Pontificia Universidad Católica del Ecuador
Facultad De Ingeniería

TEMA:

Minería de Datos para Segmentación de clientes en el Laboratorio Clínico particular Pura Vida

AUTOR:

Miguel Dimitri Ortiz Navarrete

TUTOR:

Jhonny Vladimir Pincay Nieves

TRABAJO PREVIA A LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN SISTEMAS DE INFORMACIÓN MENCIÓN DATA SCIENCE

Quito, Enero – 2023

DEDICATORIA

A Dios, a mi esposa, mi hijo y mi hija, por todo el amor y dedicación entregado a este noble y bendecido proyecto llamado FAMILIA.

AGRADECIMIENTO

A Dios, por proporcionarme el don de la perseverancia y la disciplina en mi vida.

Al Dr. Jhonny Pincay, por su conocimiento y apoyo incondicional en el desarrollo de este proyecto.

Al Laboratorio Clínico Pura Vida en la persona de sus propietarias Lcda. Sofía Cadena y Mtr. Dora Rosero, por su apertura y colaboración en facilitar los datos para el desarrollo de este proyecto.

RESUMEN

La finalidad de este proyecto es aplicar técnicas de minería de datos que permitan obtener información que apoye a la toma de decisiones en el Laboratorio Clínico Pura Vida. El Laboratorio está ubicado en Quito, sector la Kennedy, dispone de infraestructura tecnológica para la elaboración de 203 pruebas de laboratorio clínico. El registro de la información tanto financiera como clínica se realiza en dos sistemas de información diferentes conocidos como CLINICAL Lab y SisGem respectivamente. La información de los sistemas se obtiene en formato xlsx o csv a través de los reportes que cada uno de estos sistemas posee; con esta información se inició el proceso ETL y CRISP-DM para la preparación de la data. Los datos del laboratorio son en su mayoría categóricos y de una alta dimensionalidad; en tal virtud se debe considerar modelos de agrupamiento por densidad. Se aplicó para el agrupamiento los modelos G-means, K-modes, DBSCAN. Los grupos obtenidos con estos modelos no son claros debido a la dimensionalidad de los datos y su estructura. Se aplicaron también algoritmos de asociación para identificar reglas de asociación que determinen el comportamiento del paciente. Los algoritmos de asociación implementados son A PRIORI y ECLAT, los dos algoritmos presentaron reglas de asociación a considerar por el negocio. Los algoritmos fueron desarrollados en código Python y se utilizó también BigML para identificar de forma ágil las características y patrones en la data. El análisis de esta información generó alternativas futuras para proyectos de investigación que permitan desarrollar sistemas de información que apoyen al diagnóstico y decisión de los profesionales de la salud.

ÍNDICE

1. Capítulo I. Introducción 1
   1.1 Generalidades 1
   1.2. Planteamiento del problema 2
   1.3 Objetivos 4
      1.1.1 Objetivo General 4
      1.1.2 Objetivos Específicos 4
   1.4 Alcance 4
2. Capítulo II. Revisión literaria 5
   2.1 Minería de Datos 5
      2.1.1 Técnicas de minería de datos 5
      2.1.2 Algoritmos de Clustering 7
   2.2 Segmentación de Clientes 11
      2.2.1 Criterios para la segmentación de clientes 11
      2.2.2 Técnicas de segmentación de clientes 11
      2.2.3 Análisis RFM 12
   2.3 Metodología de Minería de Datos 13
      2.3.1 SEMMA 13
      2.3.2 CRISP-DM 14
3. Capítulo III. Marco metodológico 18
   3.1 Materiales 18
   3.2 Métodos 18
      3.2.1 Metodología 18
      3.2.2 Métodos y Técnicas 19
4. Capítulo IV. Resultados 21
   4.1 Análisis del estado actual del Laboratorio Clínico Pura Vida 21
      4.1.1 Comprensión del Negocio 21
      4.1.2 Organización del laboratorio 21
      4.1.3 Problemática a resolver 22
      4.1.4 Objetivos del negocio 22

4.1.5 Criterios de éxito 22
4.1.6 Riesgos y contingencias 23
4.2 Aplicación de las técnicas de Minería de Datos para la segmentación de clientes 23
4.2.1 Comprensión de los datos 23
4.2.2 Recopilación de los datos iniciales 23
4.2.3 Descripción de los datos 24
4.2.4 Exploración de los datos 27
4.2.5 Verificación de la calidad de los datos 29
4.2.6 Preparación y muestreo de los datos 29
4.2.7 Realización del modelo 35
4.3 Evaluación del modelo creado para la segmentación de clientes .. 43
4.3.1 Validación de los pasos para la ejecución de las técnicas de modelado 43
4.4 Discusión 44
5. Conclusiones y Recomendaciones 47
5.1 Conclusiones 47
5.2 Recomendaciones 48
6. BIBLIOGRAFÍA 49

ÍNDICE DE FIGURAS

Figura 1:Clasificación de las técnicas de Machine Learning. 6
Figura 2: Clasificación de las técnicas de segmentación 12
Figura 3: Proceso CRISP-DM 15
Figura 4 Total de registros de Clientes 28
Figura 5 Total de tipos de exámenes 28
Figura 6 Total de registros de Ventas 28
Figura 7 Total de órdenes de pedido 29
Figura 8 Ejemplo de direcciones de clientes 31
Figura 9 Sectores en el mapa de Quito 32
Figura 10 Estructura de la DATA en BigML 36
Figura 11 Exámenes de áreas por edad 37
Figura 12 Clusters K-means K=5 38
Figura 13 Distribución con la data del CLUSTER-0 38
Figura 14 Clusters K-means K=8 39
Figura 15 Clusters G-means CV=5 39
Figura 16 Distribución con la data del CLUSTER-0 en G-means 40
Figura 17 Grafo reglas de asociación edad, sector, examen 41
Figura 18 Grafo reglas de asociación edad, examen, sector 42

ÍNDICE DE TABLAS

Tabla 1: Estructura archivo CLIENTES..... 25
Tabla 2: Estructura archivo VENTAS_CLI_ART 26
Tabla 3 Estructura archivo ORDEN_PEDIDO..... 26
Tabla 4 Estructura Tabla VENTAS_PED_EXA_F..... 27
Tabla 5 Estructura tabla DATOS_EXA_F 27
Tabla 6 Reglas de asociación para edad sector y examen..... 41
Tabla 7 Regla de asociación para Edad, Examen y Sector 42
Tabla 8 Reglas de asociación para Examen, Sector y Edad 43

1. Capítulo I. Introducción

1.1 Generalidades

La minería de datos conocido también como el proceso para el descubrimiento de datos o de conocimiento, aplicada en el ámbito comercial de cualquier área de comercialización, permite extraer, descubrir y almacenar información que permita identificar patrones de comportamiento de clientes.

El clustering o agrupamiento es una de las técnicas en minería de datos que más se utilizan en segmentación y clasificación de clientes, con esta técnica se busca dividir en pequeños segmentos un conjunto de datos, en el cual cada segmento contiene datos similares en su segmento y una diferencia clara con el resto de segmentos.

El agrupamiento o segmentación de clientes en el ámbito de la salud no es frecuente, podemos encontrar aplicaciones de minería de datos y agrupamiento en diversas áreas de la comercialización de productos y servicios, por ejemplo en la *comercialización de insumos de motos en el Perú (Rojas Huamán, 2020), en el mercado de la confección (Cálad Noreña, 2015), en instituciones financieras (cooperativas de ahorro y crédito) (Tamayo W., Jovel;, 2020), en empresas de telecomunicaciones (Jiménes Avalos, 2021),* en los que se ha aplicado las mismas técnicas, metodologías y algoritmos que entregan resultados adecuados a la línea de negocio, situación que no sucede en el ámbito de la salud debido al comportamiento del consumidor de este tipo de servicios. En la actualidad los servicios médicos han cambiado de un enfoque de tratamiento a uno de prevención, sin embargo, esto en el Ecuador no funciona así, seguimos trabajando en relación a los servicios médicos por tratamiento y no por prevención.

En el Ecuador a diferencia de otros países del mundo no es frecuente la aplicación de técnicas para entender el comportamiento del consumidor del sector de la salud, algunos estudios han demostrado que la reducción del 5% en la tasa de deserción de los clientes ha permitido un incremento de la ganancia en un 25% - 85% (Reichheld & Sasser, 1990). Se podría deducir que, mantener un cliente es más favorable de crear nuevos clientes en términos de una gestión estratégica de la salud.

Hoy que los clientes de servicios de salud seleccionan sus proveedores basados en los diferentes medios de información, el identificar clientes leales es fundamental, esta fidelidad se podría identificar y generar a través del análisis de la información almacenada de aquellos clientes que han mantenido una relación con la organización de manera continua e introduciendo el marketing de gestión de relaciones con los clientes por sus siglas en inglés Customers Relationship Management (CRM).

Los CRM, entre otras de sus funciones, segmentan los clientes para establecer un proceso de marketing y se están convirtiendo en un tema importante

en la gestión de la salud; esta surge de la estrategia de segmentación, orientación y posicionamiento (STP); este separa un mercado de clientes (Segmentación), selecciona el mercado objetivo (Targeting) y luego posiciona un producto o servicio (posicionamiento).

Esta propuesta tiene como objetivo aplicar una segmentación de clientes en el Laboratorio Clínico Pura Vida con la finalidad de identificar clientes leales, establecer políticas y estrategias de mercadeo, establecer predicciones respecto a la comercialización de sus servicios, al comportamiento del consumidor y a la fidelización de sus clientes; todo esto aplicando técnicas de minería de datos en las que se proteja la integridad de los clientes en base a procesos de ETL por sus siglas en inglés de Extract, Transform, y Load (Extracción, Transformación y Carga).

1.2. Planteamiento del problema

Todos los segmentos de mercado del área de la salud, antiguamente trabajan en función de lo dispuesto por el médico tratante; como el modelo hipocrático de la relación médico paciente lo manifestaba (Alvarez-Alva & Kuri-Morales, 2018), esto ya no ocurre así, hoy en día, son cada día más las personas que deciden optar por un proceso de salud por prevención. Sin embargo, todavía existen personas que lo hace porque tienen la necesidad ya sea porque existe la patología o porque el médico tratante lo exige.

En la actualidad los servicios médicos han cambiado de un enfoque de tratamiento a uno de prevención, esto significa que el mercado de tratamiento está siendo sustituido por un mercado de prevención (Alvarez-Alva & Kuri-Morales, 2018).

Los clientes de hospitales, centros de salud y especialmente en laboratorios clínicos, seleccionan a sus proveedores de servicios médicos con mayor libertad y en función de información obtenida de diversas fuentes; como lo menciona Alvarez-Alva y Kuri-Morales (2018) *"El paciente actual ha adquirido un papel de consumidor en el que compara diagnósticos, paquetes, oportunidades y hasta ofertas de servicio"*. esto preocupa a los proveedores de servicios de salud por la fácil y frecuente pérdida de clientes.

En estas condiciones el negocio de las áreas de la salud depende en un porcentaje de terceros y en otro por la decisión del paciente, en tal virtud, en el campo de la salud la posibilidad de tener un cliente está en función de los convenios y compromiso a los que se llega con los médicos, y si a esto incorporamos los resultados obtenidos en el estudio realizado por Reichheld y Sasser (1990) en el que se manifiesta que reducir en un 5% la tasa de deserción de los clientes permitirá un incremento de la ganancia de un 25% - 85%. Estos resultados del estudio son respaldados en el libro La calidad en el servicio al cliente (Vertice, 2008) página 44; por (Herrera Rangel, 2016) en su documento Diseño de un programa de retención de usuarios de tarjeta de crédito; por (Urtiz Villanueva, 2018) con su documento Estrategia de Márketing Relacional U/R 360 GYM & Fitness; por (Osorio Oncoy,

2019) con su documento Satisfacción del cliente en las agencias de viaje, Huaraz – 2017.

Los profesionales de laboratorio clínico desconocen la potencialidad y riqueza que tienen en sus datos, incluso el cómo aprovecharlos, esto se evidencia ya que en el análisis de la investigación preliminar no se pudo identificar aplicaciones de minería de datos realizados con datos de laboratorios clínicos.

El Laboratorio Clínico Pura Vida es un laboratorio privado con más de 15 años en el mercado de servicios de salud, cuenta con dos tipos de clientes; los clientes particulares, de los cuales cuentan con toda la información de los exámenes realizados y los datos del cliente; los clientes por convenio de los cuales disponen información del cliente, pero no de su ubicación geográfica. Su gerente propietaria ha evidenciado la disminución de sus clientes particulares sin poder hacer mucho al respecto porque se depende en su mayoría de la referencia que entregan los médicos y estos a su vez refieren dependiendo del porcentaje de comisión que se ofrezca.

Lamentablemente esta es la realidad de los servicios de salud en nuestro país, por eso para un laboratorio clínico que no desee gestionar estos convenios y compromisos con los médicos es necesario poder contar con alternativas que le permita aplicar planes de trabajo para retener clientes (administración o mercadeo de las relaciones con los clientes), aplicar promoción directa y fidelizar a los clientes existentes.

En base a lo descrito nos hemos planteado las siguientes interrogantes a resolver con la propuesta de este trabajo, con miras a alcanzar el objetivo de la segmentación de los clientes del Laboratorio Clínico Pura Vida

¿Es posible a través de la información de facturación pronosticar el comportamiento de los clientes de un laboratorio clínico?

¿Se puede aplicar los métodos y técnicas de la minería de datos en la segmentación de clientes de servicios de salud?

¿Es suficiente la información de ventas de servicios médicos que disponen los laboratorios clínicos para segmentar y fidelizar clientes?

¿A través del procesamiento y análisis de estos datos se puede obtener información inteligente que permita crear estrategias de mercado en servicios de salud?

Los procesos de análisis de datos ven con mucha importancia el entender el negocio, para poder aplicar algún método, metodología, técnica y sobre todo modelos de aprendizaje automático que más se ajusten a los datos y lógica del negocio en estudio. En este proyecto se busca entender el negocio de los laboratorios clínicos y evidenciar que estos modelos, técnicas y metodologías son aplicables. En esto radica la importancia de responder las preguntas planteadas en esta propuesta.

1.3 Objetivos

1.1.1 Objetivo General

Aplicar técnicas, métodos y metodologías de la ciencia de datos, que después de un análisis conceptual permitan segmentar los clientes de un laboratorio clínico.

1.1.2 Objetivos Específicos

- Analizar las técnicas y metodologías de la minería de datos para dar solución a la segmentación de clientes
- Analizar la estructura de datos de los sistemas de información y la situación actual del Laboratorio Clínico Pura Vida
- Aplicar técnicas de minería de datos para la segmentación de clientes en el Laboratorio Clínico Pura Vida
- Evaluar el modelo de datos obtenido para la segmentación de clientes en el Laboratorio Clínico Pura Vida

1.4 Alcance

Para este proyecto se dispone de los datos del laboratorio clínico, estos datos están disponibles desde el 2014 a la fecha ya que en el día a día se sigue registrando facturas por servicios del laboratorio. El laboratorio cuenta con dos tipos de clientes (particulares y convenios) para este trabajo se tomará en cuenta solo los historiales de compra de los clientes particulares debido a que a los convenios se realiza una sola factura al mes de todos los exámenes realizados.

De los clientes particulares se analizará la fecha de la venta del servicio, la cantidad por examen de laboratorio y el tipo de examen; de esta manera queda fuera del análisis la dirección del paciente o el sector del paciente para el estudio. Los resultados obtenidos serán entregados a la Gerente Propietaria del laboratorio, pero no se realizará la implementación del modelo obtenido en el laboratorio.

2. Capítulo II. Revisión literaria

2.1 Minería de Datos

Los datos que se pueden obtener de una organización contienen estructuras patrones y reglas de las cuales es posible extraer conocimiento respecto los eventos que los generaron. La disponibilidad de grandes volúmenes de datos en las organizaciones y el advenimiento de las nuevas tecnologías a transformado el análisis de los datos en el uso de técnicas y herramientas que se integran en lo que hoy se conoce como la minería de datos; esta ha pasado de entender los datos a comprender los eventos que hay detrás de estructuras y relaciones entre estos (Gironés-Roig, Casas-Roma, Minguillón-alfonso, & Caihuelas-Quiles, 2017).

Se define la minería de datos como un proceso de descubrimiento de relaciones nuevas y significativas que se identifican al analizar y examinar grandes volúmenes de datos. Para esto la minería de datos exige un trabajo multidisciplinario y basa sus orígenes en cuatro grandes áreas del conocimiento; la primera la estadística clásica y la descriptiva que abarcan conocimientos como desviación estándar, varianza, análisis de discriminantes, análisis de regresión, distribuciones, entre otros.

La Inteligencia Artificial es la segunda área, esta área del Machine Learning se construye con heurísticas e intenta aplicar el pensamiento humano como solución a problemas estadísticos.

