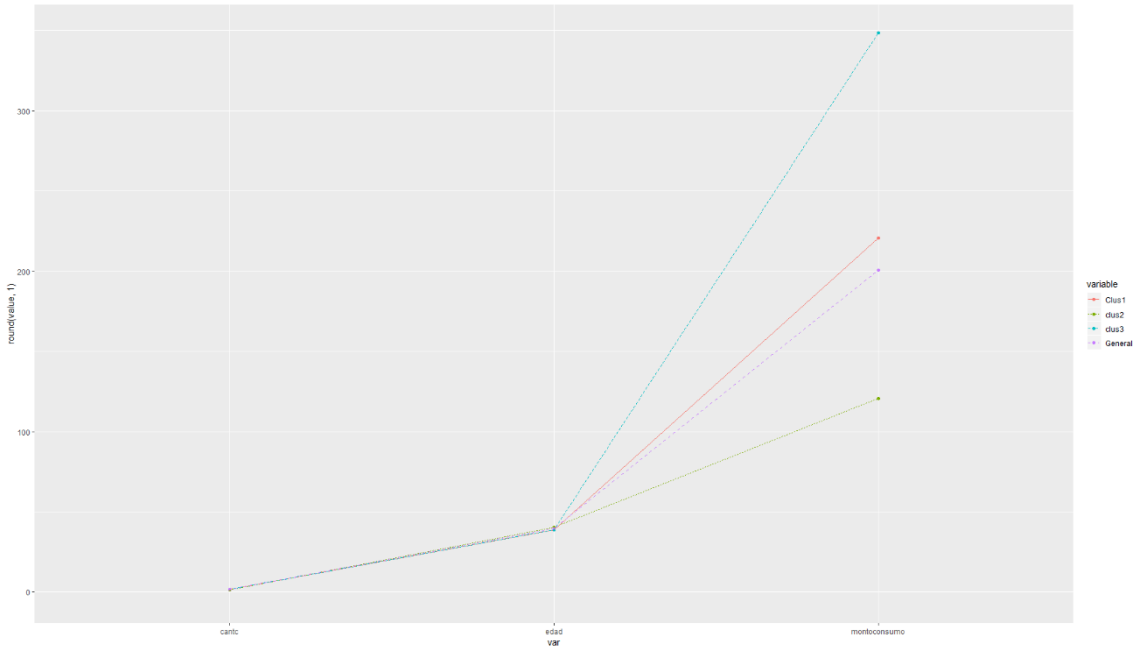


Fig. 38. INFORMACIÓN ACORDE CON VARIABLES CUANTITATIVAS

Mostrando datos característicos de cada grupo(clúster) generado, acorde con los atributos evaluados.

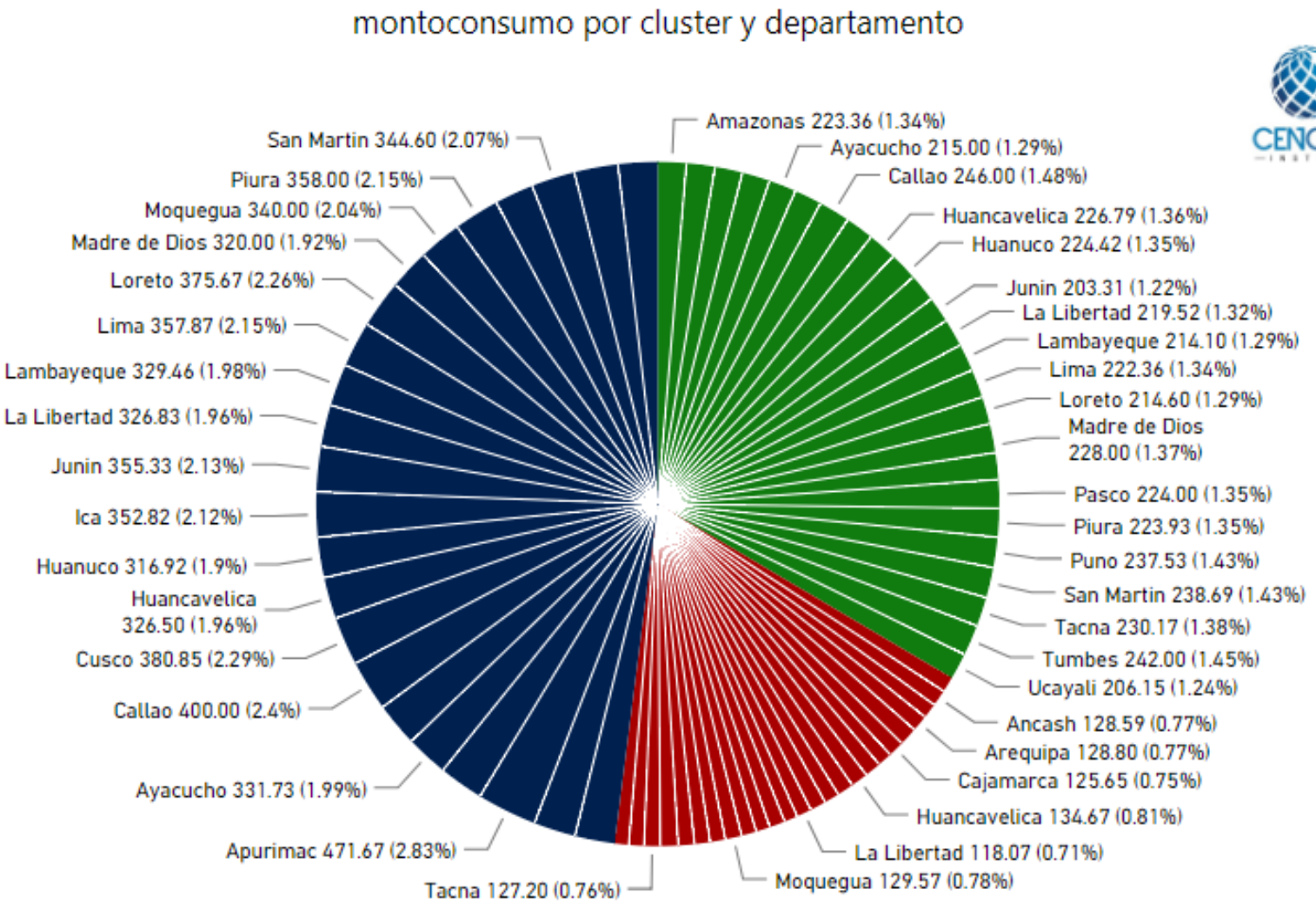


Fig. 39. DATOS AGRUPADOS POR DEPARTAMENTO, GRUPOS (CLÚSTERES)

Como resultados obtenidos del clúster 1 son:

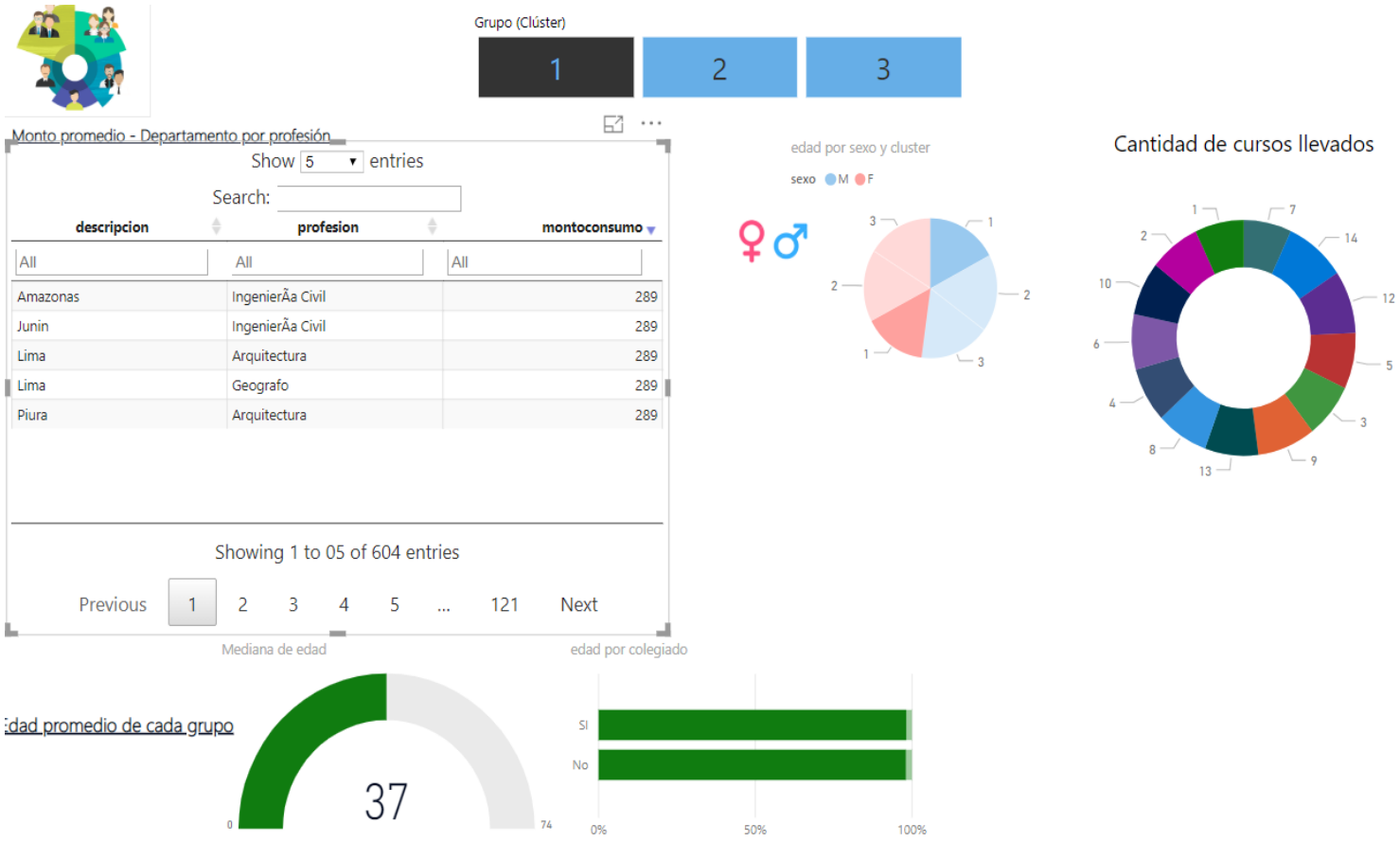
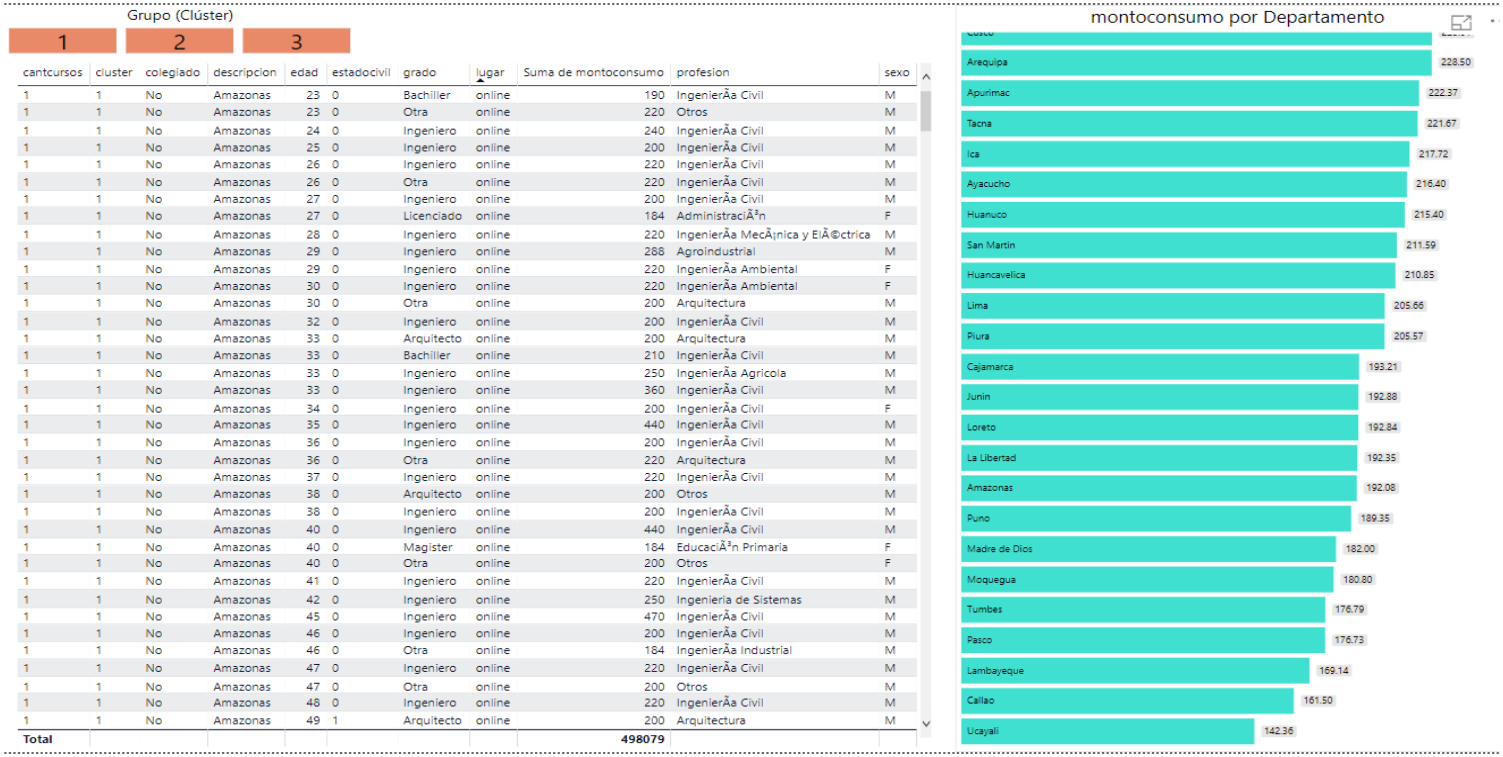