El álgebra lineal y el cálculo como un campo de la matemática es la tercera área considerada un pilar fundamental en el campo de la minería de datos, es recomendado como un prerrequisito de estudio antes de iniciar con el aprendizaje automático.

El Aprendizaje Automático (Machine Learning) que es descrita como la unión entre las tres áreas nombradas anteriormente; busca que los programas aprendan en función a los datos que son analizados, identificando relaciones entre los atributos y entidades que permiten a los expertos del negocio identificar ideas ocultas que pueden implementarse en aplicaciones inteligentes.

2.1.1 Técnicas de minería de datos

Para descubrir el conocimiento contenido en la información almacenada en grandes bases de datos se utilizan técnicas de minería de datos. El objetivo de éstas es descubrir patrones, tendencias y perfiles utilizando técnicas de reconocimiento de patrones, redes neuronales, lógica difusa entre otras técnicas de análisis de datos.

Las técnicas de minería de datos se clasifican en técnicas de aprendizaje supervisadas y de aprendizaje no supervisadas; en las técnicas supervisadas se encuentran las técnicas de Clasificación y de Regresión, dentro de las no

supervisadas se encuentra las técnicas de Clustering como se muestra en la Figura 1

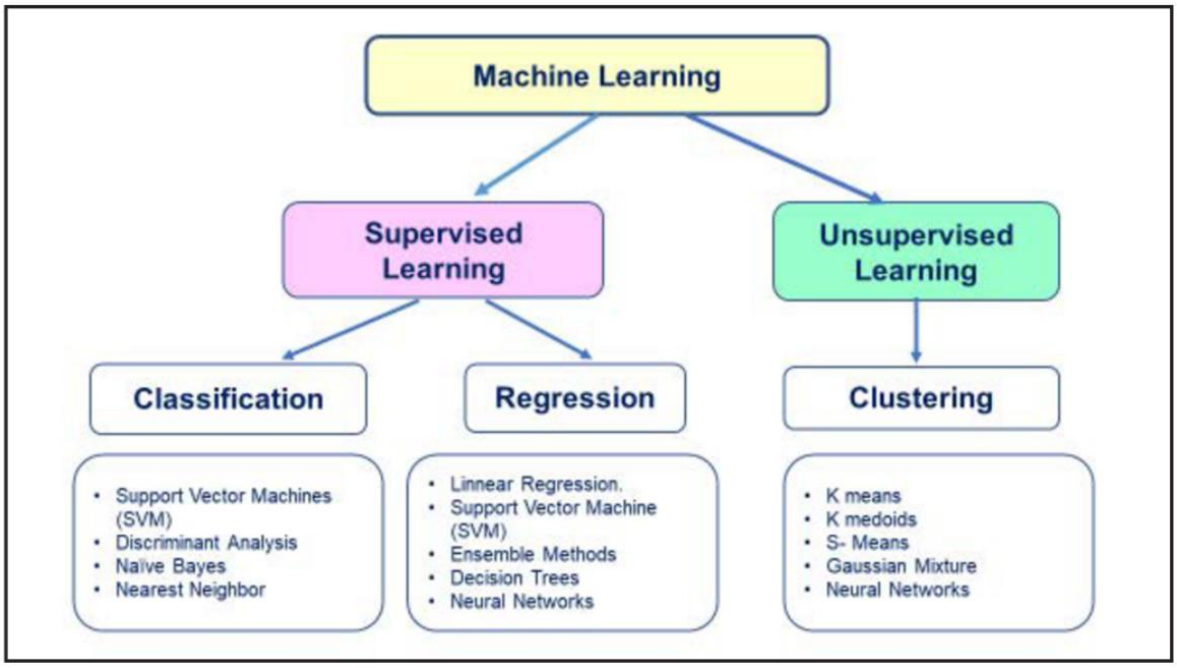


Figura 1:Clasificación de las técnicas de Machine Learning.
Fuente: (Sáiz-Manzanaree, Escobar, & Rodríguez-Medina, 2019) P-158.

Las técnicas de aprendizaje supervisado tienen como objetivo descubrir las relaciones existentes entre las variables de entrada y las variables de salida, a través de un modelo en el que se representan las relaciones encontradas. En las técnicas de aprendizaje supervisado los valores correctos son proporcionados por un supervisor.

En las técnicas de aprendizaje no supervisado no existe un supervisor y se basa en datos de entrada que se encuentren disponibles. El objetivo en esta técnica es identificar regularidades, irregularidades, relaciones, similitudes, y asociaciones que podrían ser identificadas en los datos de entrada. (Sáiz-Manzanaree, Escobar, & Rodríguez-Medina, 2019)

Con las técnicas de aprendizaje supervisado se puede aplicar modelos más complejos que permitan aprender más a diferencia de las técnicas de aprendizaje no supervisado.

a. Técnicas supervisadas

- Clasificación.

La clasificación busca un modelo válido para predecir casos futuros a partir de casos conocidos. En esta técnica se divide un conjunto de datos en grupos mutuamente excluyentes; se busca que cada miembro de un grupo esté lo más cerca a otros y otros grupos diferentes estén lo más lejos de otros; la distancia entre estos miembros de grupo se mide en función a las

variables que se quiere predecir. La técnica de clasificación permitirá identificar categorías o etiquetas en las que se clasificarán los datos utilizados en el análisis.

- Predicción

Es la actividad que predice los valores de una o varias variables, en función a un conjunto de datos. El predecir valores continuos se lo puede hacer con las técnicas de regresión (técnicas estadísticas)

- Regresión

El objetivo de esta técnica estadística es predecir una variable continua a partir del desarrollo de otra variable continua en su mayoría el tiempo.

- Lógica borrosa

Más conocida como lógica difusa, busca modelar la realidad identificando una forma más exacta y evitando el determinismo o la exactitud; permite tratar los datos categóricos de forma probabilística.

b. Técnicas no supervisadas

- Reglas de asociación

Esta técnica se aplica cuando se desea hacer análisis exploratorio; cuando se desea establecer relaciones entre distintas acciones o sucesos independientes se emplean reglas de asociación, pudiendo evidenciar como la ocurrencia de un suceso o acción puede inducir la aparición de otro suceso o acción

- Clustering o agrupamiento

El análisis de clustering o agrupamiento tiene como objetivo el agrupar o segmentar una colección de datos entre los más cercanamente relacionados. Lo realiza identificando topologías o grupos donde sus elementos guardan gran similitud entre si y grandes diferencias con otros grupos.

2.1.2 Algoritmos de Clustering

a. Algoritmos jerárquicos

El agrupamiento jerárquico está basado en la construcción de un árbol o dendograma en el que las hojas representan el conjunto de datos y buscando construir una jerarquía de particiones o grupos de datos.

Dependiendo de la manera de construcción del árbol, hay dos tipos de algoritmos de agrupamiento jerárquico que proporcionan resultados similares pero que trabajan con enfoques inversos; estos son algoritmos aglomerativo y desaglomerativo o divisivo (Gironés-Roig, Casas-Roma, Minguillón-Alfonso, & Caihuelas-Quiles, 2017); los algoritmos aglomerativos son de más fácil construcción ya que existe un solo modo de unir los conjuntos a diferencia de que existen múltiples formas de separar un conjunto.

- **Algoritmo Aglomerativo (Agglomerative Hirerchical Clustering AHC)**

Este algoritmo aplica una estrategia bottom-up es decir el árbol se construye iniciando por las hojas y finalizando en la raíz; parte de los datos fragmentados completamente y se van juntando progresivamente hasta que todos los datos pertenecen a un mismo grupo en la raíz (agrupamiento o clustering).

- **Algoritmo Desaglomerativo o Divisivo (Divisive Hirerchical Clustering DHC)**

A diferencia del algoritmo aglomerativo, esta aplica una estrategia top-down, parte de la raíz que contiene todos los datos (un solo grupo) y va haciendo divisiones recursivas hasta legar a las hojas, es decir de forma inversa al algoritmo aglomerativo (segmentación)

b. Algoritmos de particionamiento

Se definen a los algoritmos que están diseñados para clasificar o construir *k* particiones o *k* grupos de individuos (no variables). Cada partición representa un conglomerado o grupo; el algoritmo inicia eligiendo una partición de los individuos, busca el mejor agrupamiento y reubica o intercambia los objetos de un grupo a otro hasta obtener una mejor partición que actuará como criterio de parada. Intuitivamente una mejor partición debe considerar que la dispersión dentro de los grupos sea la menor posible.

Entre algunos algoritmos de este tipo se tiene: *K*-Means, *K*-Medians, *K*-Medoids,

- ***K*-Means:**

Es el método interactivo más usado, es un algoritmo de clasificación no supervisado de tipo particional, se aplica en situaciones en que todas las variables son de tipo cuantitativo, está basado en dividir un conjunto de n

observaciones en *k* grupos cuya distancia euclidiana es escogida como medida de disimilitud (Aggarwal & Reddy, 2014).

Se debe tener en cuenta que:

- Se debe definir el número de grupos o clúster (*k*) antes de ejecutar el algoritmo.
- Cada *k* está definida por un punto que se lo define como centroide del clúster (Gironés-Roig, Casas-Roma, Minguillón-Alfonso, & Caihuelas-Quiles, 2017)

Se trata de un método de agrupamiento por vecindad, parte de un número determinado de prototipos y un grupo de ejemplos a agrupar sin etiquetar. En términos generales funciona en dos fases:

1. La fase de inicialización identifica *k* puntos como centroides iniciales. Es decir, divide el conjunto de elementos en *k* grupos.
2. Es interactiva y consiste en:
    a. Calcula las distancias euclidianas de cada elemento a cada uno de los *k* centros y asigna a cada centroide los puntos del conjunto de datos más cercanos (forma grupos disjuntos).
    b. Recalcula los nuevos centroides en base a los puntos que toman parte del grupo del cual va y para el grupo al cual llega. (Gironés-Roig, Casas-Roma, Minguillón-alfonso, & Caihuelas-Quiles, 2017)

Estos dos pasos se repiten hasta que se defina un criterio de aceptación óptimo caso contrario repite el punto 2.

El comportamiento del algoritmo se define en base a cinco criterios importantes:

Inicialización de los centroides: La inicialización de centroides con *k* puntos aleatorios del conjunto de datos es el enfoque de inicialización ampliamente usado. Un problema de esta inicialización es que; diferentes ejecuciones del algoritmo conducen a diferentes modelos y a diferentes niveles de calidad del resultado; este problema es superado gracias a la simplicidad del algoritmo que permite ser ejecutado varias veces con distintos centroides para identificar el mejor resultado.

Cálculo de distancia: *k*-means generalmente utiliza la distancia euclídea, sin embargo, se puede utilizar otro tipo de métrica de similitud

Recálculo de los centroides: El valor de los centroides se calcula en función de la media de los puntos que pertenecen al segmento, por esta razón el algoritmo se aplica cuando los atributos son continuos, caso contrario se debe aplicar una transformación previa.

Criterios de parada: La condición de parada es la convergencia del algoritmo, se alcanza cuando no existe recálculo de centroides durante una interacción completa, esto provoca que no haya alteraciones en las distintas particiones grupos o clústeres (estabilidad). La condición de

parada se garantiza solo después de un número finito de interacciones y dependiendo de la métrica que se aplique en el cálculo de la distancia y de los centroides; en la práctica no es necesario que converja, bastaría estar cerca a una situación de convergencia.

Criterios para seleccionar un valor de *k*: En *k*-means se debe indicar el número de clústeres con los que se debe trabajar, este valor permite mantener una eficiencia y simplicidad elevada aunque puede ser considerado como un inconveniente importante; identificar el valor de *k* (número de particiones) no es una tarea simple y no siempre es factible extraerlo del conocimiento del dominio, el rango de valores de *k* no es muy grande por ende se puede probar diferentes valores y seleccionar el que mejores resultados entregue.

Un criterio es minimizar la suma de residuos cuadrados (RSS residual Sum of Squares) para buscar la creación de segmentos lo más compactos posible. Otro es el de maximizar la suma de distancias entre segmentos. Otra alternativa es permitir que el valor de *k* se modifique durante la ejecución del algoritmo de acuerdo a ciertos criterios como:

a) si existen dos particiones con los centros muy juntos, es mejor unir para reducir el número de peticiones
b) si el nivel de diversidad es muy elevado en la partición se puede dividir en dos particiones y se incrementa el valor de *k*.

Estos criterios pueden aplicarse para definir un valor para *k*, sin embargo, hay que considerar que los algoritmos no siempre alcanzaran un valor óptimo, esto puede darse porque el algoritmo no es capaz de encontrar un óptimo ya sea porque los datos no contienen estructuras de segmentos o porque no se ha definido adecuadamente el valor de *k* de segmentos a definir. (Gironés-Roig, Casas-Roma, Minguillón-alfonso, & Caihuelas-Quiles, 2017)

- ***K*-medians**

  Es una variante del método anterior, la diferencia está en que se basa en el cálculo de la mediana y no en el valor medio como el anterior, aunque es más complejo es mejor en algunos contextos; puede implementarse mediante distintas métricas sin embargo la más aplicada es la distancia de Manhattan.

- ***K*-medoids**

  Este método presenta mejores resultados cuando las distancias son asimétricas o existen outliers. Este algoritmo propone el recálculo de los centroides a partir de las instancias de los clústers que presentan valores distintos mínimos con respecto a los demás clústers; la identificación de estos medoids es mucho más costosa, sin embargo, hay que considerar que este método es mucho más robusto en caso de los outliers.

2.2 Segmentación de Clientes

2.2.1 Criterios para la segmentación de clientes

a. **Por criterios geográficos**: este criterio es el más utilizado, asigna una zona del mapa por vendedor (Bastos Boubeta, 2007)
b. **Por criterios relacionados con el producto**: la segmentación por este criterio es la más directa, son fácilmente cuantificables y de fácil acceso; estos criterios no profundizan en las causas que justifican el comportamiento del consumidor. (Bastos Boubeta, 2007)
c. **Por criterio socioeconómico demográfico**: toma en cuenta los perfiles socioeconómicos de los consumidores que son los que explican sus comportamientos y preferencias, las variables socioeconómicas permiten realizar una segmentación más eficaz para establecer grupos de clientes por edad, género, ingresos etc. (Bastos Boubeta, 2007)
d. **Según la frecuencia de compra**: es muy sencilla y de gran interés ya que se usa con frecuencia para crear carteras de clientes potenciales, ocasionales y fieles (Bastos Boubeta, 2007)
e. **Según su volumen de compra**: en función de este criterio se establecen grupos como: clientes por cantidades de consumo al mes, año etc. Clientes que gastan una determinada cantidad de dinero en compras por semana, mes o frecuencia de compra. (Bastos Boubeta, 2007)

2.2.2 Técnicas de segmentación de clientes

La clasificación de las técnicas de segmentación distingue técnicas predictivas y descriptivas.

- Técnicas Predictivas
  En las técnicas predictivas, las variables que intervienen pueden clasificarse como independientes o dependientes; en esta técnica se especifica el modelo en base a un conocimiento teórico previo y este debe ser contrastado antes de asumirlo como válido; la aplicación de cualquier modelo debe superar las fases de: identificación objetiva, estimación, diagnosis y predicción.
  Identificación objetiva: en función de los datos se debe aplicar normas y reglas que permitan identificar el mejor modelo.
  Estimación: cálculo de los datos de los parámetros elegidos para el modelo.
  Diagnosis: validación de la validez del modelo
  Predicción: aplicación del modelo para predecir las variables dependientes.
- Técnicas Descriptivas
  En las técnicas descriptivas todas las variables tienen el mismo estatus, no se les asignan ningún papel predeterminado, no se presupone la existencia de variables dependientes e independientes ni de un modelo previo. Los modelos se crean automáticamente en función al

reconocimiento de patrones, en las técnicas descriptivas se encuentran las técnicas de clustering y de reducción de dimensiones.

Cualesquiera de estas dos técnicas se enfocan al descubrimiento del conocimiento embebido en los datos.

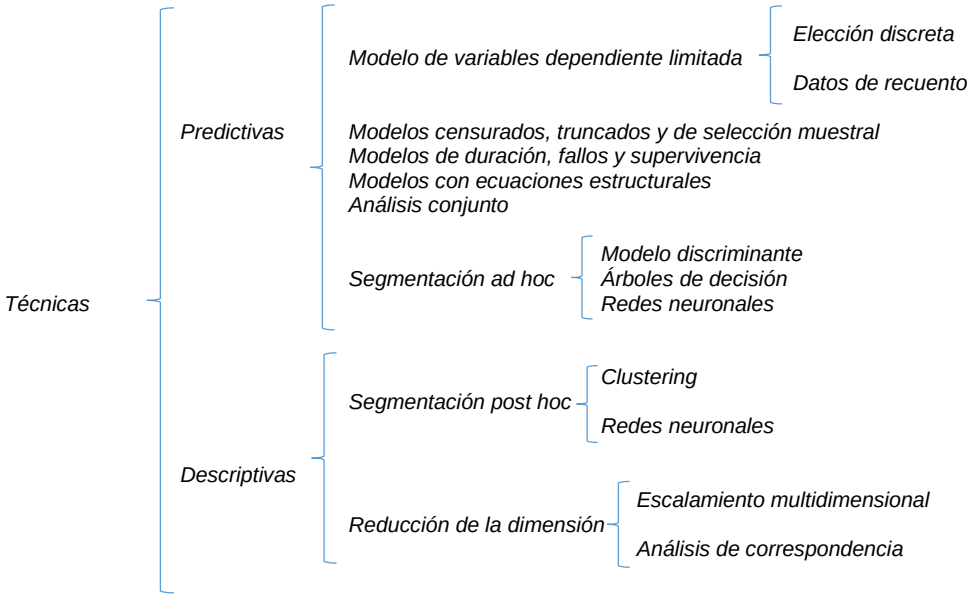


Figura 2: Clasificación de las técnicas de segmentación
Fuenate: (Pérez, 2011) P-3

2.2.3 Análisis RFM

El análisis RFM (Recency, Frecuency, Monetary) es un método de marketing utilizado para el análisis de comportamiento de los clientes en función de sus hábitos de gastos, permite identificar aquellos clientes que son capaces de responder a una nueva oferta.

Se basa en tres factores:

Actualidad. Hace referencia a que los clientes que han realizado compras recientemente tienen mayor probabilidad de volver a comprar en relación a los clientes que han realizado compras en el pasado.

Frecuencia. Determina que los clientes que han adquirido más productos tiene más posibilidades de adquirir nuevamente que aquellos que han adquirido menos productos

Valor Monetarios: Hace referencia a la cantidad total invertida, es decir los clientes que más han invertido en sus compras tienen mayor probabilidad de responder a una nueva oferta.

Su funcionamiento se basa en asignar a los clientes una puntuación en función a la fecha de compra más actual, esta puntuación está basada en una clasificación simple de puntuación de acuerdo a una categoría, por ejemplo, los clientes con

fecha de compra más reciente reciben una puntuación de 5 y los que tienen compras más antiguas reciben una puntuación de 1. De la misma forma se asigna una puntuación de 5 a los clientes con mayor frecuencia de compra y finalmente a los clientes con un valor monetario más alto reciben una puntuación de 5 y los más bajos una puntuación de 1.

Adicionalmente se considera una puntuación para el RFM combinada, no es más que la combinación de las tres puntuaciones anteriores, por ejemplo, si un cliente tiene en Actualidad = 5, Frecuencia = 3 y valor monetario = 4 entonces su RFM combinado sería 534; los clientes ideales son los que tienen un RFM combinado de 555.

Este es un método muy utilizado para realizar análisis de segmentos de clientes basado en su historial de compra; desde el punto de vista de comportamiento del cliente este método permite medir la relación del cliente con la empresa.

2.3 Metodología de Minería de Datos

2.3.1 SEMMA

Fue desarrollado por SAS Institute Inc., es un proceso de selección, exploración y modelado de grandes volúmenes de datos; permite descubrir patrones en los datos de negocios desconocidos. SEMMA es el acrónimo de 5 fases básicas del proceso como Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado) y Assess (Valoración).

Para que la minería de datos sea exitosa debe verse como un proceso, a una muestra representativa de los datos, SEMMA puede aplicar técnicas exploratorias estadísticas y de visualización, que permiten identificar seleccionar y transformar las variables predictivas en resultados garantizados por la precisión del modelo. Al evaluar los resultados de cada etapa del proceso SEMMA, se puede volver a reformular nuevas interrogantes que nos lleve a refinar mucho más los datos ya que SEMMA está definido en un siclo iterativo.

Paso 1. Muestra. En este paso se define o extrae la información lo suficientemente grande pero que sea significativa y lo suficientemente pequeña para que pueda ser manipulada rápidamente, para conseguir esto es importante aplicar técnicas de muestreo que permitan un rendimiento computacional y un costo adecuado. Se utiliza el método de descripción de datos exploratorios cuando existen nichos pequeños que no se representan en la muestra pero que son importante considerarlos; y para una mejor evaluación de la precisión es recomendable utilizar datos particionados. Los datos de entrenamiento se utilizan para el ajuste del modelo, los de Validación para la evaluación y evitar el sobreajuste y los de prueba para corroborar la calidad del modelo. (Olson, 2018)

Paso 2. Explorar. En este paso se identifican tendencias o anomalías haciendo exploraciones visual o numéricamente que permitan tener una mejor comprensión de los datos permitiendo un proceso de descubrimiento más efectivo. Si la exploración visual no revela tendencias claras se puede recurrir a técnicas

estadísticas, análisis factorial, análisis de correspondencia y la agrupación. (Olson, 2018)

Paso 3. Modificar. Según los descubrimientos dados en la fase de exploración, en esta etapa se crean, seleccionan y transforman las variables del modelo para incluir información (grupos y subgrupos), introducir nuevas variables, buscar oultliers o valores atípicos, o reducir variables para resumirlas en las más significativas; como la minería de datos es un proceso dinámico e interactivo, se puede escoger o actualizar los modelos cuando se disponga de nueva información. (Olson, 2018)

Paso 4. Modelo. Una vez que se hayan preparado los datos en los pasos anteriores, se estará listo para construir modelos de datos combinando variables que permitan predecir un comportamiento deseado o esperado; los modelos se construyen con técnicas de minería de datos como: redes neuronales artificiales, soporte vectorial, modelos logísticos y otros modelos estadísticos como series de tiempo, razonamiento basado en memoria, y el análisis de componentes especiales. Cada una de estas técnicas tiene particularidades y es apropiado para ciertas situaciones específicas en la minería de datos. (Olson, 2018)

Paso 5. Evaluar. En este paso se evalúa el modelo haciendo uso de una parte de los datos que se deja de lado en la etapa de muestreo, estos datos no son usados en las etapas de la construcción del modelo; con estos datos se evalúa que tan bien funciona el modelo y la utilidad y confiabilidad de los hallazgos. El modelo debe funcionar con los datos de evaluación igual que con los datos que se usaron para construir el modelo, así como con datos conocidos (Olson, 2018)

2.3.2 CRISP-DM

Es un proceso de 6 fases que dispone de retroalimentación en diferentes fases, por lo tanto, no se considera un proceso rígido (Olson, 2018); en la Figura 3 se puede observar cómo se interactúa en cada una de estas fases.

Comprensión Empresarial (Business Understanding): El comprender el propósito y los objetivos del estudio es fundamental en este proceso, se debe considerar la necesidad de nuevos conocimientos y de los objetivos empresariales establecidos por la gerencia, estos objetivos pueden estar en términos de: el tipo de cliente, que tipos de clientes están interesados con que productos, que valor aportan etc. Una vez identificadas las necesidades y los objetivos se debe establecer un plan y los presupuestos para llegar a deducir dicho conocimiento en términos de los responsables en generar o recopilar los datos, analizarlos e informarlos.

Comprensión de los Datos (Data Understanding): En esta etapa se incluyen las actividades de recopilación de los datos, descripción, exploración y verificación de la calidad de los datos. La fuente de los datos para la recopilación puede variar, los tipos de fuentes de datos incluyen datos demográficos, sociográficos, y transaccionales; tanto la exploración como la visualización de los datos pueden realizarse al final de esta etapa. El análisis de conglomerados puede realizarse en esta etapa.

Los datos pueden categorizarse como cuantitativos y cualitativos; los datos cuantitativos se distinguen por ser valores numéricos, estos pueden ser discretas (números enteros) o continuos (números reales); los datos cualitativos o categóricos son nominales (valores no ordenados finitos) u ordinales (valores ordenados finitos)

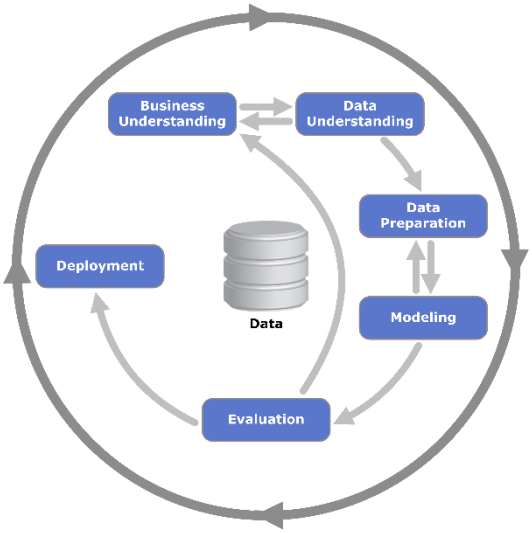


Figura 3: Proceso CRISP-DM
Fuente: Wikipedia: https://es.wikipedia.org/wiki/Archivo:CRISP-DM_Process_Diagram.png (Creative Commons)

Preparación de datos (Data Preparation):

Los datos que se utilizan para el análisis en lo general provienen de diferentes fuentes y diferentes formatos, estos deben ser convertidos a formatos electrónicos coherentes con el análisis y clasificación que se va a aplicar. El propósito de la preparación de los datos es limpiar y adecuar los datos en una mejor calidad; en general la limpieza de los datos significa validar, filtrar, completar y agregar valores faltantes, se buscan valores a típicos o redundantes que pueden sacar de contexto el análisis realizado.

Los valores atípicos pueden ser causados por diferentes razones: errores humanos, errores técnicos o debido a eventos externos que ocasionan algún cambio en los datos, sin embargo, estos deben ser analizados en función del contexto.

Entre los estadísticos comunes más utilizados y más simples de aplicar para agregar o suavizar los datos tenemos el max, min, la media y la moda; entre las herramientas de visualización podemos identificar a los diagramas de dispersión y a los diagramas de caja que generalmente son usados para filtrar los valores atípicos. Se podrían aplicar técnicas más avanzadas como, el análisis de regresión, el análisis de conglomerados, el árbol de decisiones o el análisis jerárquico; debido a lo detallado y tedioso de esta actividad generalmente la preparación de datos lleva mucho tiempo (más del 50% del todo el tiempo aplicado en la extracción de los datos).

Los datos se pueden expresar en varias formas y tipos de datos:

RANGE: son valores numéricos como entero, real, de fecha y hora

FLAG: son considerados como binarios si o no, 1 o 0 u otros datos en los que se considere dos valores.

SET: son considerados como valores múltiples tanto numéricos, de cadena o de fecha y hora

TYPELESS: para considerar otros tipos de datos.

Otras aplicaciones o herramienta de software tienen diferentes tipos de datos, los tipos de datos más comunes son: Numéricos, enteros, booleanos, categóricos, fecha, cadena y texto

Modelado (Modeling):

El modelado es la etapa donde aplicando un determinado modelo se generan resultados para diversas situaciones; en primera instancia se aplica el análisis de conglomerados o la exploración visual de los datos. Los modelos a aplicarse dependen del tipo de dato; por ejemplo, para la agrupación de datos si se conocen los grupos, el análisis de discriminantes puede ser el adecuado. Si se busca hacer una estimación y los datos son continuos o no continuos, la regresión lineal o la logística respectivamente sería la apropiada; en ambos casos también se podrían aplicar redes neuronales; en clasificación de datos, los árboles de decisión son otra herramienta útil. En términos generales lo que se busca es que el usuario trabaje con los datos para comprenderlos.

El tratamiento de los datos conocido como minería de datos es en esencia el análisis estadístico de datos muy grandes. En este tratamiento se divide este gran conjunto de datos en dos partes, una para el desarrollo y entrenamiento del modelo (training) y otra parte se reserva para probar el modelo (Test); con esto se puede obtener una prueba más convincente de la precisión del modelo.

La minería de datos se puede lograr en función de algunas técnicas como: asociación, clasificación, agrupación, predicción, patrones secuenciales, asociación difusa y secuencias de tiempo; que nos permitirán pasar a la etapa de interpretación de los datos, esta etapa es considerada muy crítica ya que busca asimilar el conocimiento de los datos extraídos.

Evaluación (Evaluation):

Es considerada una etapa crítica, pues en ella se realiza la interpretación de los datos; en la interpretación de los datos se evidencian dos problemas: el primero es como identificar o reconocer el valor comercial de los patrones de conocimiento descubiertos en el proceso de minería de datos. El segundo es identificar que herramienta de visualización se puede aplicar para mostrar de forma adecuada y entendible los resultados de la minería de datos, mientras más simple sea la interpretación gráfica más fácil entenderán los usuarios finales. Es necesario que, para obtener buenas interpretaciones en la minería de datos, se propicie la

interacción entre analistas de datos, analistas de negocio y tomadores de decisiones que conduzca a tomar decisiones productivas.

Implementación (Deployment):

La implementación es el acto de utilizar la minería de datos, es necesario monitorear el desempeño del modelo y ajustarlas a nuevas condiciones si es necesario.

3. Capítulo III. Marco metodológico

3.1 Materiales

Los materiales que se utilizarán en este proyecto están en base a las necesidades que se identifiquen en el proceso; en términos generales se hará uso de:

SQLServer: Gestor de base de datos sobre el cual se encuentra el repositorio de información del laboratorio objeto de estudio

CLINICAL Lab: Aplicación específica para la gestión de las actividades del laboratorio clínico, esta aplicación se conecta con el gestor de base de datos SQLServer en el cual se almacena toda la información el laboratorio

SisGem: Aplicación específica para la gestión de los procesos financieros del laboratorio se integra con CLINICAL Lab en las órdenes de pedido de los exámenes a realizar al paciente.

DBMS_Oracle: Gestor de base de datos a utilizar para la preparación y limpieza de los datos obtenidos de los sistemas de gestión del laboratorio objeto de estudio. La distribución de Oracle a usar será la XE que es de distribución gratuita. Se usará esta base de datos para almacenar la información contenida en archivos Excel y proporcionados por el laboratorio clínico.

Linux-Oracle: sistema operativo estructurado y mejorado por Oracle, sobre este sistema operativo se instalará la base de datos con la cual se trabajará la preparación y limpieza de los datos

Python: Lenguaje de programación que se utilizará para construir los algoritmos para el análisis de los datos.

BigML: Plataforma de machine learning que permite realizar análisis predictivo de forma sencilla y sin conocimiento de programación.

3.2 Métodos

3.2.1 Metodología

El procedimiento que se aplicará para la segmentación de clientes en el Laboratorio Clínico Pura Vida se apoya en la metodología CRISP-DM, la cual proporciona un enfoque estructurado para la planeación y gestión de la minería de datos.

De acuerdo a las etapas y secuencia que establece la metodología, se iniciará con la ***comprensión del negocio*** y de los datos, para esto se realizará un acercamiento con los responsables de las diferentes áreas del laboratorio directamente relacionadas con la segmentación de clientes. El objetivo de esta primera actividad es identificar las necesidades, el contexto en el que se desempeña el laboratorio y comprender las operaciones y el funcionamiento que rige el negocio

del laboratorio clínico desde un punto de vista de los datos; que permita identificar la completitud y validez de los datos.

En la segunda etapa se continuará con la ***preparación de los datos,*** para esto se extraerán los datos necesarios del gestor de bases de datos que contiene la información del laboratorio; luego se realizará un análisis de relaciones identificando atributos relacionales que permitan estructurar un modelo relacional de los datos a aplicar en el estudio propuesto. Con la estructura de datos propuesta se procederá a transformar y cargar los datos a las estructuras establecidas para aplicar el algoritmo seleccionado.

En la tercera etapa se continuará con el **modelamiento**; en esta etapa se pretende aplicar el algoritmo seleccionado en función de las variables seleccionadas previamente y validadas por los responsables del laboratorio clínico

3.2.2 Métodos y Técnicas

Para conocer la situación del laboratorio se usará la técnica de la observación y la entrevista a través de la cual se espera entender los datos y la información que maneja.

Los datos se obtendrán directamente de los sistemas del laboratorio en estudio, en formato Excel; todo este proceso se trabajará bajo las especificaciones de ETL (Extracción, Transformación y Carga) como se definen en la Figura 4.

La extracción de los datos se realizará del sistema SisGem para obtener datos de facturación de las pruebas de laboratorio realizadas a los clientes y del sistema CLINICAL Lab para obtener los datos de los pedidos y los exámenes realizados por el paciente.

La transformación se realizará usando la base de datos ORACLE-XE (libre) sobre una máquina virtual con ORACLE-LINUX en la que se crearán las estructuras de las tablas en las cuales se realizará toda la transformación y limpieza de los datos para posterior generación de los archivos csv con los que se aplicarán los modelos definidos en este estudio.