Fig. 40. RESULTADOS DEL CLÚSTER 1

- ✓ El 68.27% de 1163 clientes son de ingeniería que no cuentan con un colegiado y, el 31.73% son de otras especialidades.
- ✓ El rango de edades de este grupo 1, son: 21 – 63 años, con una frecuencia de clientes de 37 años.

- ✓ El monto promedio del grupo 1: s/. 220.78 soles.
- ✓ Clientes principales son del distrito de Callao con 1.48% de la muestra en total, los cuales son clientes no colegiados con un promedio de edad 37 años.
- ✓ El sexo que predomina el clúster 1 es masculino con un 17.17% del total de la muestra.
- ✓ El departamento de Lambayeque, donde cuenta con convenios con el Colegio de Ingenieros, tiene una aceptación 1.29% en total de la muestra, con un promedio de consumo s/. 214 soles.

Resultados del clúster 2:



cantcursos	cluster	colegiado	descripcion	edad	estadocivil	grado	lugar	Suma de montoconsumo	profesion	sexo
1	1	No	Amazonas	23	0	Bachiller	online	190	IngenierÃa Civil	M
1	1	No	Amazonas	23	0	Otra	online	220	Otros	M
1	1	No	Amazonas	24	0	Ingeniero	online	240	IngenierÃa Civil	M
1	1	No	Amazonas	25	0	Ingeniero	online	200	IngenierÃa Civil	M
1	1	No	Amazonas	26	0	Ingeniero	online	220	IngenierÃa Civil	M
1	1	No	Amazonas	26	0	Otra	online	220	IngenierÃa Civil	M
1	1	No	Amazonas	27	0	Ingeniero	online	200	IngenierÃa Civil	M
1	1	No	Amazonas	27	0	Licenciado	online	184	AdministraciÃ³n	F
1	1	No	Amazonas	28	0	Ingeniero	online	220	IngenierÃa MecÃ¡nica y ElÃ©ctrica	M
1	1	No	Amazonas	29	0	Ingeniero	online	288	Agroindustrial	M
1	1	No	Amazonas	29	0	Ingeniero	online	220	IngenierÃa Ambiental	F
1	1	No	Amazonas	30	0	Ingeniero	online	220	IngenierÃa Ambiental	F
1	1	No	Amazonas	30	0	Otra	online	200	Arquitectura	M
1	1	No	Amazonas	32	0	Ingeniero	online	200	IngenierÃa Civil	M
1	1	No	Amazonas	33	0	Arquitecto	online	200	Arquitectura	M
1	1	No	Amazonas	33	0	Bachiller	online	210	IngenierÃa Civil	M
1	1	No	Amazonas	33	0	Ingeniero	online	250	IngenierÃa Agrícola	M
1	1	No	Amazonas	33	0	Ingeniero	online	360	IngenierÃa Civil	M
1	1	No	Amazonas	34	0	Ingeniero	online	200	IngenierÃa Civil	F
1	1	No	Amazonas	35	0	Ingeniero	online	440	IngenierÃa Civil	M
1	1	No	Amazonas	36	0	Ingeniero	online	200	IngenierÃa Civil	M
1	1	No	Amazonas	36	0	Otra	online	220	Arquitectura	M
1	1	No	Amazonas	37	0	Ingeniero	online	220	IngenierÃa Civil	M
1	1	No	Amazonas	38	0	Arquitecto	online	200	Otros	M
1	1	No	Amazonas	38	0	Ingeniero	online	200	IngenierÃa Civil	M
1	1	No	Amazonas	40	0	Ingeniero	online	440	IngenierÃa Civil	M
1	1	No	Amazonas	40	0	Magister	online	184	EducaciÃ³n Primaria	F
1	1	No	Amazonas	40	0	Otra	online	200	Otros	F
1	1	No	Amazonas	41	0	Ingeniero	online	220	IngenierÃa Civil	M
1	1	No	Amazonas	42	0	Ingeniero	online	250	Ingenieria de Sistemas	M
1	1	No	Amazonas	45	0	Ingeniero	online	470	IngenierÃa Civil	M
1	1	No	Amazonas	46	0	Ingeniero	online	200	IngenierÃa Civil	M
1	1	No	Amazonas	46	0	Otra	online	184	IngenierÃa Industrial	M
1	1	No	Amazonas	47	0	Ingeniero	online	220	IngenierÃa Civil	M
1	1	No	Amazonas	47	0	Otra	online	200	Otros	M
1	1	No	Amazonas	48	0	Ingeniero	online	220	IngenierÃa Civil	M
1	1	No	Amazonas	49	1	Arquitecto	online	200	Arquitectura	M
Total								**498079**		

Fig. 41. RESULTADOS DEL CLÚSTER 2

- ✓ El grupo(clúster) 2 cuenta con un promedio de monto consumo de s/. 120.57 soles, un 17.98% de sexo masculino y 16.68% sexo femenino.
- ✓ Como departamento principal de consumo es Ayacucho con promedio de consumo s/. 143.26 soles, con profesión notoria de ingeniería civil, arquitectura y docencia.
- ✓ La edad promedio del grupo 2 es de 39 años, con un intervalo de 19 – 60 años de clientes.

✓ En el departamento de Lambayeque tiene un promedio consumo de s/. 105.23 soles, como principales clientes son de profesión ingeniería civil, arquitectura, docentes.

✓ Las profesiones de administración y contabilidad tienen impacto en los departamentos de Tumbes, Amazonas, Lima, Ica, Arequipa, Cajamarca, Huancavelica, Ucayali, Moquegua con un intervalo de monto promedio de s/.130 y s/.162 soles.

Resultados del clúster 3:

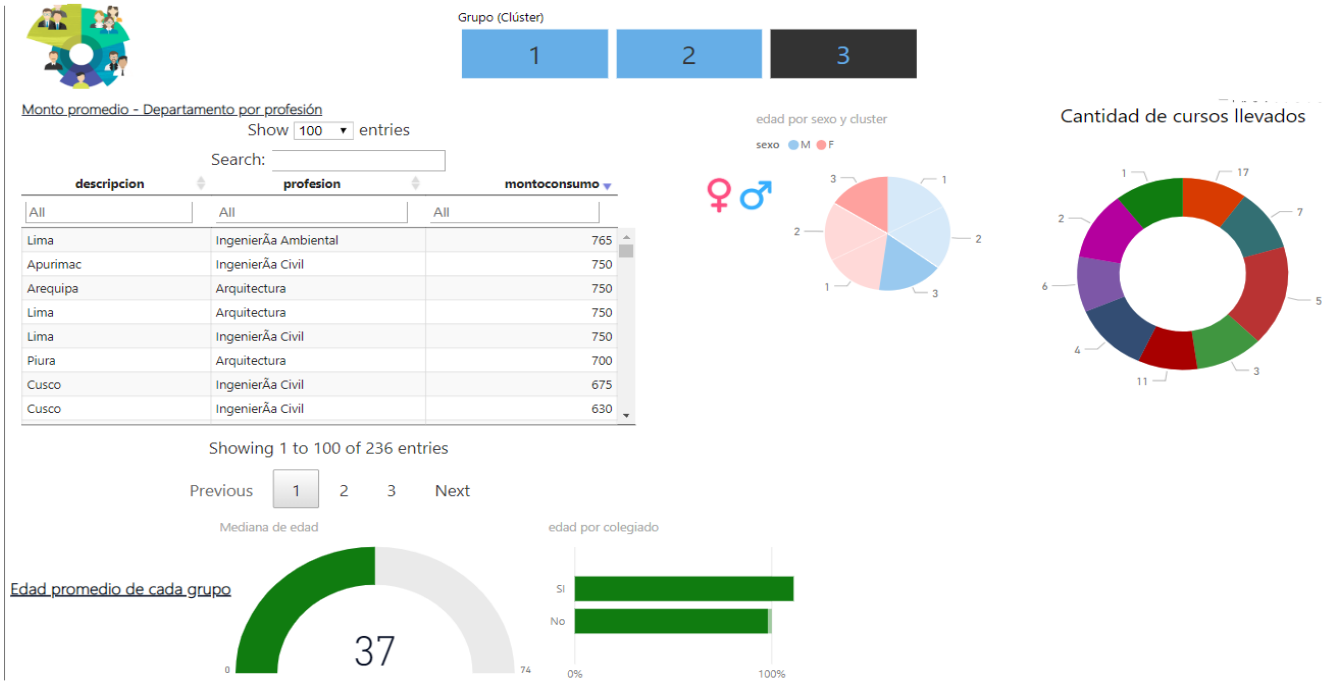


Fig. 42. RESULTADOS DEL CLÚSTER 3

✓ Su monto promedio de consumo es de s/348.50 soles, con una edad promedio de 37 años.

- ✓ Departamento de mayor consumo es Apurímac, con un promedio de s/. 471.67 soles con una mayoría de sexo masculino y de profesión ingeniería civil y docencia.
- ✓ La provincia Callao del departamento de Lima, es el segundo consumidor con un monto promedio de s/400 soles y de profesión ingeniería civil con sexo masculino.
- ✓ El departamento de Lambayeque cuenta con monto promedio de consumo de s/.329 soles, con un intervalo de edad 23 – 66 años, como profesiones ingeniería civil, arquitectura, docentes, enfermería, contadores y administradores.
- ✓ Se desataca en grupo de consumo con 5 cursos llevados siendo el 16.92% del grupo 3, con un promedio de consumo s/.548 soles.

4.1.6. Iteración #6: Distribución

Esta fase de distribución está compuesta por el monitoreo y mantención del producto que se ha realizado a través de la herramienta Rstudio para la transformación de datos, Power BI para graficar los resultados obtenidos y mostrarlos en un dashboard al usuario final que en este caso son los dueños de la empresa de capacitaciones online. Teniendo como interfaz segmentación y manual de usuario para su respectivo uso.