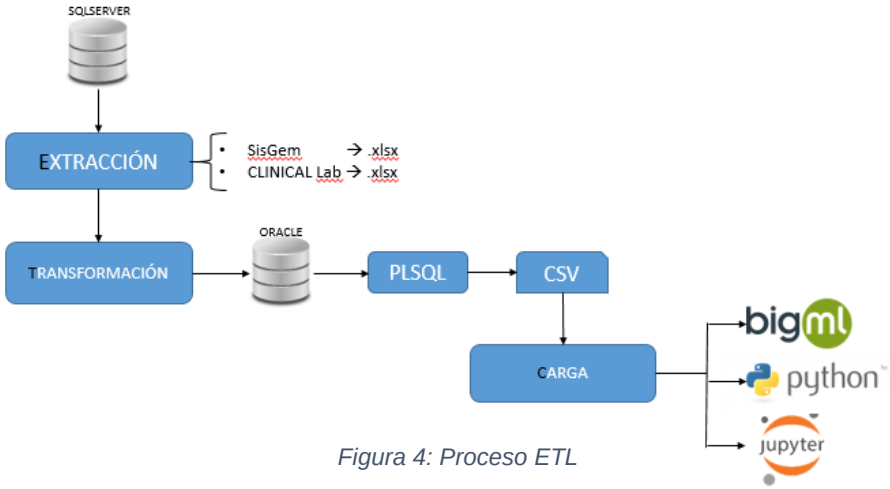


Figura 4: Proceso ETL

Se utilizarán datos de 5 años de facturación y realización de pruebas de laboratorio, desde el 2017 al 2021 incluyendo los cuatro primeros meses del 2022, estos datos hacen referencia a las pruebas facturadas a los clientes con el sector geográfico y los precios de cada prueba de laboratorio realizada; también se consideran archivos con las pruebas realizadas por pacientes, su ubicación geográfica y la fecha en que fue realizada la prueba.

La salud es un mercado que no se comporta como el resto de mercados:

1. No es un mercado promocional ya que no es posible considerar promociones por exámenes realizados o por temporadas,

2. La cultura del consumidor no es por temporada o moda, la salud del paciente ecuatoriano no es de prevención, en la mayoría de los casos es de corrección.

Considerando estos dos aspectos se aplicarán modelos de segmentación geográfica (clustering) aprovechando la ubicación de los pacientes, modelos de asociación considerando las pruebas de laboratorio, la edad del paciente y las fechas en las que han sido realizadas.

Es difícil evaluar los resultados obtenidos después de la aplicación del algoritmo no supervisados, por esta razón se aplicarán técnicas e índices que permitan identificar la validez de un agrupamiento y del clúster obtenidos, así como hiperparámetros como el número de clusters. Respecto a las reglas de asociación se aplicará las métricas de SUPPORT, CONFIDENCE y LIFT (Zumel, Nina; Mount, John;, 2019).

En este proyecto se aplicarán técnicas de validación interna que miden el agrupamiento únicamente con información de los datos y sin necesidad de información externa, sin embargo, se recurrirá a evaluación cualitativa en la que se buscará la opinión de expertos; aunque este método ha sido analizado en varias publicaciones científicas la mayoría concuerda en que aplicarlo permite integrar la novedad en el proceso (Bao, Mao, Zhu, Xiao, & Xu, 2021).

4. Capítulo IV. Resultados

4.1 Análisis del estado actual del Laboratorio Clínico Pura Vida

4.1.1 Comprensión del Negocio

Se analizó el negocio a través de entrevistas realizadas a la propietaria del laboratorio y con la observación se buscó entender la lógica del negocio.

El Laboratorio Pura Vida es un laboratorio clínico que se encuentra en el mercado desde el 2006; da servicios de exámenes de laboratorio clínico, salud ocupacional, seguridad e higiene industrial y gestión ambiental a la sociedad, se contabilizan más de 100 pruebas de laboratorio en las áreas de Química Clínica, Hormonas, Hematología clínica, Serología, Urianálisis, Coproanálisis, Inmunología, Microbiología, Marcadores Tumorales, Infecciosas, Autoinmunidad, Coagulación, Anatomía Patológica, marcadores Cardiovasculares, Drogas, Pruebas Especiales entre otros. En cada una de estas áreas se dispone de las pruebas de laboratorio específicas.

La misión de Pura vida es:

"Brindar excelencia en servicios de salud integral, mediante un trabajo honesto, eficaz y oportuno. Combinando de manera óptima el recurso humano y tecnológico, para lograr que nuestros clientes alcancen un excelente estado de bienestar."

Su visón es:

"Llegar a ser una de las principales empresas proveedoras de servicios integrados a nivel local y nacional, tanto en el área de Medicina Ocupacional como en el área de Medicina Preventiva, a través de la implementación de programas que brinden óptima calidad, tanto a médicos, pacientes, empresas y público en general."

Para estar alineados a su misión y visión el laboratorio ha invertido en infraestructura tecnológica necesaria para garantizar calidad y precisión en los resultados obtenidos en sus pruebas aplicadas; para esto corren controles de calidad continuamente. Más del 70 % de sus pruebas son realizadas con el uso de tecnología solo en casos especiales realizan las pruebas manualmente.

4.1.2 Organización del laboratorio

El laboratorio cuenta con las siguientes áreas:

- Gerencia / Propietario
- Contabilidad
- Tecnología
- Procesamiento

- Servicios generales
    - Mensajería
    - Limpieza
- Recepción

Las áreas que generan la información y que están involucrados para este proyecto son:

- Gerencia
- Tecnología
- Procesamiento

4.1.3 Problemática a resolver

La situación económica del país ha ocasionado que las personas dediquen muchas horas del día a su trabajo, esto ha llevado a establecer malos hábitos alimenticios, a realizar poca o ninguna actividad física, causando problemas en la salud que deben ser identificados a través de pruebas de laboratorio, siendo este servicio necesario y prioritario para las personas, el laboratorio en estudio no ha presentado un crecimiento y una participación en el mercado que le permita una tranquilidad económica.

El laboratorio dispone de información de sus exámenes aplicados y facturados desde el 2006 fecha de inicio de sus actividades sin embargo no se ha realizado nada con esos datos. Su Gerente busca que con el análisis de esta información se pueda identificar condiciones, eventos o situaciones que orienten a la toma de decisiones para mantener e incrementar los clientes del laboratorio.

4.1.4 Objetivos del negocio

Los objetivos planteados en base a la situación y problemática del laboratorio son:

- Crear grupos de clientes en base a su comportamiento en el pedido de exámenes
- Identificar características de los pacientes que llegan a realizarse un examen de laboratorio

4.1.5 Criterios de éxito

La información generada por la facturación de los servicios de laboratorio y obtenidos del sistema **SisGem** será utilizado para la agrupación de clientes en base a su comportamiento de compra. Hay que tomar en cuenta que el cliente al que se le factura el servicio no siempre es el paciente.

La agrupación de estos pacientes nos permitirá establecer el nivel de lealtad de estos. Para poder obtener este cruce de información entre la facturación a los clientes y los exámenes realizados por los pacientes se tuvo que extraer información

del sistema **CLINICAL Lab** el cual contiene los pedidos de exámenes y los exámenes realizados. La integración entre estos dos sistemas se da por el número de la orden de pedido.

El acceso a los datos se realizará a través de los reportes del sistema, el soporte por parte de los propietarios del sistema para entender la estructura relacional de la base de datos es facturada, la Gerencia del laboratorio no dispone de presupuesto para esta actividad.

4.1.6 Riesgos y contingencias

Los riesgos y contingencias que se analizan son los relacionados con la obtención, manipulación y limpieza de los datos; se busca que cualquier acción que lleve a una modificación o alteración de la información sea consciente y evaluada.

Pérdida de información: los datos originales se mantendrán en los archivos csv o xlsx obtenidos de los sistemas, para mantener su integridad se montarán estos datos a las tablas construidas en ORACLE-XE para el proceso ETL.

Problemas con la integridad de los datos obtenidos: se obtendrán todos los posibles reportes del sistema que nos permitan identificar columnas de clave primaria que nos permitan integrar la información y unificarla en tablas finales para el análisis.

Datos Faltantes: en función del análisis de la integridad de los datos se identificarán los datos faltantes y se solicitarán reportes nuevos al propietario del software en base al presupuesto de este proyecto.

Consistencia de la información: se realizarán análisis muestrales para identificar que los datos tengan consistencia para esto se planificarán reuniones continuas con la gerencia.

4.2 Aplicación de las técnicas de Minería de Datos para la segmentación de clientes

Esta actividad se realiza en base a la metodología CRISP-DM y comprenda fases como comprensión de los datos, preparación de los datos, modelado y evaluación.

4.2.1 Comprensión de los datos

La comprensión de los datos se realiza en base a varias actividades como recopilación, descripción, exploración y verificación de la calidad de los datos.

4.2.2 Recopilación de los datos iniciales

Se recopilaron los datos a través de los reportes que proporcionan los sistemas del laboratorio; los datos se obtuvieron desde el 2.017 al 2.021 en archivos

csv y xlsx; estos corresponden a información de clientes (reiterando que el cliente no siempre es el paciente), registros de facturación (en 2 diferentes reportes), órdenes de pedidos de exámenes.

Se analizó la información obtenida del sistema para identificar un atributo integrador de la información que nos permita construir un archivo integrado para el análisis de datos; y un modelo que nos permita mayor comprensión de sus relaciones.

Los archivos de los reportes obtenidos se relacionaron y permitieron construir el modelo de datos como se muestra en la Figura 5.

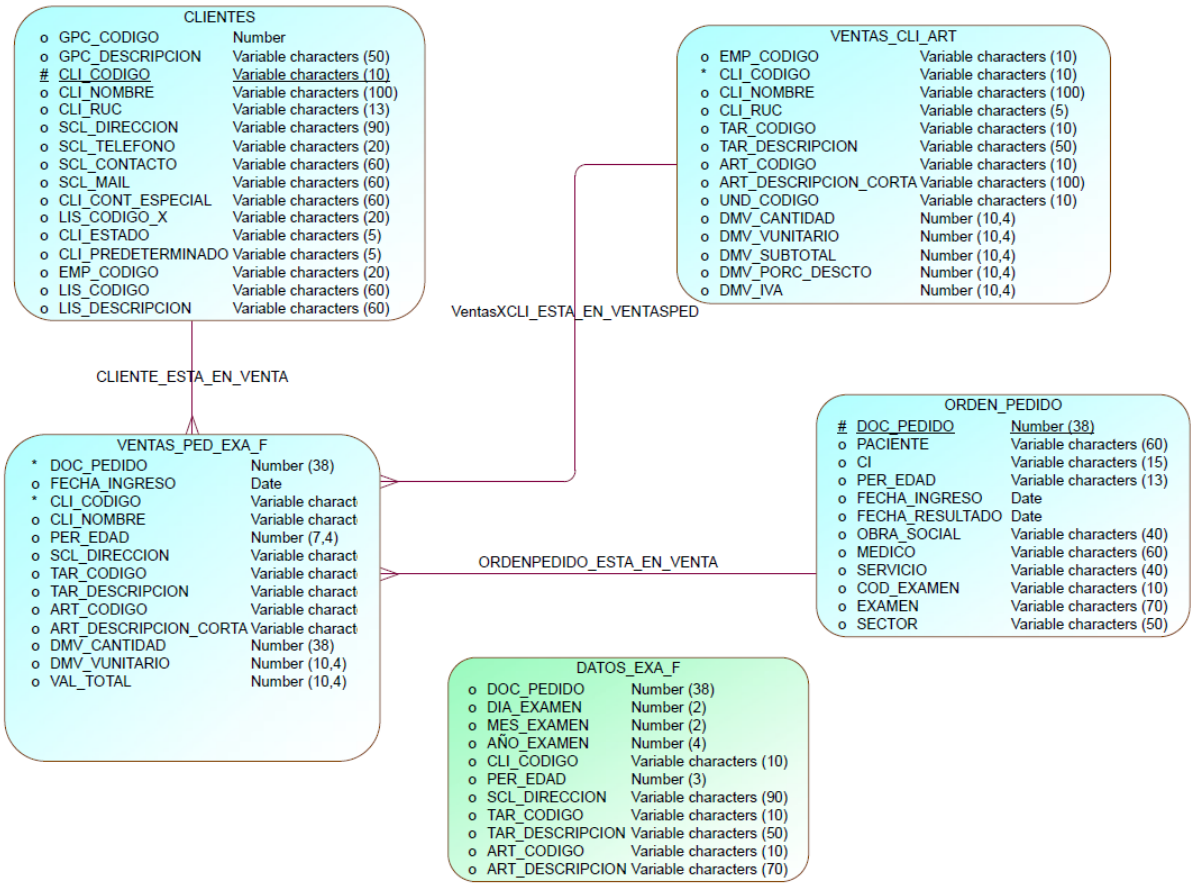


Figura 5: Modelo de Datos

4.2.3 Descripción de los datos

De los sistemas de información de los laboratorios se obtuvieron los siguientes archivos xlsx:

Tabla de CLIENTES: El archivo original xlsx contiene 10.966 registros con la información de clientes del laboratorio, de este archivo se subió a la tabla clientes todos los registros en la misma estructura.

ATRIBUTO	DESCRIPCIÓN
GPC_CODIGO	N/A
GPC_DESCRIPCIÓN	N/A
CLI_CODIGO	Código del cliente (no siempre el paciente)
CLI_NOMBRE	Nombre del cliente
CLI_RUC	Cédula o RUC del cliente
SCL_DIRECCIÓN	Dirección del cliente
SCL_TELEFONO	N/A
SCL_CONTACTO	N/A
SCL_MAIL	N/A
CLI_CONT_ESPECIAL	N/A
LIS_CODIGO_X	N/A
CLI_ESTADO	N/A
CLI_PREDETERMINADO	N/A
EMP_CODIGO0	N/A
LIS_CODIGO	N/A
LIS_DESCRIPCION	N/A

Tabla 1: Estructura archivo CLIENTES

Tabla VENTAS_CLI_ART: el archivo original xlsx dispone de 24.383 registros con información referente a las pruebas de laboratorio que han sido facturadas a los clientes, aquí es importante recordar que los clientes no necesariamente son los pacientes. Todos los registros fueron subidos a la tabla

ATRIBUTO	DESCRIPCIÓN
EMP_CODIGO	N/A
CLI_CODIGO	Código del cliente (no siempre el paciente)
CLI_NOMBRE	Nombre del cliente
CLI_RUC	Cédula o RUC del cliente
TAR_CODIGO	Código del área de exámenes
TAR_DESCRIPCION	Descripción del área de exámenes
ART_CODIGO	Código del artículo o prueba (examen)

ART_DESCRIPCION_CORTA	Descripción del artículo (examen)
UND_CODIGO	N/A
DMV_CANTIDAD	Cantidad de exámenes facturados
DMV_VUNITARIO	Precio unitario del examen
DET_MOVIMIENTO_DMV_SUBTOTAL	N/A
DET_MOVIMIENTO_DMV_PORC_DSCTO	N/A
DET_MOVIMIENTO_DMV_IVA	N/A

Tabla 2: Estructura archivo VENTAS_CLI_ART

Tabla ORDEN_PEDIDO: el archivo original xlsx dispone de 129.148 registros con información referente a los exámenes realizados a cada paciente, el área a la que pertenece el examen, la fecha del pedido, la edad del paciente. Todos los registros fueron subidos a la tabla

ATRIBUTO	DESCRIPCIÓN
ORDEN	Número de pedido de exámenes (DOC_PEDIDO)
PACIENTE	Nombre del paciente
CI	Cédula de identidad del paciente
PER_EDAD	Edad del paciente
FECHA_INGRESO	Fecha del pedido de examen
FECHA_RESULTADO	N/A
OBRA_SOCIAL	N/A
MEDICO	N/A
SERVICIO	N/A
COD_EXAMEN	Código del examen realizado
EXAMEN	Nombre del examen realizado
SECTOR	Área o sector al que pertenece el examen realizado

Tabla 3 Estructura archivo ORDEN_PEDIDO

Tabla VENTAS_PED_EXA_F: esta tabla fue reconstruida con información integrada de las dos tablas relacionadas que se muestra en el modelo.

ATRIBUTO	DESCRIPCIÓN
DOC_PEDIDO	Número de pedido de exámenes
FECHA_INGRESO	Fecha del pedido de examen

CLI_CODIGO	Código del paciente
CLI_NOMBRE	Nombre del paciente
PER_EDAD	Edad del paciente
SCL_DIRECCION	Dirección del paciente
TAR_CODIGO	Código del área del examen
TAR_DESCRIPCION	Descripción del área del examen
ART_CODIGO	Código del examen
ART_DESCRIPCION_CORTA	Descripción del examen
DMV_CANTIDAD	Cantidad de exámenes realizados
DMV_VUNITARIO	Precio unitario del examen realizado
VAL_TOTAL	Valor total de los exámenes realizados

Tabla 4 Estructura Tabla VENTAS_PED_EXA_F

Tabla DATOS_EXA_F: esta tabla se creó como repositorio definitivo de los datos, base para la generación del archivo csv que se usará en el modelado

ATRIBUTO	DESCRIPCIÓN
DOC_PEDIDO	Número de pedido de exámenes
DIA_EXAMEN	Día de la fecha del pedido de examen
MES_EXAMEN	Mes de la fecha del pedido de examen
AÑO_EXAMEN	Año de la fecha del pedido de examen
CLI_CODIGO	Código del paciente
PER_EDAD	Edad del paciente
SCL_DIRECCION	Dirección del paciente
TAR_CODIGO	Código del área del examen
TAR_DESCRIPCION	Descripción del área del examen
ART_CODIGO	Código del examen
ART_DESCRIPCION_CORTA	Descripción del examen

Tabla 5 Estructura tabla DATOS_EXA_F

4.2.4 Exploración de los datos

Para un mejor análisis de los datos, el autor de este proyecto optó por subir la información a la base de datos ORACLE-XE para que a través del uso del PLSQL obtener información general del estado inicial de estos datos.