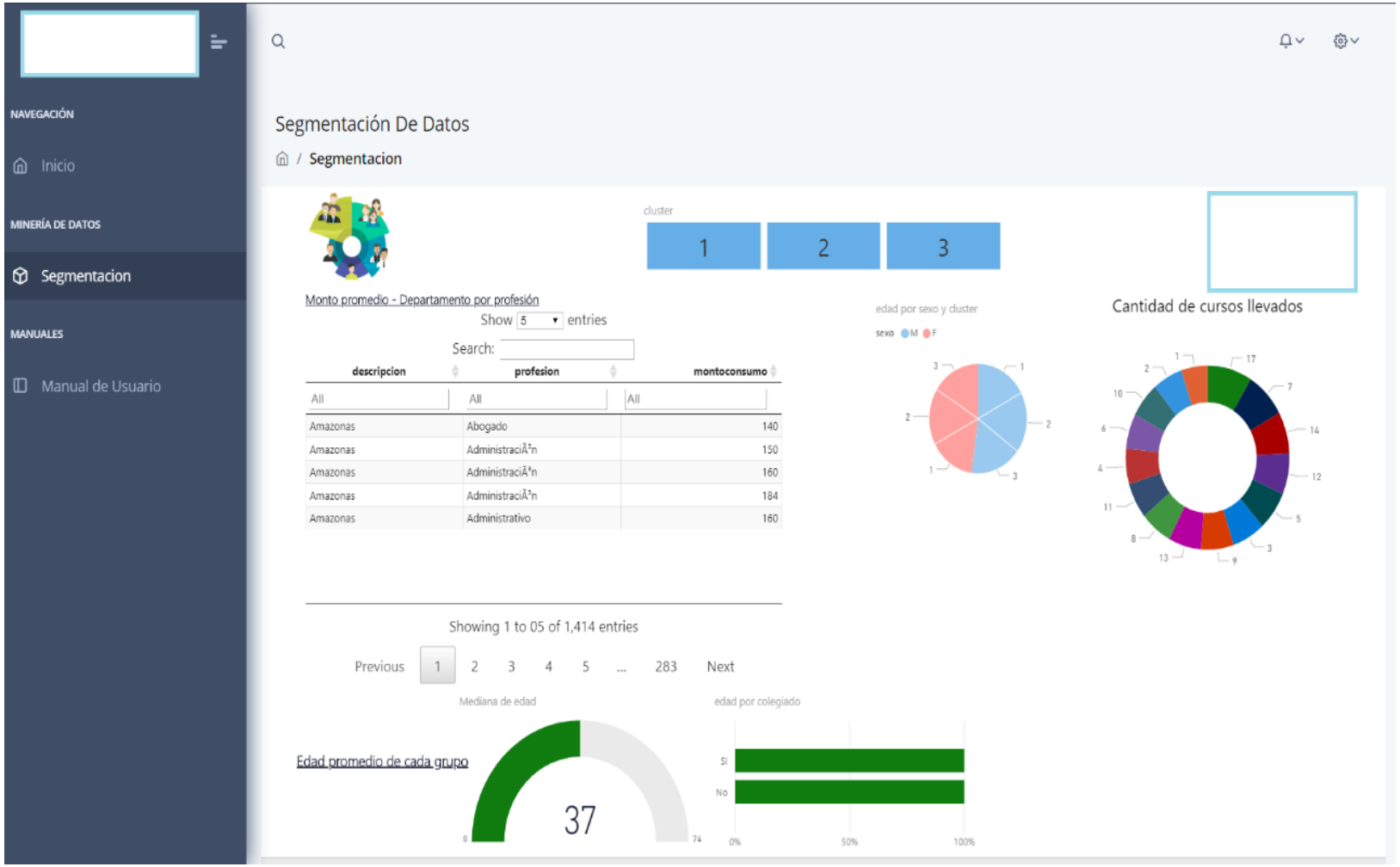


Fig. 43. INTERFAZ DE VISUALIZACIÓN DE RESULTADOS

4.2. En base a los objetivos de la investigación

4.2.1. Realizar un análisis exploratorio a la base de datos

Se desarrollo mediante herramientas de exploración a la base de datos PostgreSQL y en contraste con la base de datos Mysql alojado en la web, para determinar qué calidad de datos se disponía. Ver resultados de iteración n° 02 comprensión de datos en exploración de datos de la metodología CRISP-DM.

4.2.2. Identificar las variables que definan las características del cliente

Se identifico las variables acordes a la disponibilidad de datos y segmentación de clientes por estilo de vida se tomó como referencia el autor Rolando Arellano. Ver resultados de iteración n° 03 en selección de datos, en tabla n° VI de la metodología CRISP-DM.

4.2.3. Identificar técnicas de algoritmos para segmentar clientes de acuerdo con sus características

Se identificaron las técnicas a utilizar por segmentación(clustering) de minería de datos, por ser de tipo descriptivo. Ver resultados de la iteración n° 04 modelado, en selección de técnica de la metodología CRISP-DM.

4.2.4. Analizar los clientes de acuerdo con la segmentación

Conforme a los resultados obtenidos de la aplicación de algoritmos de clustering siendo 3 grupos(clúster) formados por homogeneidad de variables, se analiza sus respectivas variables cualitativas y cuantitativas. Ver iteración n° 05 evaluación de resultados de la metodología CRISP-DM.

4.2.5. Sugerir patrones en base a la solución de minería de datos

Acorde al análisis realizado en la fase n°05 de evaluación de resultados de la metodología CRISP-DM, se establece recomendaciones a la empresa para la toma de decisiones en sus respectivas campañas de publicidad, promoción de cursos(eventos). Ver Recomendaciones.

V. DISCUSIÓN

La presente investigación estuvo enfocada en el problema del negocio sobre el desconocimiento de grupos de clientes, para lo cual se planteó como hipótesis el desarrollo de una solución de minería de datos para la determinación de segmentos de clientes en la empresa de capacitaciones online siendo contrastada con los resultados obtenidos mediante la herramienta de minería de datos como el clustering.

Afirmando la investigación de Jiménez [9] "[…] Se eliminaron los clientes no reales que han sido creados para pruebas del sistema, […] también del cliente", en el proceso de limpieza de datos, menciona la inconsistencia de datos, encontrándose con valores nulos afectando la información a evaluar. Tras realizar un análisis exploratorio a la base de datos PostgreSQL se verificó la calidad de datos disponibles contando con una cantidad de clientes y como resultado se tuvo que realizar una limpieza de datos puesto que, contenían datos nulos y no coherentes con el número de DNI del cliente; por lo que se realizó un proceso de actualización utilizando base de datos externa como de la RENIEC.

Además, se identificaron las características del cliente, contando con un total de 3643 clientes que, a través de un proceso de limpieza de datos, se obtuvieron 2482 clientes seleccionados, los cuales generan ingresos a la empresa. Se detalla en la fase de preparación de datos de la metodología CRIPS-DM ver iteración 4.1.3, considerando los tipos de segmentación por el autor Rolando Arellano. Asimismo, afirmó la investigación realizada por el autor Ramírez [9] presentando variables mixtas para la creación de un modelo crediticio basado en sus transacciones históricas realizadas por los clientes, proporcionando un modelo a evaluar.

Se aplicaron técnicas de segmentación (clustering) para el procesamiento de los datos K-means, método de distancia para la evaluación del clúster óptimo, así mismo; se obtuvieron 3 grupos(clústeres) agrupándose por homogeneidad hallando sus características para determinar el patrón del cliente. Por consiguiente, se afirmó el estudio realizado por el autor Riquelme [5], en el que empleó el método jerárquico y la técnica k-means logrando obtener grupos para su evaluación del modelo creado.

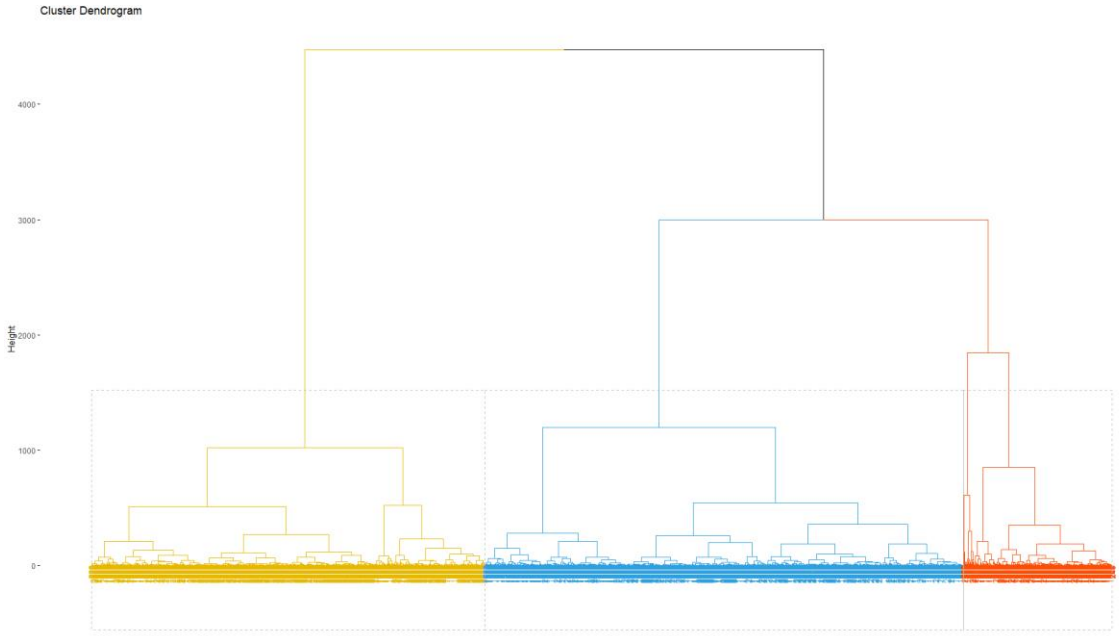


Fig. 44. Aplicando técnica de minería de datos k-means

Dicho lo anterior, se procede a analizar los grupos (clústeres) encontrados determinando sus características, ver iteración n° 05 evaluación de resultados de la metodología CRISP-DM. Por otra parte, se coincide en aspectos encontrados de las características de la investigación de Chamba [7], que expresan las características de los clientes empleando técnicas de minería de datos acorde con ciertos productos de la empresa investigada, utilizando un modelo RFM con el objetivo de determinar características del cliente en base a las similitudes de compra. Esto quiere decir, que el cliente es evaluado según las transacciones realizadas, por lo que en esta investigación se evaluó a las matrículas de cursos realizados por parte del cliente.

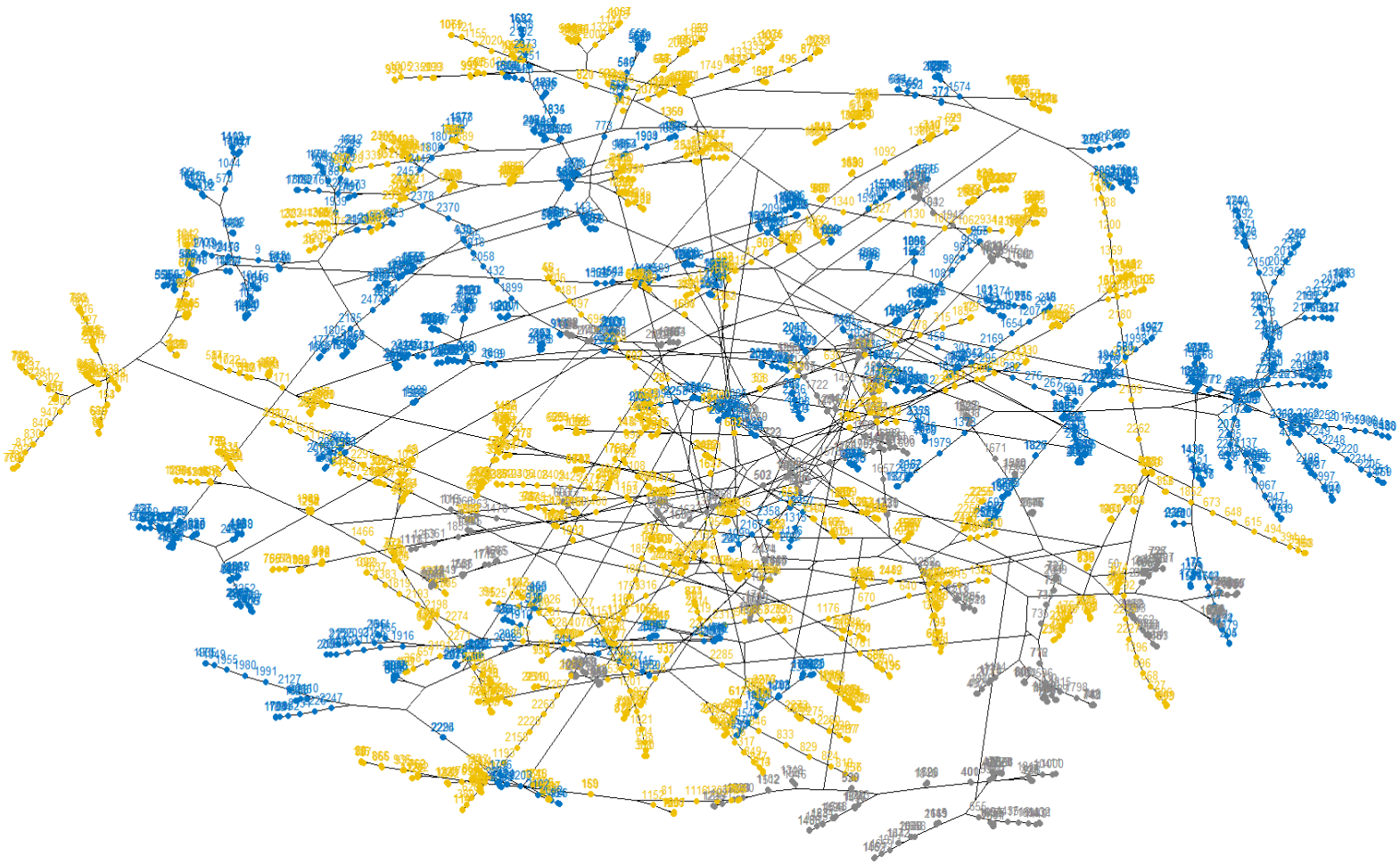


Fig. 45. Gráfica circular de agrupación por técnica de minería de datos

	edad	montoconsumo	cantcursos	cluster
1	40	225	10	1
2	40	225	10	1
3	58	140	2	2
4	32	209	4	1
5	32	209	4	1
6	33	272	3	1
7	33	272	3	1
8	47	180	1	1
9	62	100	11	2
10	39	270	1	1
11	33	315	2	3
12	63	200	2	1
13	30	220	1	1
14	33	300	1	3
15	34	175	1	1
16	35	180	1	1
17	60	200	1	1
18	61	298	3	3
19	27	150	1	2
20	58	289	3	3
21	62	267	3	1
22	42	240	1	1
23	45	340	1	3
24	33	150	1	2
25	35	180	1	1
26	37	155	2	2
27	33	180	1	1
28	28	180	1	1
29	48	260	2	1
30	48	260	2	1
31	37	207	1	1
32	52	320	1	3
33	50	171	1	1
34	55	250	3	1
35	55	250	3	1
36	48	350	1	3
37	64	240	3	1
38	62	180	1	1
39	54	250	1	1
40	46	200	1	1
41	56	150	1	2
42	58	150	1	2

Fig. 46. Agrupación por homogeneidad.

Finalmente, se establece recomendaciones para la empresa de capacitaciones online (ver recomendaciones), con el objetivo de crear estrategias en sus campañas de lanzamientos de cursos (eventos) acorde con los resultados obtenidos por los grupos (clústeres). Asimismo, la investigación por Grández [8] que mediante datos históricos de los clientes se determina cierto patrón de comportamiento que ayuda a ofrecer ciertos productos acorde con las características encontradas.

VI. CONCLUSIONES

Las conclusiones al finalizar este presente trabajo de investigación son las siguientes:

1. Los datos almacenados en la empresa de capacitaciones online en Chiclayo, correspondiente en el periodo 2014-2018, siendo 5 años, lo cual permitió realizar un análisis exploratorio de todas las tablas conformadas y realizar un análisis de contraste a la base de datos alojado en la web. Tomando en cuenta para la verificación de datos disponibles en su respectiva evaluación de segmentación de clientes.

2. En el análisis exploratorio a los datos disponibles, se consideraron variables mixtas (cualitativas y cuantitativas) para este tipo de segmentación de clientes, tomando en cuenta los estilos de vida mencionado por el autor Rolando Arellano. Construyendo un nuevo modelo para su respectiva evaluación.

3. Conforme al nuevo modelo generado, se buscó algoritmos para la segmentación aplicando k-means, k-medoids y de distancia. La comparación de datos aplicando matriz de distancia, matriz de correlación entre clústeres a variables cuantitativas para la obtención del número de clúster. El índice de distancia entre datos como: suma de error, métodos de distancia permitió determinar 3 grupos (clústeres) óptimos para su evaluación.

4. Para el análisis de los resultados se obtuvieron 3 grupos de clientes mediante la aplicación de segmentación de minería datos y se determinó la edad promedio, el monto promedio, profesión con mayor demanda, departamentos con su respectivo monto promedio, el sexo, cantidad de cursos llevados con su respectivo promedio de consumo acorde con los grupos generados por la segmentación de minería de datos.

5. La herramienta R permitió el proceso de minería de datos en la segmentación de clientes, encontrando grupos formados por homogeneidad. Facilitando por sus librerías ya que implementan funciones estadísticas, el manejo de datos con la aplicación de algoritmos, por lo tanto, R es una herramienta importante en el proceso de minería de datos. Este proceso se realizó con una serie de pasos siguiendo la metodología CRIPS-DM.

6. Las recomendaciones encontradas para cada grupo de clientes en base a homogeneidad de datos, le permitirá a la empresa de capacitaciones online Chiclayo elaborar estrategias en promoción de cursos(eventos), descuentos, lanzamientos de cursos acorde al departamento y promedio de monto de consumo.

VII. RECOMENDACIONES

1. Se recomienda efectuar un buen registro de sus clientes de la empresa en el sistema que dispone; puesto que permite la alimentación a la solución de minería de datos; para garantizar la calidad de datos, sin que estos estén nulos.
2. Se sugiere que el registro de sus clientes sea validado automáticamente por su aplicación con la base de datos de la RENIEC, para mantener una calidad de datos, y el proceso sea eficiente.
3. Tener en consideración los futuros cambios en la estructura de la base de datos, para dar un oportuno mantenimiento a la solución de minería de datos.
4. Se debe considerar los resultados obtenidos por la solución de minería de datos, en la cual, se acercará en los envíos de correos y publicidad de acuerdo con los grupos de clientes de la empresa, para evitar problemas en los servicios y problemas legales.
5. Se sugiere que la base de datos se aloje en un servidor para que en un futuro se realice aplicaciones de predicciones de clientes acorde a cursos(eventos) lanzados en su momento aprovechando los datos para generar valor agregado.

VIII. LISTA DE REFERENCIAS

[1] C. Perazo, "La minería de datos desaprovechada", El Cronista, 03 Septiembre 2013. [En línea]. Disponible en: https://www.cronista.com/itbusiness/La-mineria-de-datos-desaprovechada-20130903-0005.html [Accedido: 10 Mayo 2018].

[2] M. A. García, "El Data Mining," UH Noticias el económico, 14 Julio 2017. [En línea]. Disponible en: https://ultimahora.es/noticias/economico/2017/07/14/279653/data-mining.html . [Accedido: 24 Junio 2018].

[3] E. Comercio, "Ventajas de utilizar la tecnología Data Lake en las empresas," 29 Noviembre 2017. [En línea]. Disponible en: https://elcomercio.pe/especial/zona-ejecutiva/negocios/5-ventajas-utilizar-tecnologia-data-lake-empresas-noticia-1992761 . [Accedido: 10 Mayo 2018].

[4] D. Gestión, "Empresas peruanas están cambiando el enfoque en cómo miden su información," 10 Junio 2014. [En línea]. Disponible en: https://gestion.pe/tecnologia/empresas-peruanas-cambiando-enfoque-miden-informacion-62569. [Accedido: 10 Julio 2018].

[5] G. A. Riquelme A, "Desarrollo de estrategias de fidelización mediante análisis multivariante de clúster de los mejores clientes activos de comercial Kaufmann S.A.", Tesis pregrado, Universidad Austral de Chile. 2014. [En línea]. Disponible en: http://cybertesis.uach.cl/tesis/uach/2014/bpmfcir594d/doc/bpmfcir594d.pdf

[6] S. A. Orellana M., "Segmentación de perfiles de consumo utilizando variables latentes para el mercado de clientes regulados de Chilectra.", Trabajo fin de grado, Universidad De Chile, Chile. 2016. [En línea]. Disponible en: http://repositorio.uchile.cl/bitstream/handle/2250/139903/Segmentacion-de-perfiles-de-consumo-utilizando-variables-latentes-para-el-mercado-de-clientes.pdf?sequence=1&isAllowed=y . [Accedido: 25-Agos-2018].

[7] S. F. Chamba J., "Minería de datos para segmentación de clientes en la empresa tecnológica Master PC", Tesis pregrado, Universidad Nacional de Loja, Ecuador. 2015. [En línea]. Disponible en:

https://dspace.unl.edu.ec/jspui/bitstream/123456789/10462/1/Chamba%20Jim%C3%A9nez,%20Sairy%20Fernanda.pdf .[Accedido: 30-May-2019]

[8] M. A. Grández M., "Aplicación de minería de datos para determinar patrones de consumo futuro de clientes de una distribuidora de suplementos nutricionales.", Tesis pregrado, Universidad San Ignacio de Loyola, Lima. 2017. [En línea]. Disponible en: http://repositorio.usil.edu.pe/handle/USIL/2763 .[Accedido: 30-May-2019]

[9] M. L. Ramírez S., "Identificación de perfiles de clientes crediticios aplicando técnicas de segmentación y regresión logística multinomial.", Tesis pregrado, Universidad Nacional Agraria La Molina, Lima. 2014. [En línea]. Disponible en: http://repositorio.lamolina.edu.pe/handle/UNALM/2280. [Accedido: 30-May-2019]

[10] M. A. Calderón P. y P. A. K. G. Vega A., "Sistema de información para la recomendación de productos basado en patrones de comportamiento y localización visual de una canasta de productos en un supermercado.", Tesis pregrado, Pontificia Universidad Católica del Perú, Lima. 2015. [En línea]. Disponible en: http://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/6040 . [Accedido: 30- May-2019].

[11] M. J. Cotrina G y W. H. Siancas G., "Aplicación de minería de datos usando reglas de asociación para el análisis de ventas en una empresa de Retail.", Tesis pregrado, Universidad Señor de Sipán, Pimentel. 2016. [En línea] Disponible en: http://repositorio.uss.edu.pe/handle/uss/300 . [Accedido: 30-May-2019].

[12] L. H. Jiménez B., "Aplicación de un sistema de alerta temprana basada en la minería de datos para identificar patrones delictivos en la ciudad de Chiclayo.", Tesis pregrado, Universidad Católica Santo Toribio de Mogrovejo, 2015. [En línea]. Disponible en: http://tesis.usat.edu.pe/bitstream/20.500.12423/543/1/TL_Jimenez_Berrios_LeslyHaymet.pdf . [Accedido: 30-May-2019]

[13] E. H. Cubas, "Implementación de un modelo de data mining para optimizar el proceso de toma de decisiones de la gestión académica en el instituto

superior de administración gerencial y sistemas – ISAG.", Universidad Señor de Sipán, Pimentel, 2016.

[14] R. Fernández V, "Segmentación de mercados y clasificación", en Segmentación de Mercados, 2da. Ed. México, International Thomson, 2002.

[15] L. C. Schiffman, "¿Qué es segmentación de mercado?", en Comportamiento del consumidor, 8va. Ed., México, Prentice Hall, 2005. [En línea]. Disponible en: https://bit.ly/2JQB11w [Accedido: 30-May-2019]

[16] R. Arellano C, "La segmentación de mercados y el posicionamiento", en Marketing enfoque América Latina, 1ra. Ed. México, McGraw-Hill Interamericana, 2000, pp. 481-484.

[17] J. A. Lara T, Minería de datos, España: Ed. Centro de Estudios Financieros, 2014.

[18] "El Algoritmo K-means aplicado a clasificación y procesamiento de imágenes", Universidad de Oviedo, 2015. [En línea]. Disponible en: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html#El-algoritmo-k-means-aplicado-a-clasificaci%C3%B3n-y-procesamiento-de-im%C3%A1genes .[Accedido: 6-jun-2019]

[19] "Métodos jerárquicos de análisis Clúster", Universidad de Granada, España. [En línea]. Disponible en: https://www.ugr.es/~gallardo/pdf/cluster-3.pdf . [Accedido 8- Jun-2019].

[20] "Data Science vs Data Analytics: parecidos, pero no iguales", *Telefónica*, 2017. [En línea], Disponible en: https://empresas.blogthinkbig.com/data-science-vs-data-analitycs/ . [Accedido: 6-jul-2019]

[21] M. Alvarez, "Qué es Python", Desarrollo web, 2003. [En línea]. Disponible en: https://desarrolloweb.com/articulos/1325.php .[Accedido: 7-jul-2019]

[22] "Introducción a R", R Development Core Team, 2000. [En línea]. Disponible en: https://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf .[Accedido: 6-jul-2019]

[23] R. Chalmeta y C. Oscar, "Methodology for the extraction of Enterprise Knowledge from Data", Scielo Analytics, Vol. 17, n° 2-2006, pp. 81-88, 2006.