El archivo de clientes que se extrajo del sistema dispone de 7.520 clientes del laboratorio, cada uno de estos registros contienen información cualitativa de los clientes. En primera instancia se debe trabajar en una depuración y transformación ya que existen clientes sin información clara, las edades con la palabra años, meses y días, las direcciones de forma específica en algunos casos en otros sin dirección.

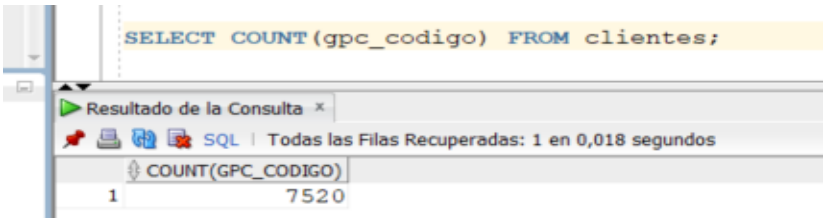


Figura 4 Total de registros de Clientes

El archivo de exámenes no se consideró en el proceso de carga porque el resto de archivos ya tenían atributos del código y descripción de los exámenes, sin embargo, sirvió para identificar que el laboratorio realiza 203 tipos diferentes de exámenes cada uno de estos agrupados en un total de 20 áreas de exámenes. Los exámenes más demandados son el de Biometría Hemática, Elemental y microscópico de orina y Glucosa en ayunas

Figura 5 Total de tipos de exámenes

El archivo de ventas por examen dispone de 182.425 registros en este archivo se dispone de la cédula de identidad o RUC del cliente, pero no se dispone de la orden de pedido de exámenes para integrar con los exámenes realizados, recalco nuevamente que el paciente no siempre es el cliente.

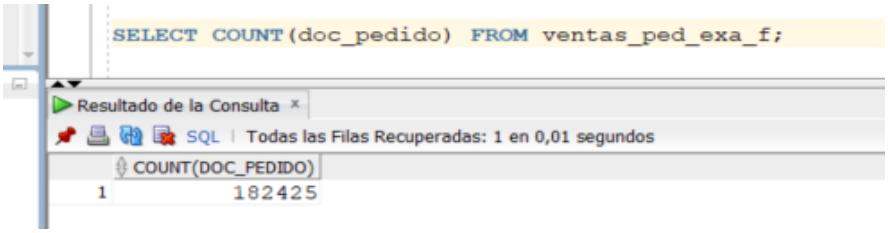


Figura 6 Total de registros de Ventas

El archivo de órdenes de pedido dispone de las órdenes de pedido y de los exámenes solicitados a esta orden, la integridad con las ventas se debe realizar a través del campo cédula o del nombre del paciente. Este archivo dispone de 129.147 registros.

Figura 7 Total de órdenes de pedido

4.2.5 Verificación de la calidad de los datos

La base de datos del laboratorio posee datos categóricos de gran variedad que podría complicar el análisis de agrupamiento que se puede realizar en este estudio.

El archivo de clientes específicamente en el atributo de dirección debe ser transformado de una dirección específica a un sector de la ciudad para disminuir el rango de alternativas; es necesario establecer sectores en función de la necesidad del laboratorio.

La edad del paciente que se encuentra en el archivo de órdenes de pedido es de tipo texto ya que para el reporte de los resultados de los exámenes es necesario especificar si son años, meses o días que tiene el paciente; esta variable categórica debe ser transformada a una variable numérica.

Se analizaron dos reportes más emitidos por el sistema para definir consistencia, pertinencia y confiabilidad de la información entregada, estos dos archivos hacen referencia a un informe general de ventas y a un informe de ventas con o sin factura; identificando que la diferencia de información corresponde a facturación a convenios y empresas; esta facturación es especial ya que los pacientes no son del laboratorio, son del convenio o de la empresa, en este caso no se los crea como pacientes y clientes del laboratorio, los pacientes y clientes son los convenios y las empresas. Es necesario un análisis y depuración de esta información.

La información se encuentra dispersa en los diferentes archivos, es necesario un análisis de integridad que permita identificar atributos relacionales y generar un solo archivo de datos para el análisis.

4.2.6 Preparación y muestreo de los datos

Para una adecuada preparación y muestreo de los datos es necesario realizar varias actividades de selección de los datos, limpieza de los datos, construcción de nuevos datos si es necesario, integración de estos datos y definición de formatos. Estas actividades no se realizaron en forma secuencial, se fueron realizando conforme se identificaba la necesidad de transformación o limpieza; esto no afectó el proceso ya que las tablas creadas no responden a una integridad referencial.

4.2.6.1 Selección de los datos

Para la ejecución de esta actividad se considera los objetivos del negocio y los criterios de éxito definidos para este proyecto (ver puntos 4.1.4 y 4.1.5).

De los registros de clientes se han descartado los registros que hacen referencia a los convenios y empresas, así como el cliente “Consumidor Final”.

De la tabla de clientes solo se seleccionó los atributos de código del cliente, nombre del cliente y su dirección, el resto de atributos no son necesarios para este análisis

De los registros de ventas se han seleccionado los registros que corresponden a pacientes del laboratorio, descartando del análisis los registros de ventas que hacen referencia a convenios y empresas; se eliminó también todos los registros de ventas relacionados con el cliente “Consumidor Final”

De la tabla de ventas a clientes se seleccionó los atributos como código del paciente, nombre del paciente, cédula del paciente, código del área del examen aplicado, descripción del área, código del examen, descripción del examen, cantidad de exámenes realizados y el valor unitario del examen.

De los registros de órdenes de pedido se eliminaron todos los registros que hacen referencia en el campo cédula, al código del convenio o empresa; se eliminaron también registros de órdenes de pedido realizadas a niños que no reportan el número de cédula y en su lugar se registra la cédula del representante adicionando un 02 al final de la cédula, adicionalmente porque son casos mínimos en relación a la data.

De la tabla de órdenes de pedido se seleccionaron los atributos de número de orden, nombre del paciente, cédula de identidad del paciente, edad del paciente, fecha del pedido, el código del examen solicitado, la descripción del examen y la descripción del área del examen.

4.2.6.2 Limpieza de los datos

Las actividades requeridas para la limpieza de los datos fueron realizadas con PLSQL en la herramienta SQLDEVELOPER de ORACLE; todas estas actividades se realizaron en tablas temporales creadas en el DBMS ORACLE para luego integrarlas en las tablas definitivas; luego; con esta misma herramienta, se exportó a formato csv la tabla en la que se integran todos los datos para el análisis.

Las direcciones de los clientes en el archivo original son muy específicas como se muestra en la figura de ejemplos de direcciones de clientes; estas generan una gran variedad de alternativas de ubicación (alta dimensionalidad). Se procede a establecer sectores con la gerencia y se identifica un mapa que se ajuste de la mejor manera a las necesidades de sectorización del laboratorio (ver Figura 6).

GPC_CODIG	GPC_DESCR	CLI_CODIGO	CLI_NOMBRE	CLI_RUC	SCL_DIRECCION
1	PUBLICO GE	6108	ALVAREZ MARTHA	0918280249	CENTRO HISTORICO
1	PUBLICO GE	6742	ALVAREZ MESA ARACELY IVETH	1600305237	CALIFORNIA ALTA
1	PUBLICO GE	9067	ALVAREZ MONTALVO SIXTA XIMENA	1715988018	EFRAIN ARMAS Y VIRGEN DE MONSERRAT
1	PUBLICO GE	9780	ALVAREZ MUÑOZ OSCAR EMILIO	1713750485	LA MAGDALENA OE 6-391 Y ZAMORA
1	PUBLICO GE	3542	ALVAREZ NARVAEZ JULIA INES	1711453892	AV 10 DE AGOSTO Y AV 6 DE DICIEMBRE
1	PUBLICO GE	7851	ALVAREZ NARVAEZ OSWALDO DAVID	1717535122	10 DE AGOSTO Y 6 AV. 6 DE DICIEMBRE
1	PUBLICO GE	9375	ALVEAR ACOSTA ANA MARIA	1710572452	EL CONDADO
1	PUBLICO GE	9376	ALVEAR ACOSTA XIMENA DE LOS ANGELES	1706019955	EL CONDADO
1	PUBLICO GE	10888	ALVEAR CABRERA JOSE LUIS	0401191564	AV. REAL AUDIENCIA Y MURIALDO
1	PUBLICO GE	7060	ALVEAR CALDERON MARIA CRISTINA	1712215043	EL QUICENTRO
1	PUBLICO GE	9524	ALVEAR HERRERA LUIS RAMIRO	1709134447	CARAPUNGO AV. LUIS PACARI Y PASAJE MACHA LILIA N6-199
1	PUBLICO GE	6471	ALVEAR OLMEDO MERCEDES ANDREA	1715424626	LA KENEDY
1	PUBLICO GE	2367	ALVEAR ROMERO VICTOR MANUEL	1704890829	APARICIO RIVADENEIRA E3-31

Figura 8 Ejemplo de direcciones de clientes

Se procede a identificar el sector al que pertenece cada dirección haciendo uso del GOOGLE MAPS y sustituyendo en la tabla clientes por el sector identificado en el mapa; existen clientes de provincia que reportan su provincia como dirección, a estos se les mantuvo esa dirección, de la misma manera sucede con pacientes de los estados Unidos de América, en total se registran 66 sectores incluidos los 32 del mapa; la instrucción de PLSQL utilizada para esta sustitución es:

```
SQL> UPDATE clientes SET scl_direccion = 'KENNEDY' WHERE scl_direccion
like '%RAMON BORJA%';
```

En el mismo análisis se eliminaron clientes con direcciones nulas y con errores e inconsistencias en la dirección.

La edad de los pacientes registrada en la orden de pedido, contienen texto como “AÑOS”, “MESES” y “DIAS” esto debido a que al momento de reportar los resultados del paciente es necesario reportar la edad con esta aclaración como parámetro importante para los médicos, este atributo se debe corregir eliminando estas palabras en el atributo PER_EDAD.

Para modificar las edades es necesario primero identificar si es “AÑOS” entonces solo se selecciona los dos primeros caracteres, si es “MESES” estos dos caracteres deben ser divididos entre 12 y si son “DIAS” debe ser dividido entre 365.

```
SQL> INSERT INTO EXAMENES_F SELECT ORDEN, substr(per_edad,1,2),
fecha_ingreso, COD_EXAMEN, EXAMEN, SECTOR   FROM EXAMENES WHERE
per edad like '%AÑOS%';
```

```
SQL> INSERT INTO EXAMENES_F SELECT ORDEN, substr(per_edad,1,2)/12,
echa_ingreso, COD_EXAMEN, EXAMEN, SECTOR FROM EXAMENES WHERE per_edad
like '%MESES%';
```

```
SQL> TO EXAMENES_F SELECT ORDEN, substr(per_edad,1,2)/365,
fecha_ingreso, COD_EXAMEN, EXAMEN, SECTOR FROM EXAMENES WHERE per_edad
like '%DIAS%';
```

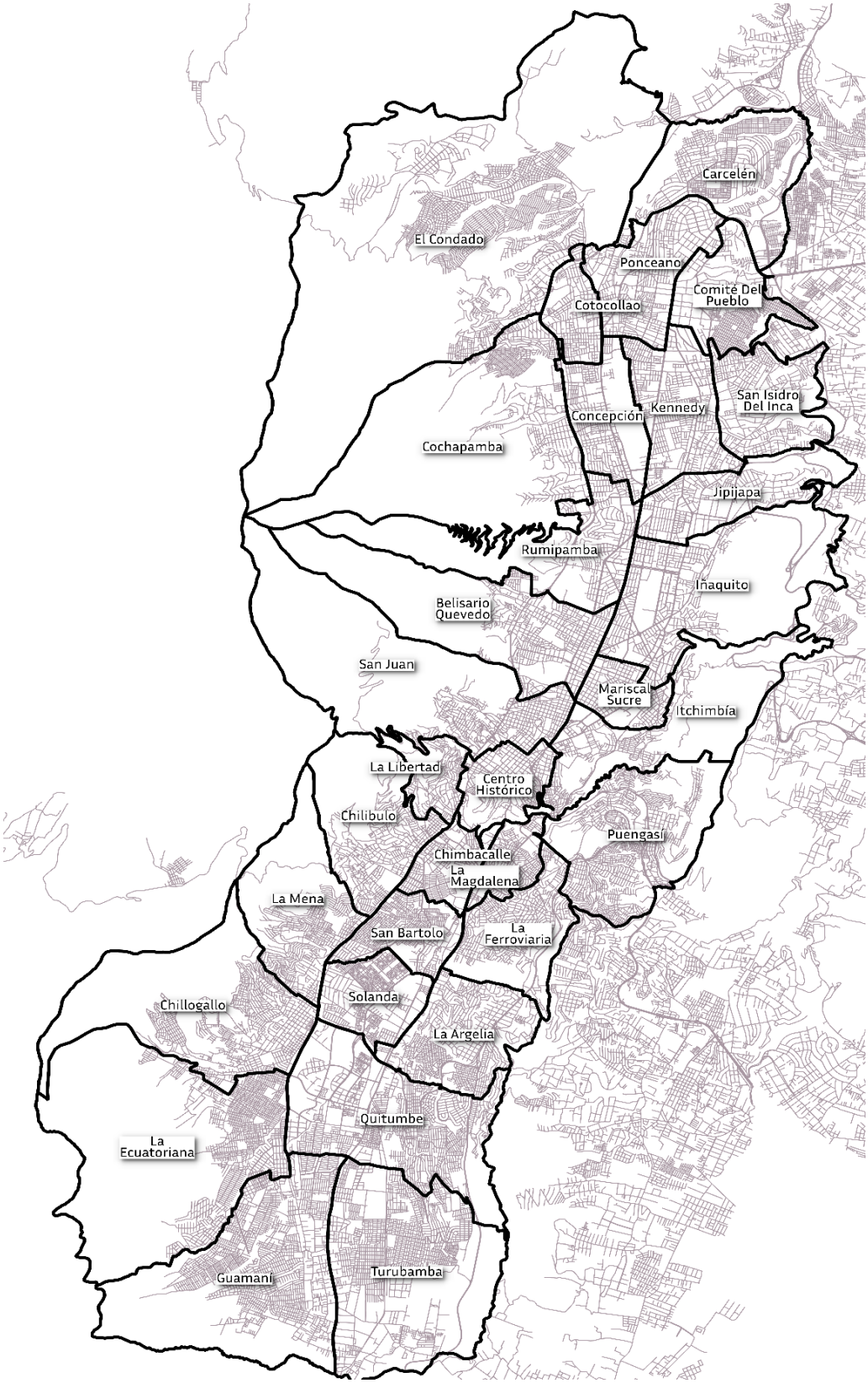


Figura 9 Sectores en el mapa de Quito

Fuente: https://es.wikipedia.org/wiki/Quito#/media/Archivo:Parroquia_Urbanas3.png

Se eliminan registros de las órdenes de pedido con edad del paciente nulas.

Entre los diferentes tipos de exámenes que se realizan en el laboratorio se identifica un examen con descripción “EXAMENES REALIZADOS”, al consultar con el analista de laboratorio se informa que fue erróneamente creado para ingresar exámenes PCR debido a que el laboratorio no los realiza. En esta misma descripción se incluyen los exámenes especiales que se realizan a los convenios y empresas por esta razón no se pudo diferenciar los PCR. Estos registros fueron eliminados de la tabla.

```
SQL> DELETE FROM EXAMENES WHERE examen = 'EXAMENES REALIZADOS';
```

Se identifican en la tabla de ventas registros duplicados, se consulta con la gerencia y se definen como registros que el sistema genera con el fin de controlar los materiales y reactivos que se utilizan en cada prueba (examen) y que deben ser descargados del inventario; todos estos registros se eliminan de la tabla. La eliminación de estos registros se complica pues son registros exactos en su totalidad, es decir de 3, 4 y en algunos casos están 5 registros duplicados se debe considerar solo 1; para eliminar estos registros no se puede utilizar la instrucción DELETE; se debe crear una función en PLSQL que identifique que: si ya se insertó un registro de estos a la tabla, no permita insertar otro. Evitando el ingreso de registros duplicados. El código desarrollado es el siguiente:

```
CREATE OR REPLACE FUNCTION  fn_cursor RETURN SYS_REFCURSOR
IS
      TYPE ref_cursor IS REF CURSOR;
        v_ref ref_cursor;
BEGIN
     OPEN v_ref FOR SELECT DOC_PEDIDO, FECHA_INGRESO, CLI_CODIGO, PER_EDAD,
     SCL_DIRECCION, TAR_CODIGO, TAR_DESCRIPCION, ART_CODIGO,
     ART_DESCRIPCION_CORTA FROM ventas_ped_exa_f;
     RETURN v_ref;
END;
/
```

```
SET SERVEROUTPUT ON;
DECLARE
    ref_cursor SYS_REFCURSOR;
    TYPE ventas_record IS RECORD(
        DOC_PEDIDO            NUMBER(38),
        FECHA_INGRESO         DATE,
        CLI_CODIGO            VARCHAR2(10),
        PER_EDAD              VARCHAR2(13),
        SCL_DIRECCION         VARCHAR2(90),
        TAR_CODIGO            VARCHAR2(10),
        TAR_DESCRIPCION       VARCHAR2(50),
        ART_CODIGO            VARCHAR2(10),
        ART_DESCRIPCION_CORTA VARCHAR2(70)
        );
    ventas ventas_record;
    cont NUMBER;
BEGIN
    ref_cursor := fn_cursor;
    LOOP
        FETCH ref_cursor INTO ventas;
        EXIT WHEN ref_cursor%NOTFOUND;
        SELECT COUNT(doc_pedido) INTO cont from datos_exa_f
        WHERE dia_examen = EXTRACT(DAY FROM ventas.fecha_ingreso)
        AND mes_examen = EXTRACT(MONTH FROM ventas.fecha_ingreso)
        AND año_examen = EXTRACT(YEAR FROM ventas.fecha_ingreso)
        AND cli_codigo = ventas.cli_codigo
        AND art_codigo = ventas.art_codigo;
        if cont = 0 THEN
            INSERT INTO datos_exa_f VALUES(ventas.doc_pedido, EXTRACT(DAY
                  FROM ventas.fecha_ingreso),
                  EXTRACT(MONTH FROM ventas.fecha_ingreso), EXTRACT(YEAR
                  FROM ventas.fecha_ingreso), ventas.cli_codigo,
                  ventas.per_edad, ventas.scl_direccion,ventas.tar_codigo,
                  ventas.tar_descripcion, ventas.art_codigo,
                  ventas.art_descripcion_corta);
        end if;
    END LOOP;
END;
```

4.2.6.3Integración de datos

Para la generación del archivo definitivo se integró la información de “ventas por cliente y artículo” con las “ventas con factura” y las “órdenes de pedido”; esta integración permite obtener los exámenes realizados por paciente descartando los registros que hacen referencia a clientes facturados.

```
SQL> INSERT INTO ventas_ped_exa SELECT a.doc_pedido, c.fecha_ingreso,
b.cli_codigo, b.cli_nombre, c.per_edad, b.tar_codigo, b.tar_descripcion,
b.art_codigo, b.art_descripcion_corta, b.dmv_cantidad, b.dmv_vunitario,
(b.dmv_cantidad*b.dmv_vunitario) AS TOTAL
FROM ventas_con_sin_fac a, ventas_cli_art b, orden_pedido c
WHERE a.doc_cliente_final = b.cli_nombre
AND c.orden=a.doc_pedido
AND c.paciente=b.cli nombre;
```

Esta integración inicial en la que se obtuvieron los pacientes y sus exámenes, se cruzó con la tabla clientes para obtener los sectores de ubicación de los pacientes.

```
SQL> INSERT INTO ventas_exa_sec_f SELECT a.doc_pedido, a.doc_fecha,
a.cli_codigo, a.cli_nombre, b.scl_direccion, a.tar_codigo,
a.tar_descripcion, a.art_codigo, a.art_descripcion_corta, a.dmv_cantidad,
a.dmv_vunitario, val_total
FROM ventas_ped_exa a, clientes b
WHERE a.cli_codigo = b.cli_codigo;
```

Con esta información integrada se generó la tabla definitiva que se exportará a formato csv con la cual se realizará el análisis.

```
SQL> INSERT INTO datos_exa_f SELECT EXTRACT(DAY FROM fecha_ingreso),
EXTRACT(MONTH FROM FECHA_INGRESO), EXTRACT(YEAR FROM fecha_ingreso),
Cli_codigo, per_edfad, scl_direccion, tar_codigo, tar_descripcion,
art_codigo, art_descripcion_corta
FROM vesntas_exa_sec_f;
```

4.2.7 Realización del modelo

Se inició el proceso de análisis haciendo uso de BigML. Este dashboard facilita la visualización e identificación de indicadores claves(PKI), métricas y datos fundamentales que permitan conocer el estado de la información y de la empresa.

4.2.7.1Agrupamiento

Actividades realizadas:

1. Se montó a BigML el archivo csv DATOS_EXA_F
2. Se cambiaron los tipos de datos de los atributos DOC_PEDIDO, AÑO_EXAMEN, CLI_CODIGO, TAR_CODIGO, ART_CODIGO de numérico a categóricos definiéndolos como valores discretos y se actualizó el dataset
3. Generamos el dataset con la opción de configure dataset con el nombre de DATOS_EXA_F_DS
4. BigML identificó a las variables SCL_DIRECCION, ART_CODIGO y ART_DESCRIPCION_CORTA como no preferida esto se debe a que estas variables son identificativas en su conjunto, es decir entre las tres variables definen un registro único
5. Identificamos que en la data no existen valores perdidos ni errores

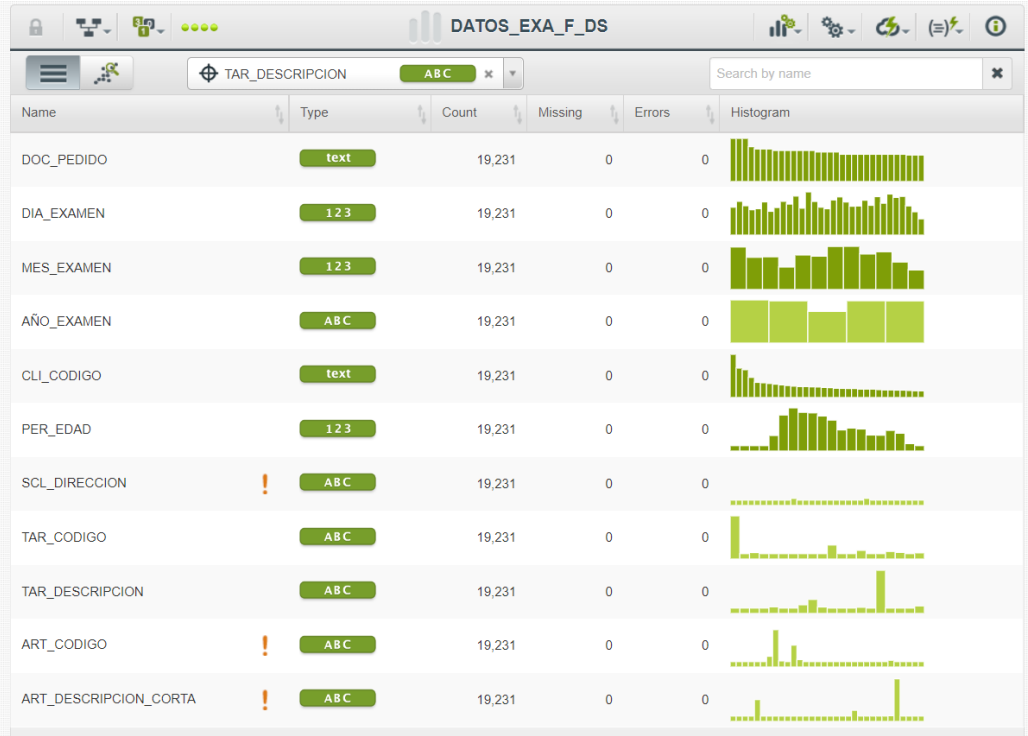


Figura 10 Estructura de la DATA en BigML

En estos datos se puede visualizar los diagramas de frecuencias de cada variable, se evidencian variables con datos dispersos.

Podemos identificar que el mes de diciembre son los meses más bajos para el laboratorio y enero julio y agosto los más altos; que para el laboratorio el año 2019 fue un año de menor producción en relación al 2017, 2018, 2020 y 2021; las edades en las que más exámenes se realizan son de los 25 a los 55 años; el sector con mayor cantidad de exámenes realizados es KENNEDY. Se puede analizar también

los estadísticos básicos de cada una de estas variables; de acuerdo al punto de vista de la gerencia, esto se debe a que el laboratorio se encuentra en ese sector; los sectores que le siguen son los de SAN ISIDRO DEL INCA y PONCEANO definiendo como oportunidad para poner sucursales del laboratorio.

Se puede identificar en la Figura 8 que personas de todas las edades se realizan con mayor frecuencia exámenes del área de Química Clínica

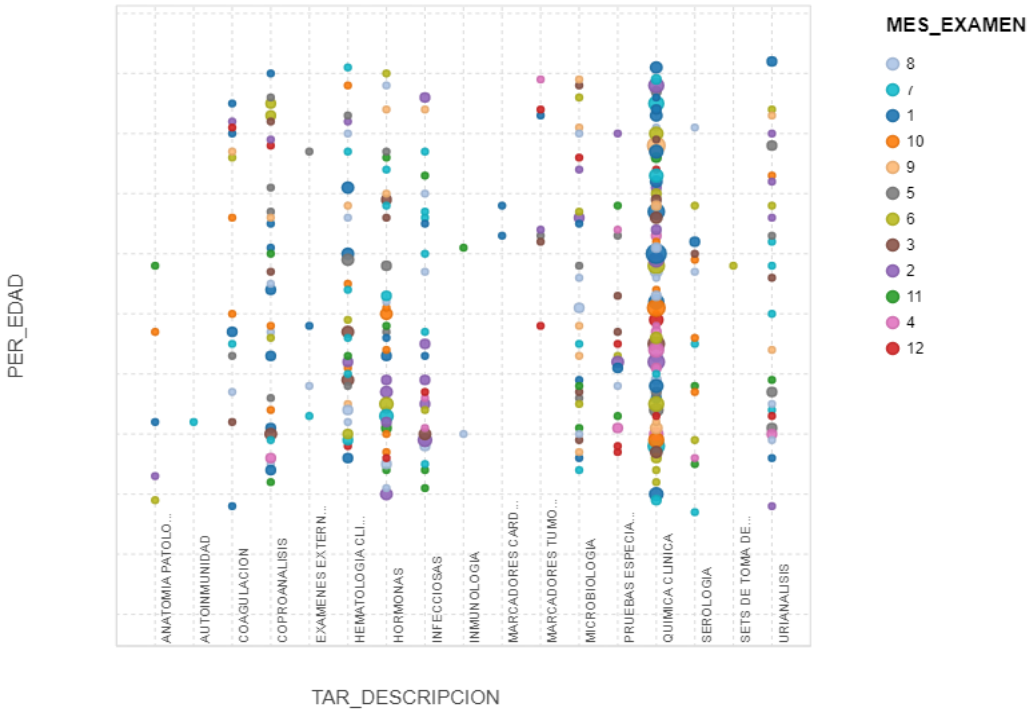


Figura 11 Exámenes de áreas por edad

Esta información fue analizada por médico especialistas (análisis de expertos), los cuales validan en función de estudios realizados que el sedentarismo y la mala alimentación de las nuevas generaciones ha ocasionado que exámenes que históricamente se realizan en su mayoría a personas mayores de 50 años hoy se estén aplicado a jóvenes desde los 19 años.

Para la agrupación se utilizaron en BigML dos modelos, el K-means y el G-means.

Se realizó un primer ensayo con un valor de K=5 (ver Figura 9), no son claros los clústeres obtenidos.

Se trabajó con el conjunto de datos del cluster-0 de este primer agrupamiento (ver Figura 10), con el fin de tener algún dato que permita entender cuáles son las condiciones de agrupación del modelo

En primera instancia se podría considerar que se está clasificando por área del examen ya que la mayor concentración está en Química Clínica

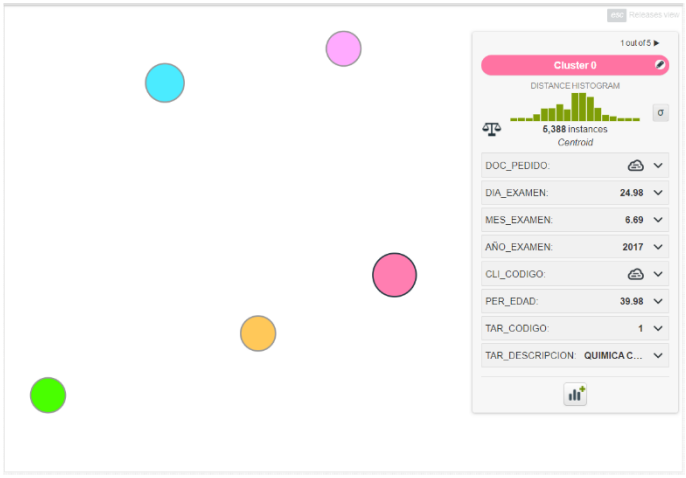


Figura 12 Clusters K-means K=5

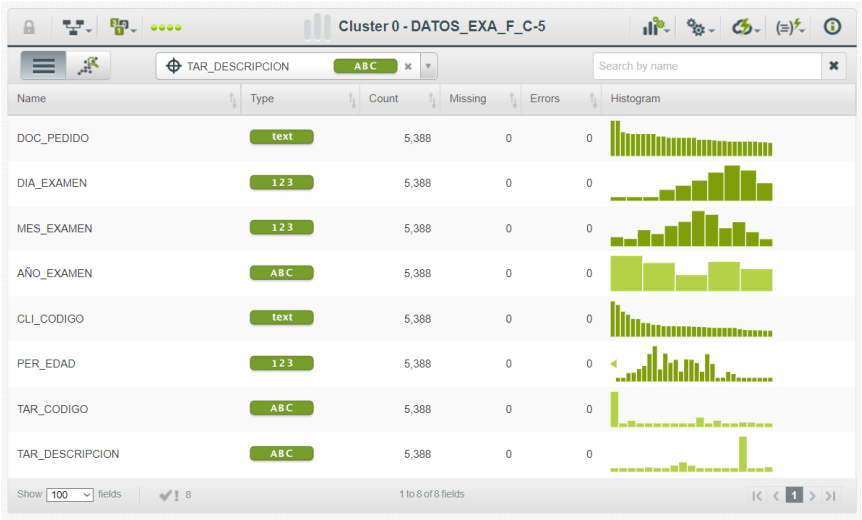


Figura 13 Distribución con la data del CLUSTER-0

Se probó con K=8, los clústeres siguen sin ser claros

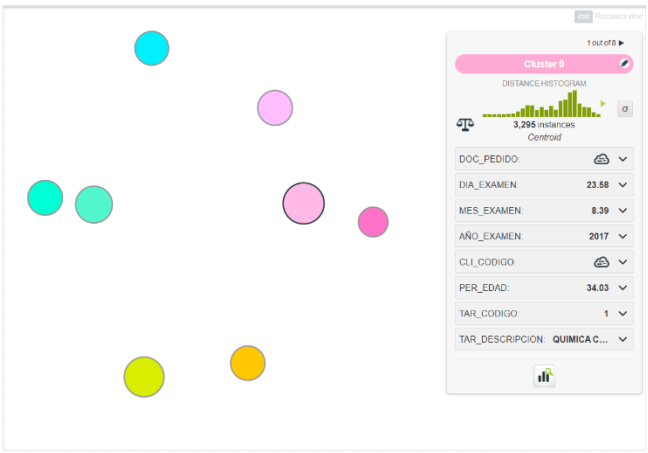


Figura 14 Clusters K-means K=8

G-means es un algoritmo más avanzado, es jerárquico y realiza subdivisiones de 2 en 2 de los datos usando K-means, realiza este proceso hasta que se dejan de cumplir ciertos criterios estadísticos establecidos por un critical values que es un parámetro que indica si queremos más o menos grupos

Se realizó un primer ensayo con un valor de critical values = 5, las agrupaciones obtenidas no son claras, presenta solo dos clústeres.