[24] J. Hernandez Orallo, J. Ramirez Quintana y C. Ferrari Ramirez, Introducción a la minería de datos, Madrid: PEARSON EDUCACIÓN S.A, 2004.

[25] UNAL, “Metodología SEMMA”, 2016, [En línea]. Disponible en: http://disi.unal.edu.co/~eleonguz/cursos/md/presentaciones/Sesion5_Metodologias.pdf. [Accedido: 30- jun-2018].

[26] R.d.U.e.C.d.l Computación, “Estudio comparativo de metodologías para minería de datos”, 2011. [En línea]. Disponible en: http://sedici.unlp.edu.ar/handle/10915/20034 .[Accedido: 30-jun-2018].

[27] IBM, Manual CRISP-DM de IBM SPSS Modeler, Estados Unidos: IBM Corporation, 2012. [En línea] Disponible en: http://www.crips-dm.org .[Accedido: 30-Jun-2018].

[28] R. H. Sampieri, C. F. Collado y P. B. Lucio, *Metodología de la investigación*. 4ta ed. México: McGraw-Hill, 2006.

[29] Real Academia Española. [En línea] Disponible en: https://dle.rae.es/?id=5ASmP2Z . [Accedido: 10 -Jun-2019].

IX. ANEXOS

ANEXO N° 01

INSTRUMENTOS DE RECOLECCIÓN DE DATOS

ENTREVISTA AL DUEÑO DE LA EMPRESA DE CAPACITACIONES ONLINE

Se realiza la entrevista con el objetivo de conocer el negocio e identificar el problema a solucionar.

Fecha: / /

Nombre del entrevistado:

Cargo en la empresa:

Preguntas para responder:

1. ¿Qué servicios brinda la empresa?
2. ¿De qué manera brindan su servicio de educación?
3. ¿Qué tipo de cursos brinda?
4. ¿a qué personas va dirigido estos cursos?
5. ¿Qué medios de publicidad utiliza para la promoción de dichos cursos?
6. ¿Qué técnicas utiliza para lanzar dichos cursos(eventos)?
7. ¿Cuánto es el minino de clientes matriculados debe haber por curso?
8. ¿Qué herramientas utiliza para captar clientes?
9. ¿Realiza un benchmarking?
10. ¿Qué debilidades presenta en la captación de clientes?
11. ¿Qué plataformas utilizan los clientes matriculados?
12. ¿Cuál es el tiempo de cada curso (evento)?
13. ¿Cuál es su valor agregado en diferencia de la competencia?
14. ¿Qué segmento del mercado es su objetivo?
15. ¿Cuáles son los criterios de selección de los ponentes?
16. ¿Pueden identificar sus clientes potenciales?

OBSERVACIÓN

Se observaron los correos de distintas cuentas de la empresa en las cuales son enviados masivamente a cierto curso (evento), mostrando el estado por combinación de correos si estos son abiertos, no abiertos o Spam.

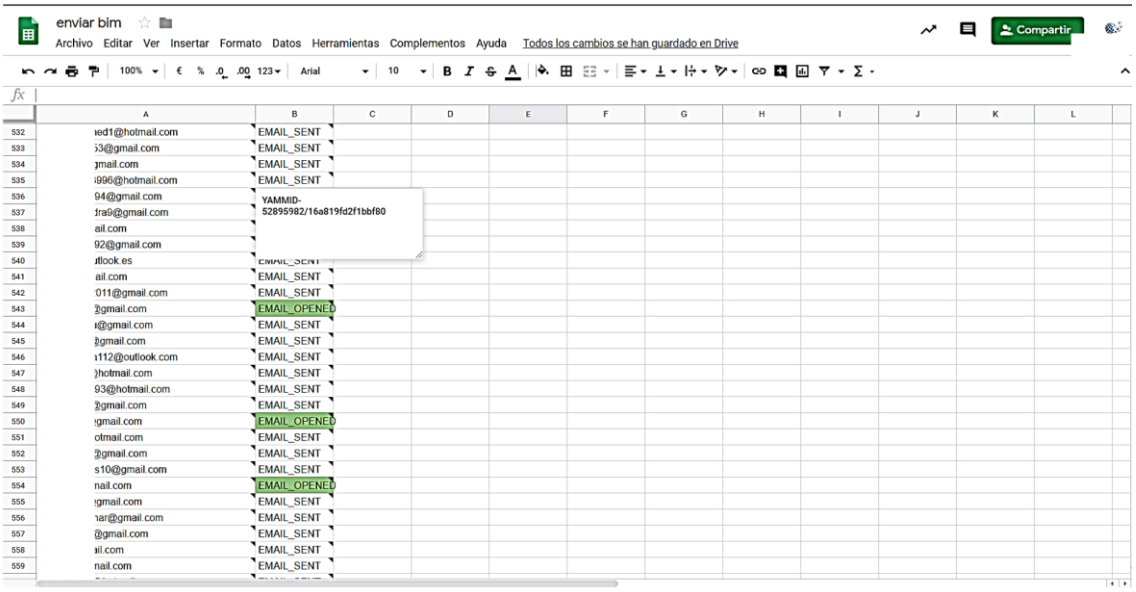


Fig. 47. CORREOS ENVIADOS MASIVAMENTE

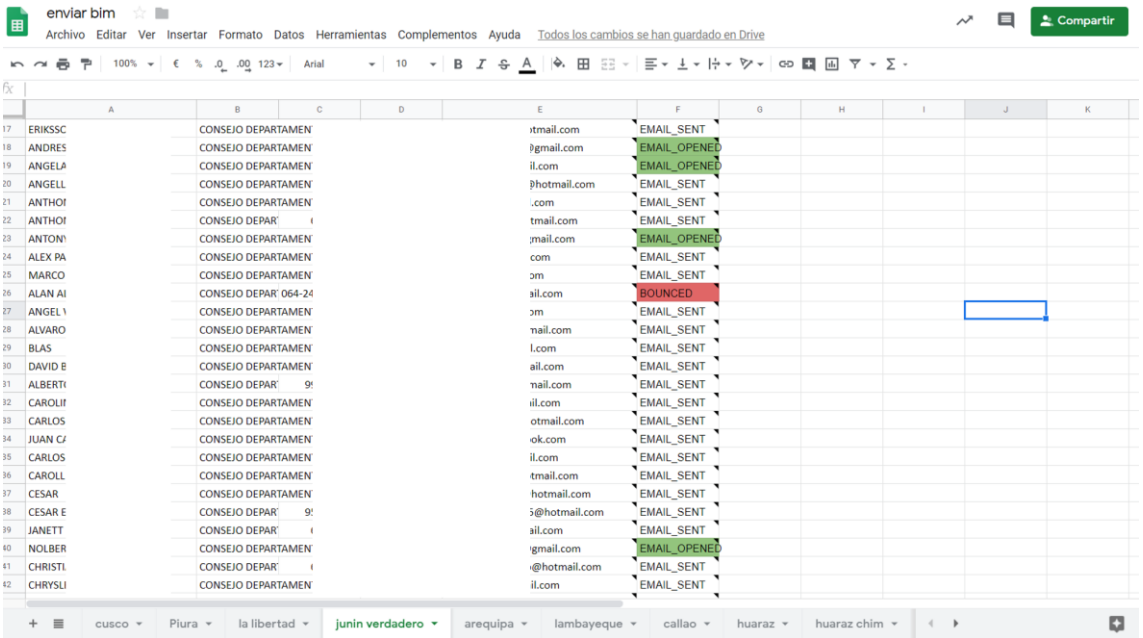


Fig. 48. CORREOS ENVIADOS, ENTRE 40% NO ABIERTOS

Clientes matriculados acorde a un curso, con publicidad enviada.

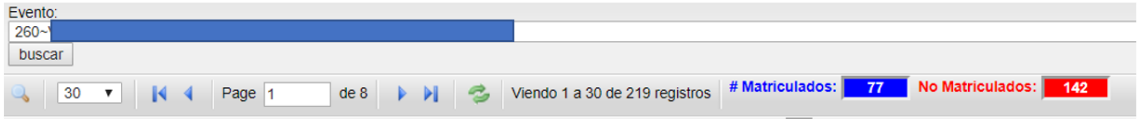


Fig. 49. CLIENTES INTERESADOS Y MATRÍCULADOS

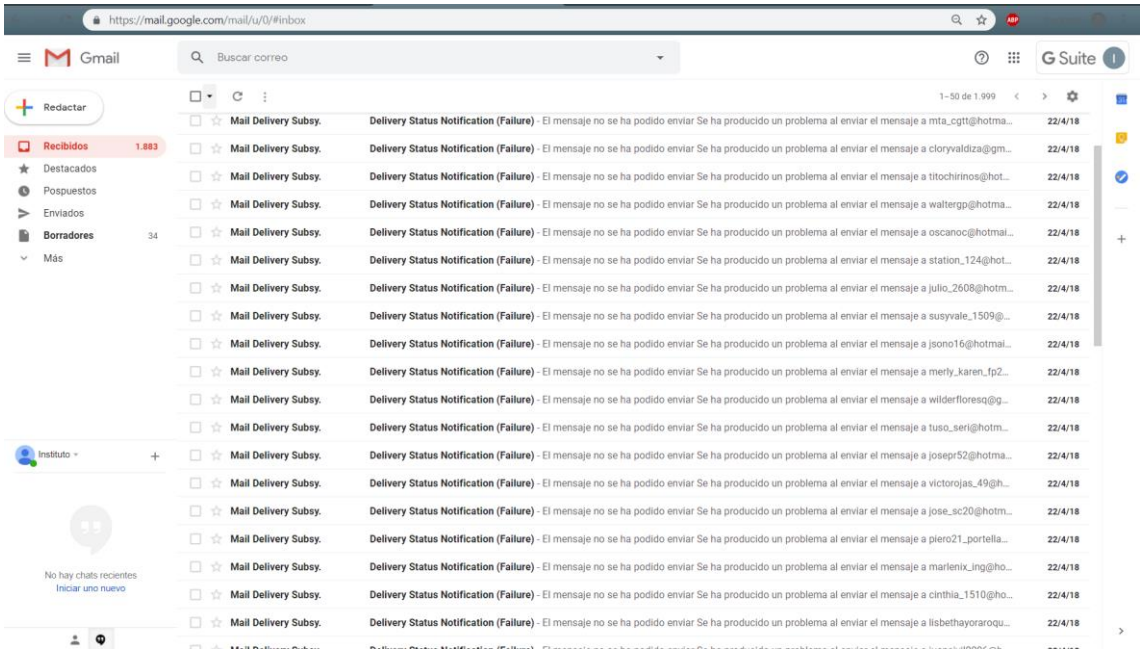


Fig. 50. CORREOS ENVIADOS QUE GENERAN SPAM

ANEXO N° 02

ANÁLISIS DE RIESGOS

1. **Datos generales**
   - **Tesista** **:** Dany Yesenia Gelacio Mendoza
   - **Fecha inicial** **:** 20 de agosto de 2018
   - **Fecha final** **:** 12 de julio de 2019
2. **Alcance del proyecto**

   Se desarrolla en la empresa de capacitaciones online Chiclayo para la segmentación de clientes, con la finalidad de buscar segmentos de clientes haciendo uso de minería de datos modelo descriptivo.
3. **Interesados (Stakeholders)**

   Durante el desarrollo de la presente tesis se ha identificado a los siguientes interesados:
   - **Internos**

TABLA IX
INTERESADOS INTERNOS

Interesado	Participación
Gerente General	Representante de la empresa de capacitaciones online.
Accionista	Interesado en el crecimiento de la empresa en el rubro de educación vía online.

4. Etapa de desarrollo

Para el desarrollo de la solución de la presente tesis se ha tomado en cuenta las etapas de la Metodología CRISP-DM, en la cual; se consideran los riesgos y sus respectivas acciones de contingencia.

- **Riesgos y Contingencia**

  Siendo de referencia PMBOOK, que el intervalo es lo siguiente: [0-20] es muy bajo, [21-40] es bajo, [41 – 60] es medio, [61-79] es alto y [81-100] es muy alto.