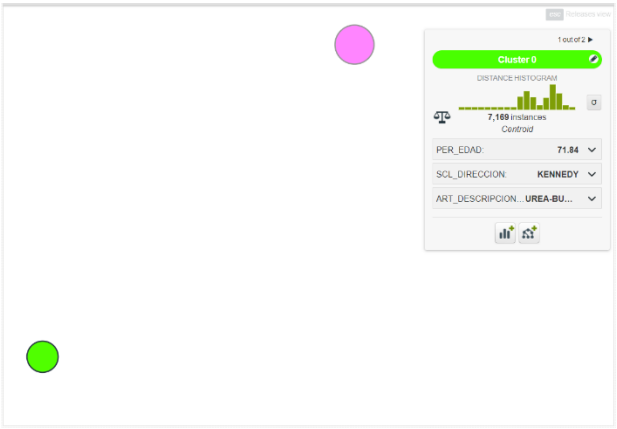
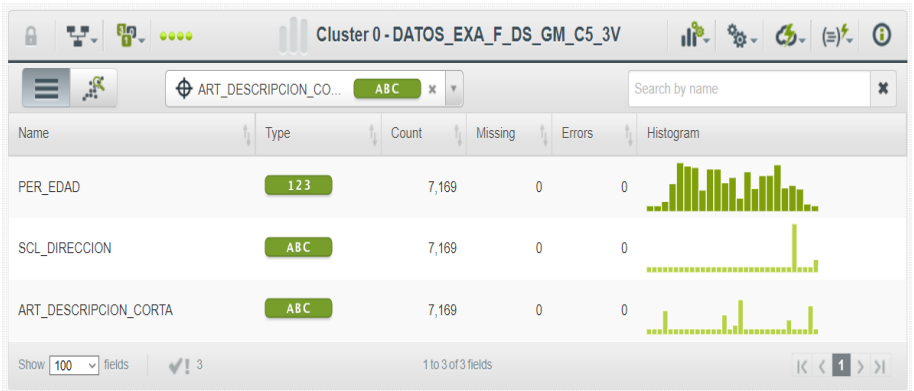


Figura 15 Clusters G-means CV=5

De igual forma se trabaja con el conjunto de datos del cluster-0 para verificar alguna característica de clasificación



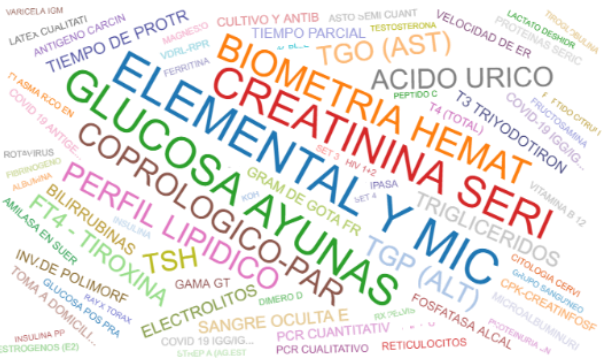


Figura 16 Distribución con la data del CLUSTER-0 en G-means

Con los resultados obtenidos, lo ideal sería aplicar la prueba del codo para identificar el valor ideal de K, sin embargo, el hecho de disponer de datos discretos y datos de alta dimensión se considera como mejor alternativa aplicar el método K-modes; sin embargo, por temas de validación se aplicó la prueba del codo como se observa en el ANEXO 3.

Entre las diferencias importantes entre K-means y K-modes que podemos citar tenemos: sustituir la media por la moda, reemplazar la distancia euclidiana por la medida de disimilitud de coincidencia simple (distancia Hamming); con este modelo se esperaría obtener agrupaciones más claras. Este modelo se aplicó utilizando Python cuyo código se presenta en el ANEXO 1. Los clústeres siguen sin ser claros.

Se probó también DBSCAN, modelo basado en densidad, son modelos muy eficientes para encontrar agrupaciones de alta densidad y valores atípicos porque no solamente se preocupan en etiquetar los datos categóricos, sino también en usar procedimientos diferentes de cálculo de la distancia entre los elementos; en este caso usaremos la distancia de Gower para calcular la similitud o disimilitud entre observaciones. El resultado de la aplicación de este modelo está en el ANEXO 2. Los clústeres siguen siendo confusos

4.2.7.2Reglas de asociación

Otro de los modelos importantes en algoritmos no supervisados es el análisis de reglas de asociación; se hizo uso de BigML para aplicar diferentes alternativas de reglas de asociación; en todos los ensayos para el análisis se consideraron las 20 reglas con el Lift más alto.

Reglas de asociación para Edad, Sector, Examen: para esta prueba se utilizó como consecuente el examen realizado; se obtuvieron un total de 101 reglas.

Antecedent	Consequent	Leverage	Lift
SCL_DIRECCION = RUMIPAMBA	ART_DESCRIPCION_CORTA = ACETIL COLINEST	0.00031	63.3989
SCL_DIRECCION = CHILLOGALLO	ART_DESCRIPCION_CORTA = RX PELVIS	0.00036	60.66562
SCL_DIRECCION = CHILLOGALLO	ART_DESCRIPCION_CORTA = RX COL.LUMBOSACRO	0.00036	60.66562
SCL_DIRECCION = CHILLOGALLO	ART_DESCRIPCION_CORTA = CREATININA EN O	0.00036	60.66562
SCL_DIRECCION = AMAGUAÑA	ART_DESCRIPCION_CORTA = AUDIOMETRIA	0.00031	58.6311
52 < PER_EDAD <= 69 & SCL_DIRECCION = CONDADO	ART_DESCRIPCION_CORTA = PLASMA RICO EN	0.00071	55.39794
SCL_DIRECCION = BELISARIO QUEVEDO	ART_DESCRIPCION_CORTA = ANDROSTENDIONA	0.00031	48.93384
SCL_DIRECCION = AMAGUAÑA	ART_DESCRIPCION_CORTA = RAY. X CERVICAL	0.0003	41.38666
SCL_DIRECCION = RUMIPAMBA	ART_DESCRIPCION_CORTA = FOSFORO SERICO	0.0003	40.34476
SCL_DIRECCION = CHILLOGALLO	ART_DESCRIPCION_CORTA = ACIDO FOLICO	0.00035	38.60539
SCL_DIRECCION = CHILLOGALLO	ART_DESCRIPCION_CORTA = INFLUENZA A-B	0.00035	38.60539
SCL_DIRECCION = RUMIPAMBA	ART_DESCRIPCION_CORTA = AGLUTINACIONES	0.0003	34.13787
SCL_DIRECCION = CHILLOGALLO	ART_DESCRIPCION_CORTA = RAY.X COLUMNA	0.00035	32.6661
SCL_DIRECCION = CHILLOGALLO	ART_DESCRIPCION_CORTA = FIBRINOGENO	0.00035	30.33281
PER_EDAD > 69 & SCL_DIRECCION = PONCEANO	ART_DESCRIPCION_CORTA = COPROCULTIVO	0.0014	24.88529
SCL_DIRECCION = RUMIPAMBA	ART_DESCRIPCION_CORTA = MAGNESIO	0.00035	23.53444
PER_EDAD > 69 & SCL_DIRECCION = PONCEANO	ART_DESCRIPCION_CORTA = ANTIGENO CARCIN	0.00139	22.68953
PER_EDAD > 69 & SCL_DIRECCION = PONCEANO	ART_DESCRIPCION_CORTA = RETICULOCITOS	0.00139	21.429
SCL_DIRECCION = CHILLOGALLO	ART_DESCRIPCION_CORTA = VITAMINA B 12	0.00035	19.3027

Tabla 6 Reglas de asociación para edad sector y examen

Aquí se puede observar que las reglas con mayor frecuencia de ocurrencia es la prueba ACETIL COLINESTERASA (intoxicación de hígado por aspiración de componentes químicos) en el sector de RUMIPAMBA, mientras que la prueba PLASMA RICO EN PLAQUETAS (procedimiento para obtener plasma rico en plaquetas) se realiza con más frecuencia en el Condado a pacientes mayores a 52 y menores a 69 años.

Es interesante poder ver estas reglas de asociación representadas en grafos que permiten identificar gráficamente como se relacionan los datos; por tema de densidad no es posible presentar la información en el gráfico, sin embargo, los algoritmos están dispuestos en este proyecto.

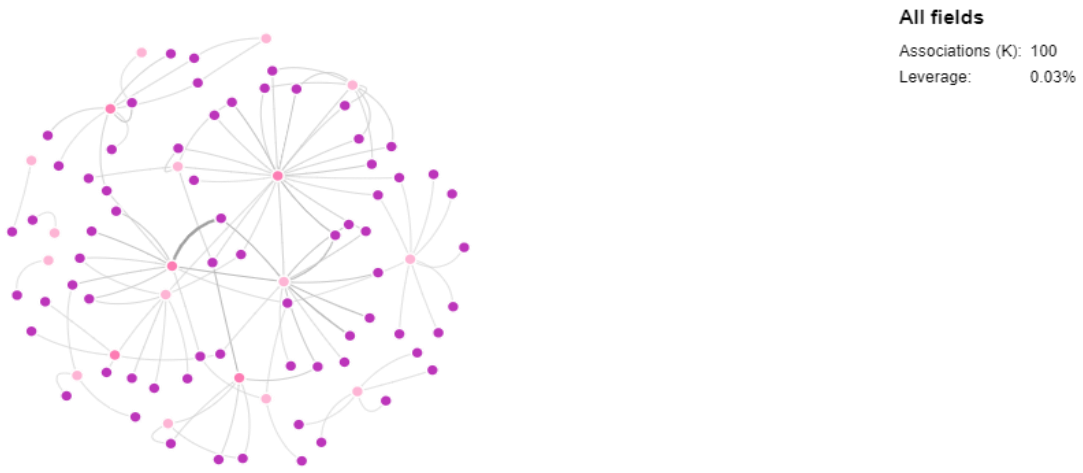


Figura 17 Grafo reglas de asociación edad, sector, examen

Reglas de asociación para Edad, Examen, Sector: para esta prueba se utilizó como consecuente el sector; se obtuvieron un total de 101 reglas.

Antecedent	Consequent	Leverage	Lift
ART_DESCRIPCION_CORTA = ACETIL COLINEST	SCL_DIRECCION = RUMIPAMBA	0.00031	63.3989
PER_EDAD > 69 & ART_DESCRIPCION_CORTA = RAY. X CERVICAL	SCL_DIRECCION = CHILLOGALLO	0.00036	60.66562
PER_EDAD > 69 & ART_DESCRIPCION_CORTA = AMILASA EN SUER	SCL_DIRECCION = CHILLOGALLO	0.00036	60.66562
ART_DESCRIPCION_CORTA = CREATININA EN O	SCL_DIRECCION = CHILLOGALLO	0.00036	60.66562
ART_DESCRIPCION_CORTA = RX PELVIS	SCL_DIRECCION = CHILLOGALLO	0.00036	60.66562
PER_EDAD > 69 & ART_DESCRIPCION_CORTA = LIPASA	SCL_DIRECCION = CHILLOGALLO	0.00036	60.66562
ART_DESCRIPCION_CORTA = RX COL.LUMBOSACRO	SCL_DIRECCION = CHILLOGALLO	0.00036	60.66562
ART_DESCRIPCION_CORTA = AUDIOMETRIA	SCL_DIRECCION = AMAGUAÑA	0.00031	58.6311
ART_DESCRIPCION_CORTA = ANDROSTENDIONA	SCL_DIRECCION = BELISARIO QUEVEDO	0.00031	48.93384
ART_DESCRIPCION_CORTA = RAY. X CERVICAL	SCL_DIRECCION = AMAGUAÑA	0.0003	41.38666
ART_DESCRIPCION_CORTA = FOSFORO SERICO	SCL_DIRECCION = RUMIPAMBA	0.0003	40.34476
ART_DESCRIPCION_CORTA = INFLUENZA A-B	SCL_DIRECCION = CHILLOGALLO	0.00035	38.60539
ART_DESCRIPCION_CORTA = ACIDO FOLICO	SCL_DIRECCION = CHILLOGALLO	0.00035	38.60539
52 < PER_EDAD <= 69 & ART_DESCRIPCION_CORTA = SET 3	SCL_DIRECCION = CONDADO	0.00071	37.12548
ART_DESCRIPCION_CORTA = AGLUTINACIONES	SCL_DIRECCION = RUMIPAMBA	0.0003	34.13787
ART_DESCRIPCION_CORTA = RAY.X COLUMNA	SCL_DIRECCION = CHILLOGALLO	0.00035	32.6661
ART_DESCRIPCION_CORTA = FIBRINOGENO	SCL_DIRECCION = CHILLOGALLO	0.00035	30.33281
ART_DESCRIPCION_CORTA = MAGNESIO	SCL_DIRECCION = RUMIPAMBA	0.00035	23.53444
ART_DESCRIPCION_CORTA = VITAMINA B 12	SCL_DIRECCION = CHILLOGALLO	0.00035	19.3027
ART_DESCRIPCION_CORTA = SET 3	SCL_DIRECCION = CONDADO	0.00069	18.56274

Tabla 7 Regla de asociación para Edad, Examen y Sector

En este análisis es importante resaltar que en el sector de CHILLOGALLO es muy frecuente exámenes de CREATININA EN ORINA (funcionamiento renal) al igual que sucedió en el análisis anterior y no hay una participación importante en algún rango de edades.

En el grafo se puede identificar lo dicho.

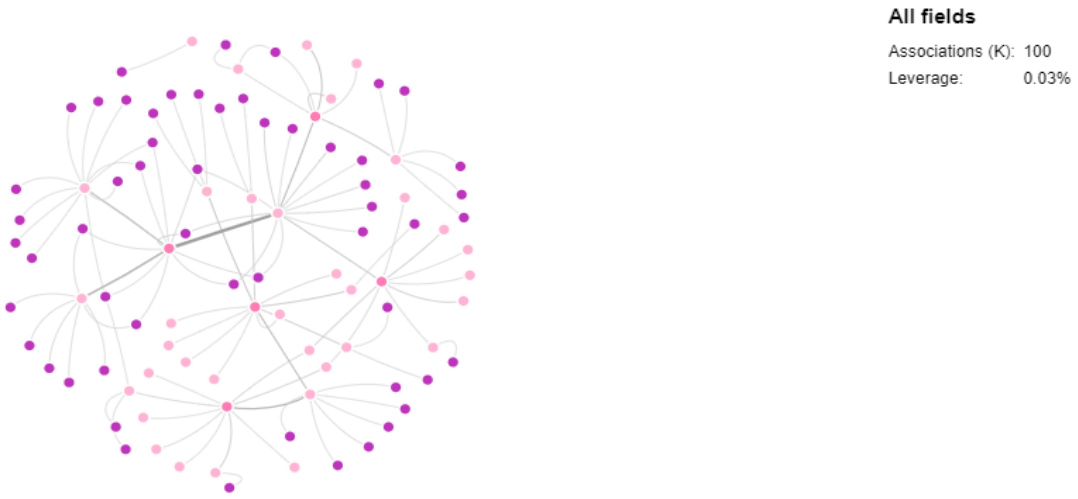


Figura 18 Grafo reglas de asociación edad, examen, sector

Reglas de asociación para Examen, Sector, Edad: para esta prueba se utilizó como consecuente la edad; se obtuvieron un total de 85 reglas.

Antecedent	Consequent	Leverage	Lift
SCL_DIRECCION = CARCELEN	42 < PER_EDAD <= 52	0.00038	5.03561
SCL_DIRECCION = PONCEANO & ART_DESCRIPCION_CORTA = ANTIGENO PROSTA	52 < PER_EDAD <= 69	0.001	5.03457
SCL_DIRECCION = LUMBISI	42 < PER_EDAD <= 52	0.00523	4.95693
ART_DESCRIPCION_CORTA = DEHIDROEPIANDRO	31 < PER_EDAD <= 42	0.0005	4.92471
ART_DESCRIPCION_CORTA = SCREENING DE DR	31 < PER_EDAD <= 42	0.00037	4.92471
SCL_DIRECCION = AMBATO	PER_EDAD <= 31	0.0005	4.86984
SCL_DIRECCION = ATUNTAQUI	PER_EDAD <= 31	0.00041	4.86984
SCL_DIRECCION = PINTAG	PER_EDAD <= 31	0.00037	4.86984
ART_DESCRIPCION_CORTA = VITAMINA B 12	PER_EDAD > 69	0.00086	4.71778
ART_DESCRIPCION_CORTA = ACIDO FOLICO	PER_EDAD > 69	0.0004	4.49312
SCL_DIRECCION = CHILLOGALLO	PER_EDAD > 69	0.01159	4.47469
ART_DESCRIPCION_CORTA = COPROCULTIVO	PER_EDAD > 69	0.00113	4.46413
SCL_DIRECCION = PONCEANO & ART_DESCRIPCION_CORTA = BILIRRUBINAS	PER_EDAD > 69	0.00113	4.46413
SCL_DIRECCION = PUEMBO	PER_EDAD <= 31	0.0004	4.42713
ART_DESCRIPCION_CORTA = RETICULOCITOS	PER_EDAD > 69	0.00129	4.39327
SCL_DIRECCION = PONCEANO & ART_DESCRIPCION_CORTA = SANGRE OCULTA E	PER_EDAD > 69	0.0012	4.36097
SCL_DIRECCION = AMAGUAÑA	PER_EDAD <= 31	0.00292	4.33535
SCL_DIRECCION = PONCEANO & ART_DESCRIPCION_CORTA = HELICOBACTER PY	PER_EDAD > 69	0.00119	4.23637
SCL_DIRECCION = GUAYAQUIL	31 < PER_EDAD <= 42	0.00063	4.14713
ART_DESCRIPCION_CORTA = ANTIGENO CARCIN	PER_EDAD > 69	0.0011	4.07024

Tabla 8 Reglas de asociación para Examen, Sector y Edad

En este análisis la regla de asociación tiene un valor de Lift bajo, por tal razón no podría considerarse como regla con un índice de ocurrencia favorable.