  Entre los riesgos identificados en esta etapa se mencionan:

TABLA X
RIESGOS IDENTIFICADOS ETAPA I

Código del riesgo	Descripción del riego	Fase afectada	Causa Raíz	Acciones de contingencia	Estimación Probabilidad	Probabilidad por impacto	Prioridad	Nivel de riesgo
RE1 – 001	Pérdida de información	Comprensión del negocio	Pérdida de la base de datos	Realizar copias de seguridad en repositorios en línea.	60	80	70	ALTO
RE1-002	Información no brindada por la empresa	Comprensión del negocio	Información no brindada	Realizar un acta de aprobación	60	80	70	
RE1-003	Información brindada inoportuna	Comprensión del negocio	Que la empresa no brinde información	Realizar entrevistas con el accionista	70	70	70	
RE1-004	Inconsistencia de información	Comprensión del negocio	Información escasa	Realizar un análisis del negocio y procesos	20	40	30	Bajo

TABLA XI
RIESGOS IDENTIFICADOS ETAPA II

Código del riesgo	Descripción del riego	Fase afectada	Causa Raíz	Acciones de contingencia	Estimación Probabilidad	Probabilidad por impacto	Prioridad	Nivel de riesgo
RE2-001	Datos faltantes en la información	Comprensión de datos	Datos faltantes en la base de datos	Realizar un seguimiento de dicha información	60	80	70	Alto
RE2-002	Datos incomprensibles	Comprensión de datos	Que los datos se presenten de manera incoherente	Realizar una exploración de los datos con la aplicación de escritorio	30	50	40	Medio
RE2-003	Problemas de hardware	Comprensión de datos	Incidentes de hardware, como fallas en disco duro.	Realizar mantenimiento en cierto periodo	20	80	50	

TABLA XII
RIESGOS IDENTIFICADOS EN ETAPA III

Código del riesgo	Descripción del riego	Fase afectada	Causa Raíz	Acciones de contingencia	Estimación Probabilidad	Probabilidad por impacto	Prioridad	Nivel de riesgo
RE3-001	Datos no disponibles	Preparación de datos	Datos incoherentes	Establecer comparaciones con base de datos externas	40	80	60	Medio
RE3-002	Base de datos externas no disponibles	Preparación de datos	Datos atípicos	Realizar consultas a base externas para la limpieza de datos	50	80	65	
RE3-003	Orígenes de datos	Preparación de datos	Alojamiento de datos	Realizar un análisis de registro y encontrar la base de datos adecuada	20	40	30	Bajo

TABLA XIII
RIESGOS IDENTIFICADOS ETAPA IV

Código del riesgo	Descripción del riego	Fase afectada	Causa Raíz	Acciones de contingencia	Estimación Probabilidad	Probabilidad por impacto	Prioridad	Nivel de riesgo
RE4-001	Retrasos en las fases de ejecución	Modelamiento	Desconocimientos de técnicas de segmentación	Establecer tiempos y buscar a expertos en técnicas	40	80	60	Medio
RE4-002	Selección de técnicas	Modelamiento	Selección de algoritmo	Buscar expertos y leer la documentación	40	90	65	
RE4-003	Cambios continuos de datos y modelo	Modelamiento	Nuevos requerimientos	Realizar un análisis del negocio, establecer criterios de segmentación	20	40	30	Bajo

ANEXO N° 03

ANÁLISIS DE COSTO

PARTE PRESUP.	DESCRIPCIÓN	CANT.	UNIDAD DE MEDIDA	PRECIO UNITARIO	PRECIO TOTAL	SUBTOTAL
				(S/.)	(S/.)	(S/.)
2.3.2 2.2	SERVICIOS DE TELEFONIA E INTERNET					S/. 360,00
2.3.2 2.2 1	Servicio de Telefonía Móvil	12	GLOBAL	10,00	120,00	
2.3.2 2.2 3	Servicio de Internet	12	GLOBAL	20,00	240,00	
2.3.1 5.1	MATERIALES Y UTILES DE OFICINA					S/. 100,00
2.3.1 5.1 1	Set de tintas para impresora	1	UNIDAD	50,00	50,00	
2.3.1 5.1 2	Papelería en general, útiles y materiales de oficina	1	GLOBAL	50,00	50,00	
2.3.1 1.1	ALIMENTOS Y BEBIDAS					S/. 275,00
2.3.1 1.1 1	Almuerzos	20	UNIDAD	10,00	200,00	
	Bebidas	50	UNIDAD	1,50	75,00	
2.1.2 1.2	SERVICIOS					S/. 750,00
2.1.2 1.2 1	Movilidad para traslado	150	UNIDAD	5,00	750,00	
2.6.3 2.3	ADQUISICION DE EQUIPOS INFORMATICOS Y DE COMUNICACIONES					S/. 1173,00
	USB 16 GB	1	UNIDAD	25,00	25,00	
	Impresora	1	GLOBAL	320, 00	320, 00	
	Laptop (uso)	1	GLOBAL	69, 00	828,00	
TOTAL, PRESUPUESTO DEL PROUCTO ACREDITABLE						S/. 2658,00

PARTE PRESUP.	DESCRIPCIÓN	CANT.	UNIDAD DE MEDIDA	PRECIO UNITARIO	PRECIO TOTAL	SUBTOTAL
				(S/.)	(S/.)	(S/.)
2.6.6 1.3	ACTIVOS INTANGIBLES					S/. 971.4
2.6.6 1.3 2	Hosting	1	UNIDAD	47.95	575.4	
	Power BI PRO($9.99)	1	MENSUAL	33	396	
TOTAL PRESUPUESTO DEL PROUCTO TECNOLÓGICO						S/. 971.4

Total, del presupuesto:

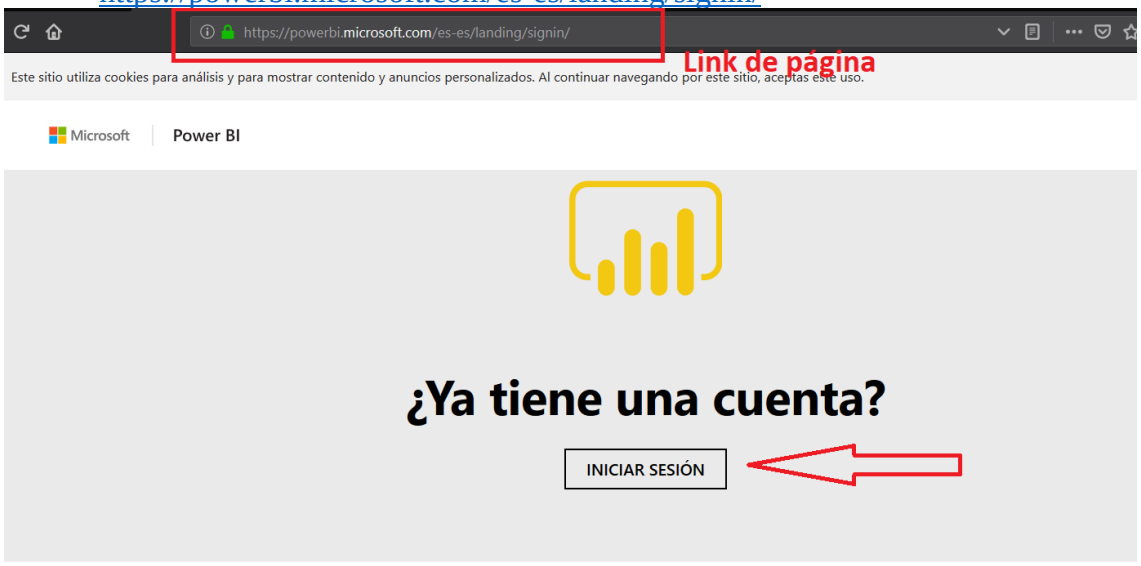
ITEM	DESCRIPCIÓN PRESUPUESTO	SUBTOTAL
		(S/.)
1	TOTAL PRESUPUESTO DEL PROUCTO ACREDITABLE	2658,00
2	TOTAL PRESUPUESTO TECNOLÓGICO	971,4
TOTAL PRESUPUESTO		S/. 3629,4

ANEXO N° 04

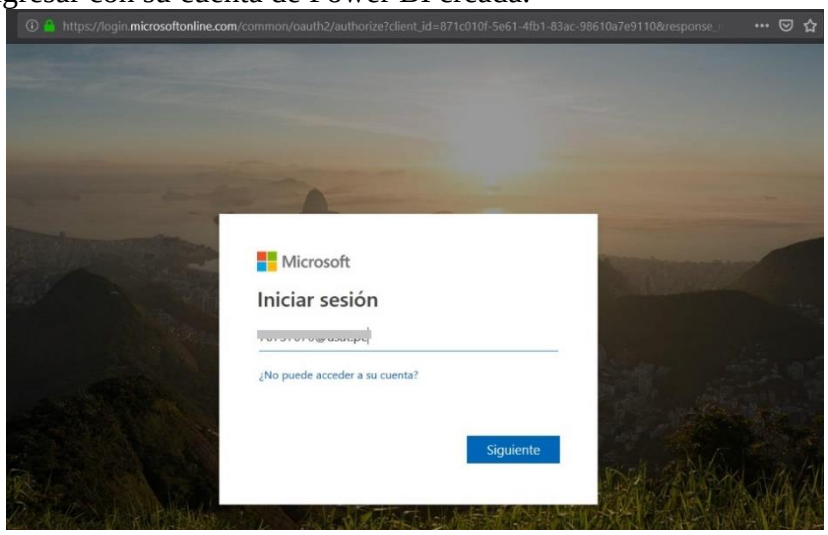
MANUAL DE USUARIO

1. REGISTRO DE CUENTA EN POWER BI

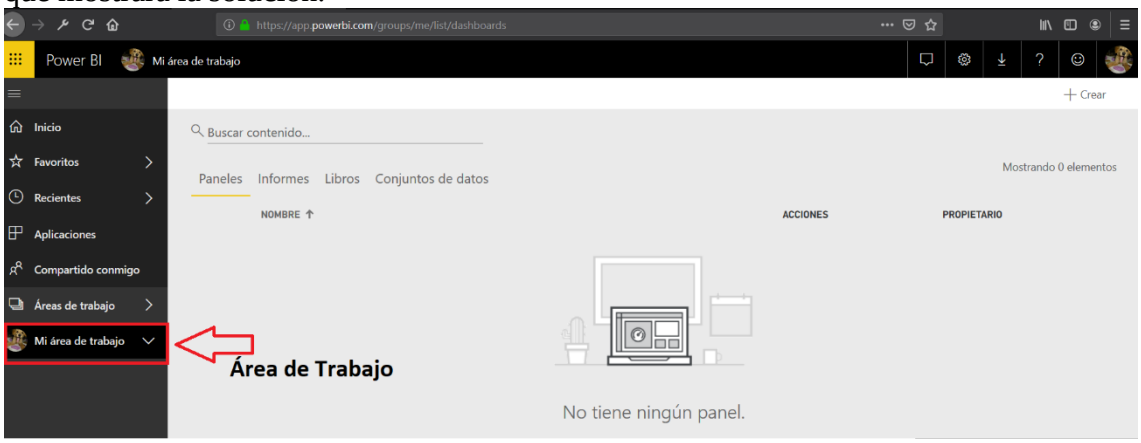
1.1 Ingresar a la página de Power BI en el siguiente enlace: https://powerbi.microsoft.com/es-es/landing/signin/

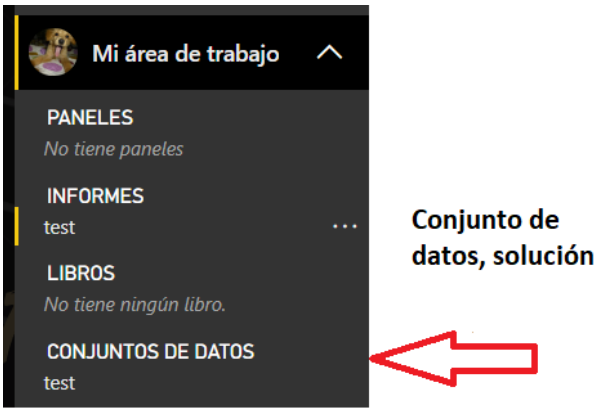


1.2. Ingresar con su cuenta de Power BI creada.



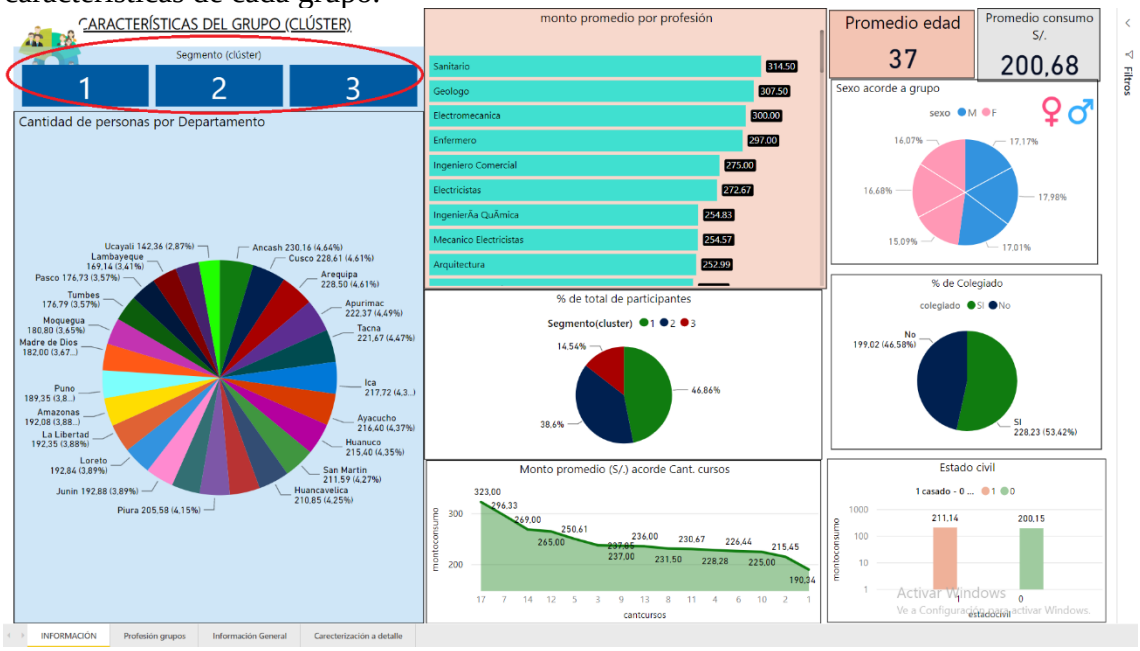
1.3. Se mostrará el panel de Power BI, nos ubicaremos en el área de trabajo; en la que mostrará la solución.



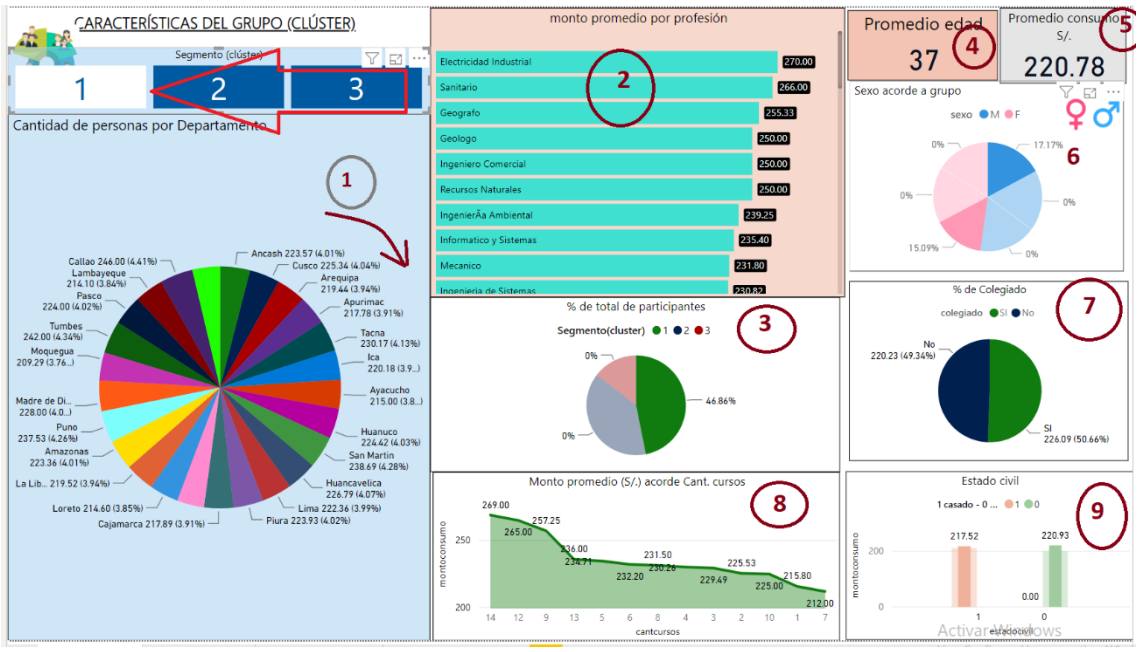


2. VIZUALIZACIÓN DE CARACTERÍSTICAS DE SEGMENTOS DE CLIENTES

2.1 Visualización de n° de segmentos de clientes generados por la solución de minería de datos. Seleccionar un determinado segmento y este actualizará las características de cada grupo.

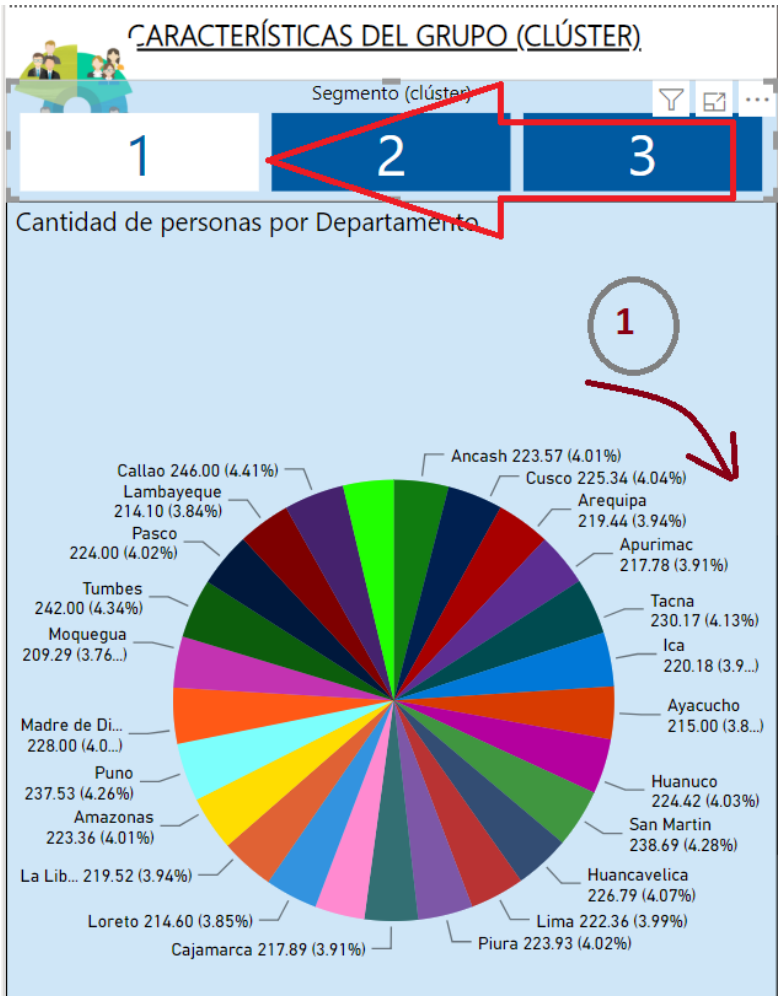


2.2 Lectura de características de acuerdo con un grupo (clúster) seleccionado.

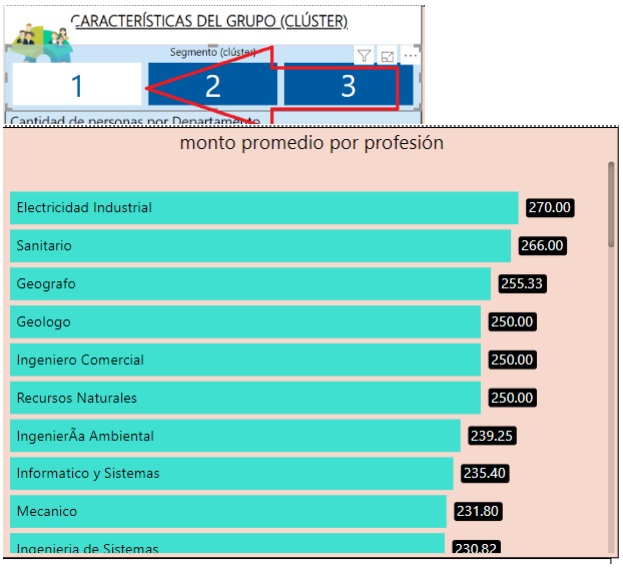


2.2.1 Lectura de características de la información mostrada.

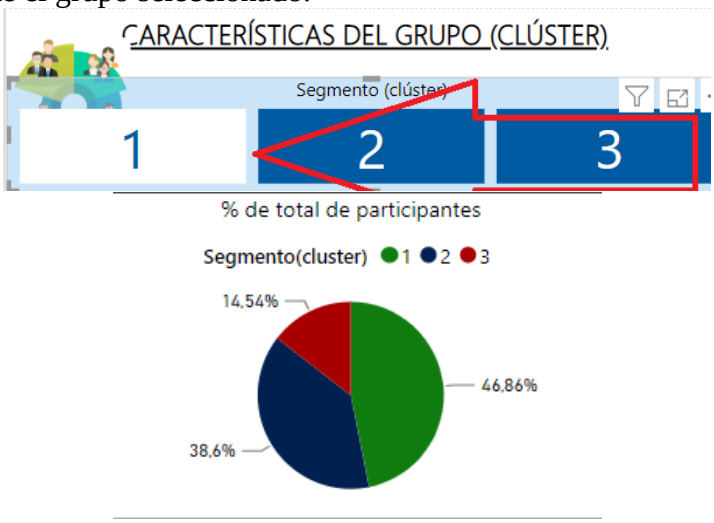
1. Grafica circular, visualización del porcentaje de clientes de acuerdo con el departamento inscrito. Su lectura es sentido horario, puesto que muestra el mayor porcentaje que acogida por departamento y su monto promedio de consumo.



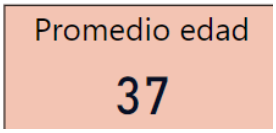
2. Visualización de profesiones con su respectivo monto promedio



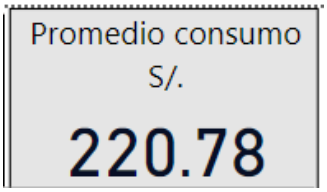
3. Gráfica circular, mostrando el porcentaje del total de participantes que represente el grupo seleccionado.



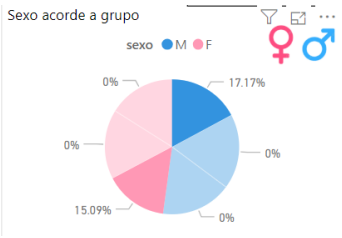
4. Visualización de edad promedio de clientes de acuerdo con el grupo seleccionado.



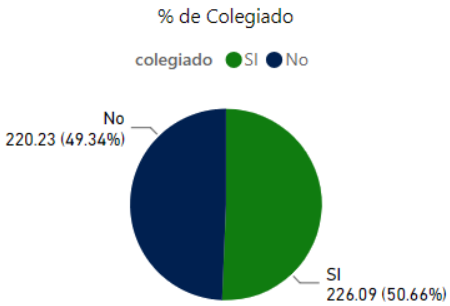
5. Monto promedio de consumo de clientes acorde al segmento seleccionado.



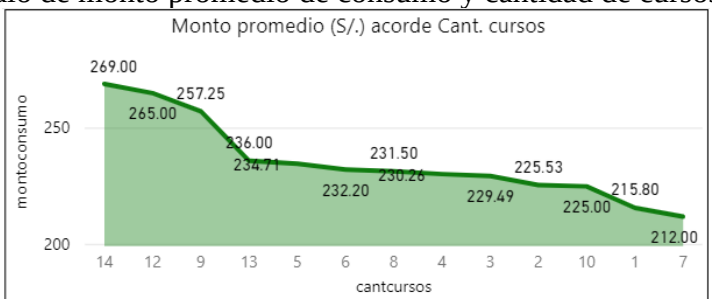
6. Gráfica circular de porcentaje de participantes según su sexo acorde con el grupo seleccionado.



7. Gráfica circular mostrando el porcentaje de participantes que si son colegiados.



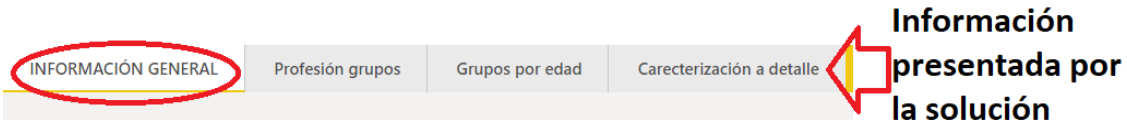
8. Intervalo de monto promedio de consumo y cantidad de cursos.



9. Grafica del estado civil, acorde con su monto promedio de consumo de segmento seleccionado.

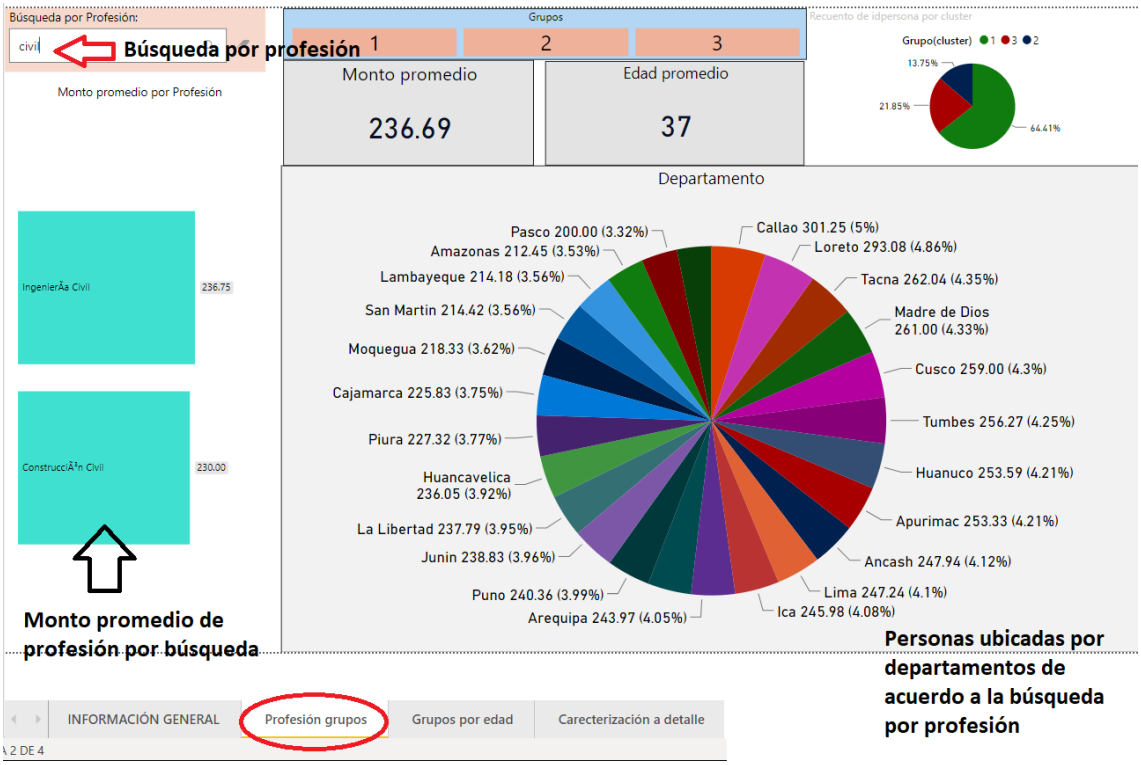
3. PANEL DE NAVEGACIÓN

Visualización de información a detalle

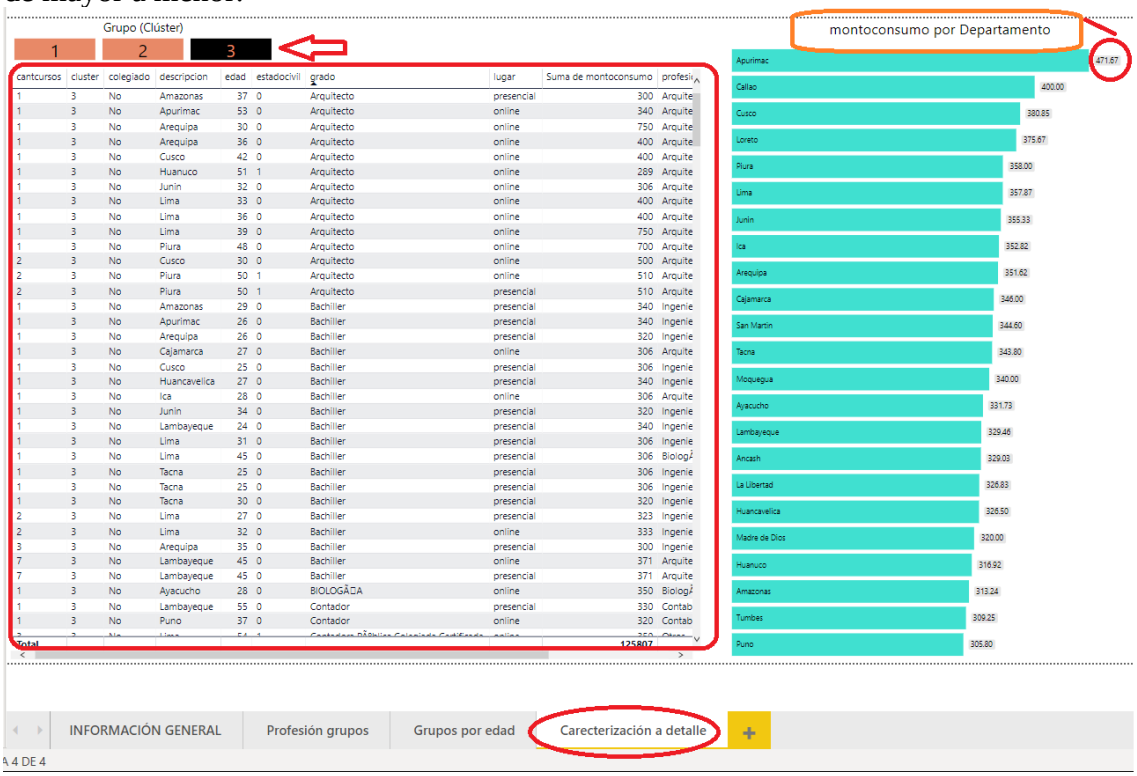


3.1 DETALLE DE INFORMACIÓN

Búsqueda por profesión, de acuerdo a los resultados muestra en que grupos de segmentación se encuentra, monto promedio de consumo de personas con aquella profesión buscada y el departamento.



En la pestaña Caracterización a detalle muestra el grupo de participantes de acuerdo al grupo seleccionado y además el monto promedio de consumo por departamento ordenado de mayor a menor.



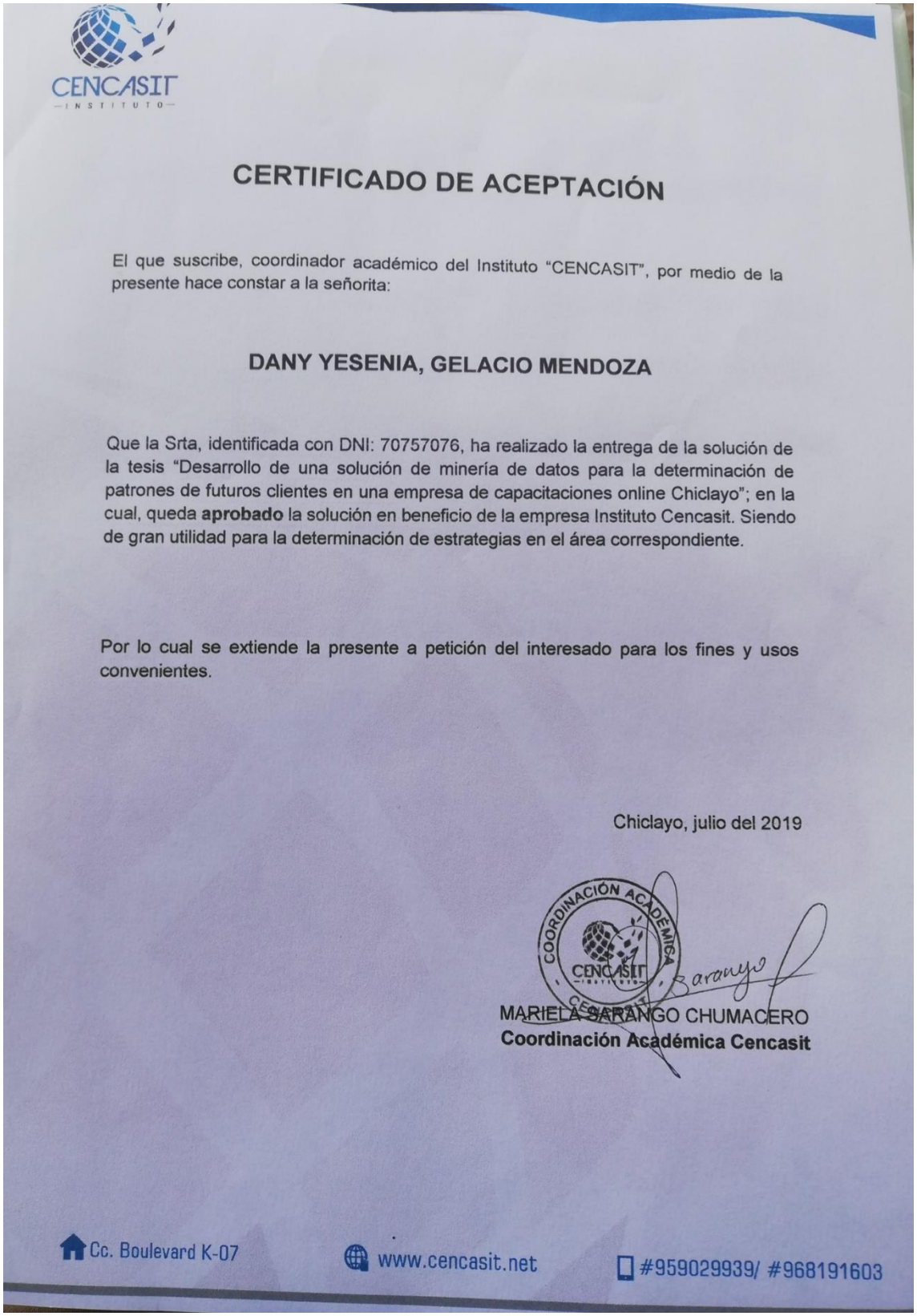
ANEXO N° 05

PLAN DE PROYECTO

Actividades	2018					2019						
Proyecto ejecución informe	ago	set	oct	nov	dic	ene	feb	mar	abr	may	jun	jul
Análisis del Problema												
Determinar los objetivos del negocio												
Evaluar la situación												
Análisis de los datos												
Recopilar los datos iniciales												
Descripción de los datos												
Revisar los datos												
Verificar la calidad de datos												
Preparación de los datos												
Selección de datos												
Limpieza de datos												
Construcción de los datos												
Integración de los datos												
Modelado												
Selección de técnica de modelado												
Construcción del modelo de pruebas												
Implementación del modelo												
Evaluación del modelo												
Evaluación												
Evaluación de los resultados												
Revisión de proceso												
Transferencia o Explotación												
Plan de monitoreo y mantenimiento												
Producción del reporte final												
Revisión del proyecto												

ANEXO N° 06

DOCUMENTO DE ACEPTACIÓN POR PARTE DE LA EMPRESA DE CAPACITACIONES ONLINE



CENCASIT
INSTITUTO

CERTIFICADO DE ACEPTACIÓN

El que suscribe, coordinador académico del Instituto "CENCASIT", por medio de la presente hace constar a la señorita:

DANY YESENIA, GELACIO MENDOZA

Que la Srta, identificada con DNI: 70757076, ha realizado la entrega de la solución de la tesis "Desarrollo de una solución de minería de datos para la determinación de patrones de futuros clientes en una empresa de capacitaciones online Chiclayo"; en la cual, queda **aprobado** la solución en beneficio de la empresa Instituto Cencasit. Siendo de gran utilidad para la determinación de estrategias en el área correspondiente.

Por lo cual se extiende la presente a petición del interesado para los fines y usos convenientes.

Chiclayo, julio del 2019

COORDINACIÓN ACADÉMICA
CENCASIT

MARIELA SARANGO CHUMACERO
Coordinación Académica Cencasit

Cc. Boulevard K-07 www.cencasit.net #959029939/ #968191603

ANEXO N° 07

INFORME DEL TURNITIN