En Python se aplicaron dos modelos de reglas de asociación para validar sy lo obtenido en BigML es real. Se aplicaron los modelos A PRIORI y ECLAT, esto se puede evidenciar en el ANEXO 4.

4.3 Evaluación del modelo creado para la segmentación de clientes

Con este punto se busca identificar el cumplimiento de la técnica de modelado aplicados en los diferentes modelos utilizados en este análisis

4.3.1 Validación de los pasos para la ejecución de las técnicas de modelado

a. Modelo de agrupamiento.

Paso 1: Se identifica los algoritmos de agrupación a aplicar en función de los datos y sus características. Se define hacer uso de herramientas que permitan hacer una visualización rápida de las características de los datos.

Paso 2: Se aplica BIGML con los algoritmos de K-means y G-means, los dos algoritmos se aplicaron a datos numéricos etiquetados; sin embargo, estos datos no demostraron una agrupación coherente para el análisis. K-means se aplicó con valor de K=5 y K=8; G-means se aplicó con critical_values = 5 y critical_values =8

Paso 3: Se aplicó la prueba del codo para identificar los valores más óptimos de K este proceso se lo realizó con código Python que se incluye en el ANEXO 3; el K óptimo obtenido fue 3 sin embargo las agrupaciones siguen siendo incoherentes.

Paso 4: Como los dos modelos anteriormente aplicados no presentan resultados adecuados, se identifican modelos que sean más eficientes con datos categóricos y se aplica el modelo K-modes y DBSCAN; estos modelos se aplican en código Python y se incluyen en los Anexo 1 y 2 respectivamente

Paso 5: por la alta direccionalidad de los datos las técnicas de visualización no fueron muy eficientes; se opta por análisis matricial del clúster obtenidos.

Paso 6: se almacenan los cuadernos de evidencia en los anexos de este trabajo.

b. Modelo de reglas de asociación

Paso 1: Se identifica los algoritmos de asociación a aplicar en función de los datos y sus características. Se define hacer uso de herramientas que permitan hacer una visualización rápida de las características de los datos.

Paso 2: Se aplica el modelo de asociación con BIGML encontrando reglas de asociación que serán analizadas con expertos (médicos)

Paso 3: Se aplicó el modelo APRIORI y el modelo ECLAT con código Python para validad las reglas de asociación obtenidas con BIGML

Paso 4: Se generaron las matrices con las reglas de asociación y gráficas de grafos de estas.

Paso 5: Se consultó con 3 médicos especialistas, respecto a las reglas de asociación con la dirección y el examen, y no existe información que permita identificar esas reglas como ciertas

4.4 Discusión

Objetivo específico 1: Analizar las técnicas y metodologías de la minería de datos para dar solución a la segmentación de clientes

En esta etapa se realizó una investigación bibliográfica respecto a las técnicas de minería de Datos, las metodologías para la extracción, transformación y carga de los datos y una más profunda respecto a los modelos que se podrían aplicar en un análisis no supervisado.

El análisis bibliográfico permitió identificar las áreas institucionales de las que se debería obtener los datos y entender su proceso.

Adicional al análisis bibliográfico se analizó al Laboratorio Clínico Pura Vida, su razón de ser, su filosofía y las actividades que realiza, con miras a identificar y

orientar respeto al propósito y los objetivos que el laboratorio debería buscar con el análisis de sus datos.

Objetivo específico 2: Analizar la estructura de datos de los sistemas de información y la situación actual del Laboratorio Clínico Pura Vida

Se analizó la funcionalidad de los sistemas de información del laboratorio con el objetivo de entender el proceso de obtención y definición de los datos que se ingresan a sus sistemas.

Se identificó que el laboratorio dispone de dos sistemas de información, uno para la parte clínica y otro para la parte financiera, estos sistemas son de fabricantes diferentes, existió un proceso de integración que permitió que estos dos sistemas compartan información.

Conjuntamente con el personal del laboratorio se obtuvo la información haciendo uso de los reportes que el sistema entrega; se analizó la información y se definieron los datos adicionales que ninguno de los reportes entregaba y que se requeriría un nuevo desarrollo.

No se permitió el acceso directo a la base de datos por parte de los desarrolladores, se trabajó únicamente con los reportes que el sistema entrega.

Con los datos en formato Excel en algunos casos y en formato csv en otros se inició el proceso de integración, transformación y carga. Es importante mencionar que este proceso fue el más duro de este trabajo por la gran inconsistencia y poca integración de esta información que ocasionó que se pierda mucha información por mala calidad de los datos.

Todo este proceso se realizó con bases de datos y PLSQL montado en una máquina virtual con Linux y Oracle y se generó el archivo csv con el cual se trabajó este análisis.

Objetivo específico 3: Aplicar técnicas de minería de datos para la segmentación de clientes en el Laboratorio Clínico Pura Vida

Se analizaron los modelos relacionados a un análisis no supervisado, se aplicaron modelos de agrupamiento como K-means, G-means, K-modes y DBSCAN; estos modelos se aplicaron con los datos generados en el proceso anterior.

Se analizaron modelos de asociación y se implementaron modelos como A PRIORI y ECLAT.

Todos estos modelos se implementaron algunos en BIGML y otros con desarrollo en Python bajo la plataforma Jupyter Notebook.

Objetivo específico 4: Evaluar el modelo de datos obtenido para la segmentación de clientes en el Laboratorio Clínico Pura Vida

Se analizaron los pasos realizados para la implementación y validación de los modelos para luego analizar los resultados obtenidos con cada uno de estos modelos.

El análisis realizado a los resultados obtenidos con BIGML se encuentran en este documento; y el análisis de los resultados de los modelos aplicados en Python se encuentran en los respectivos Jupyter Notebooks adjuntos en los anexos de este documento.

5. Conclusiones y Recomendaciones

5.1 Conclusiones

Podemos llegar a las siguientes conclusiones

- En función al análisis bibliográfico realizado, se pudo identificar la existencia de otros modelos de los cuales no se tenía conocimiento y que al aplicar en este proyecto generó expectativas para trabajos futuros
- La data obtenida de los sistemas de información del laboratorio, tiene una alta dimensionalidad que complica el análisis de agrupamiento a realizar con estos datos.
- En el mercado de la salud, es poco probable que aplicando técnicas de agrupamiento se consiga obtener información que permita tomar decisiones para mejorar su participación en el mercado.
- El conocer y entender los modelos de agrupamiento que trabajen con datos etiquetados y categóricos es fundamental al momento de realizar un análisis de minería de datos.
- El uso de dashboard permite agilizar la comprensión de las relaciones y características de los datos, sin embargo, no se puede dejar de lado el uso de un lenguaje de programación que nos permita especificar de forma más detallada las métricas a usar en los modelos.
- La aplicación del proceso ETL, de las actividades de limpieza y de los modelos de agrupamiento, permitió identificar la importancia que tiene establecer políticas de calidad de información en las organizaciones
- La aplicación de los modelos de agrupamiento no permitió identificar grupos con característica claras, debido a la gran dimensionalidad de los datos
- La aplicación de los algoritmos de asociación permitió identificar reglas de asociación interesantes, sin embargo, en algunos de los casos los expertos no pueden dar una certeza de su ocurrencia por ejemplo la frecuencia de ciertos exámenes realizados en determinados sectores.
- Las bases de datos y el lenguaje de sentencias de búsqueda SQL son una gran alternativa para realizar el proceso ETL y la limpieza de los datos.
- La integración de los sistemas de información se debe realizar con responsabilidad, no se puede decidir gestionar una actividad en una aplicación que se generó para otro proceso. Esto disminuye la calidad de los datos y genera inconsistencia en la información.
- Si bien es cierto por los tipos de datos no se ha podido obtener resultados más relevantes, para proyectos futuros se trabajará con resultados de exámenes que permita proponer soluciones tecnológicas que apoyen a la decisión de un diagnóstico.

5.2 Recomendaciones

- Es necesario mantener claros los objetivos planteados en un proyecto de minería de datos; los volúmenes de la información, la capacidad de integración y combinación de estos, nos llevarán fácilmente a perder esta orientación.
- Se debe realizar un análisis minucioso de los datos en el que se contemple la aplicación de técnicas controladas y validadas para la extracción, la limpieza, la transformación y la carga de los datos antes de aplicar los modelos seleccionados
- Es recomendable aplicar diferentes modelos en cualquier tipo de minería que se realice, esto nos permitan analizar y validar que la información obtenida es la correcta.
- Se recomienda para proyectos futuros, trabajar con datos que hagan referencia a los resultados obtenidos en las pruebas de laboratorio; sin considerar datos personales de los pacientes se podría obtener información importante para el desarrollo de aplicaciones de apoyo a la toma de decisiones de los médicos.
- Se recomienda identificar con la data del laboratorio, propuestas de investigación que identifiquen nuevo conocimiento que se puede obtener de los resultados de una prueba de laboratorio.

6. BIBLIOGRAFÍA

Aggarwal, C. C., & Reddy, C. K. (2014). *DATA CLUSTERING Algorihms and Applications.* Boca Raton FL: CRC Press A CHAPMAN & HALL BOOK.

Alvarez-Alva, R., & Kuri-Morales, P. (2018). *Salud Pública y Medicina Preventiva.* Ciudad de México-México: El Manual Moderno S.A. de C.V.

Bao, F., Mao, L., Zhu, Y., Xiao, C., & Xu, C. (2021). An Improved Evaluation Methodology for Mining Association Rules. *axioms MDPI*, 1-16. Obtenido de https://doi.org/10.3390/axioms11010017

Bastos Boubeta, A. I. (2007). *Fidelización de clientes: Introducción a la venta personal y a la dirección de ventas.* A Coruña - españa: Ideas Propias Editorial S.L.

Cálad Noreña, F. (2015). Segmentación de clientes automatizada a partir de técnicas de minería de datos (K-means clustering). *Tesis de Ingeniería*. Antioquia: Escuela de Ingeniería de Antioquia.

Gironés-Roig, J., Casas-Roma, J., Minguillón-alfonso, J., & Caihuelas-Quiles, R. (2017). *Minería de Datos Modelos y algoritmos.* Barcelona España: Editorial UOC.

Gironés-Roig, J., Casas-Roma, J., Minguillón-Alfonso, J., & Caihuelas-Quiles, R. (2017). *Minería de datos Modelos y Algoritmos.* Barcelona: Editorial UOC.

Herrera Rangel, L. (25 de 03 de 2016). *Diseño de un programa de retención de usuarios de tarjetas de crédito*. Obtenido de Repositorio Universidad Nacional de Colombia: https://repositorio.unal.edu.co/handle/unal/59088

Jiménes Avalos, K. J. (2021). *Repositorio Universidad Nacional de Colombia.* Obtenido de Modelo para la estimación de la demanda operativa segun el perfil del cliente de una empresa de telecomunicaciones: https://repositorio.unal.edu.co/bitstream/handle/unal/81126/1128400689.2021.pdf?sequence=1&isAllowed=y

Olson, D. L. (2018). *Data Mining Models: Big Data and Business Analytics Collection.* New York: BEP Business Expert Press.

Osorio Oncoy, E. d. (02 de 04 de 2019). *Satisfación del Cliente en las Agencias de Viaje, Huaraz - 2017*. Obtenido de Repositorio Universidad de San Pedro: http://repositorio.usanpedro.edu.pe/handle/USANPEDRO/10707

Pérez, C. (2011). *Técnicas de Segmentación Conceptos, Herramientas y aplicaciones.* Esoaña: Garceta Grupo Editorial.

Reichheld, F. F., & Sasser, W. (1990). Cero deserciones: la calidad llega a los servicios. *Harvard Business Review*. Obtenido de https://hbr.org/1990/09/zero-defections-quality-comes-to-services?language=es

Rojas Huamán, A. (julio de 2020). Análisis de datos para identificar perfiles de clientes por características similares de hábitos de compras. *Tesis de Maestría*. E.T.S de Igenieros Informáticos (UPM).

Sáiz-Manzanaree, M. C., Escobar, M., & Rodríguez-Medina, L. j. (2019). *Investigación cualitativa, aplicación de métodos mixtos y de técnicas de minería de datos.* Burgos - España: Editorial Universidad de Burgos.

Tamayo W., Jovel;. (2020). Desarrollo de un modelo analítico para la segmentación de asociados en una cooperativa de ahorros y crédito. *Universidad Nacional de Colombia UN*.

Urtiz Villanueva, C. G. (27 de 10 de 2018). *Estratégia de Marketing Realacional U/R 360 Gym & Fitness*. Obtenido de https://rei.iteso.mx/bitstream/handle/11117/5756/Estrategia%20de%20marketing%20relacional%20.pdf?sequence=2

Vertice, P. (2008). *La Calidad en el Servicio al Cliente.* Malaga - España: Vértice.

Zumel, Nina; Mount, John;. (2019). *Practical Data Science with R, Second Edition (2019).* Manning.

Cluster_LAB_PV_KMODES

January 30, 2023

1 ANEXO 1

1.1 ALGORITMOS DE CLUSTERING K-modes

1.1.1 Para los datos del Laboratorio Pura Vida

Como las variables son categoricas se aplicará Kmodes Se importan las librerias que se utilizarán en la clusterización

[11]:
```
!pip install seaborn
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: seaborn in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (0.12.2)
Requirement already satisfied: numpy!=1.24.0,>=1.17 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from seaborn)
(1.24.1)
Requirement already satisfied: pandas>=0.25 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from seaborn)
(1.5.3)
Requirement already satisfied: matplotlib!=3.6.1,>=3.1 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from seaborn)
(3.6.3)
Requirement already satisfied: contourpy>=1.0.1 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from
matplotlib!=3.6.1,>=3.1->seaborn) (1.0.7)
Requirement already satisfied: cycler>=0.10 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from
matplotlib!=3.6.1,>=3.1->seaborn) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from
matplotlib!=3.6.1,>=3.1->seaborn) (4.38.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from
matplotlib!=3.6.1,>=3.1->seaborn) (1.4.4)
Requirement already satisfied: packaging>=20.0 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from
matplotlib!=3.6.1,>=3.1->seaborn) (23.0)
Requirement already satisfied: pillow>=6.2.0 in
```

```
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from
matplotlib!=3.6.1,>=3.1->seaborn) (9.4.0)
Requirement already satisfied: pyparsing>=2.2.1 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from
matplotlib!=3.6.1,>=3.1->seaborn) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from
matplotlib!=3.6.1,>=3.1->seaborn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from
pandas>=0.25->seaborn) (2022.7.1)
Requirement already satisfied: six>=1.5 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from python-
dateutil>=2.7->matplotlib!=3.6.1,>=3.1->seaborn) (1.16.0)
```

Se importan todas las librerias requeridas para el analisis

[1]:
```
import warnings
warnings.filterwarnings('ignore')

import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from matplotlib.pyplot import xticks
```

Se cargamos el archivo csv que se obtuvo despues del proceso ETL y limpieza

[2]:
```
labpv = pd.read_csv('datos_exa_f.csv', sep=",")
labpv.head()
```

[2]:

	DOC_PEDIDO	DIA_EXAMEN	MES_EXAMEN	AÑO_EXAMEN	CLI_CODIGO	PER_EDAD	SCL_DIRECCION	TAR_CODIGO	TAR_DESCRIPCION	ART_CODIGO	ART_DESCRIPCION_CORTA
0	78383	18	2	2020	9047	50	CENTRO HISTÓRICO	1	QUIMICA CLINICA	194	
1	78383	18	2	2020	9047	50	CENTRO HISTÓRICO	1	QUIMICA CLINICA	314	
2	78383	18	2	2020	9047	50	CENTRO HISTÓRICO	1	QUIMICA CLINICA	331	
3	78383	18	2	2020	9047	50	CENTRO HISTÓRICO	3	HEMATOLOGIA CLINICA	467	
4	78383	18	2	2020	9047	50	CENTRO HISTÓRICO	4	SEROLOGIA	315	

```
0       CREATININA SERI
1       UREA-BUN SUERO/
2       PERFIL LIPIDICO
3       BIOMETRIA HEMAT
4              VDRL-RPR
```

[3]:
```
labpv.columns
```

[3]:
```
Index(['DOC_PEDIDO', 'DIA_EXAMEN', 'MES_EXAMEN', 'AÑO_EXAMEN', 'CLI_CODIGO',
       'PER_EDAD', 'SCL_DIRECCION', 'TAR_CODIGO', 'TAR_DESCRIPCION',
       'ART_CODIGO', 'ART_DESCRIPCION_CORTA'],
      dtype='object')
```

Se importan las variables categoricas con las que se trabajará el modelo, en este caso se iniciará con el sector y la descripción del examen

[5]:
```
labpv_cust = labpv[['SCL_DIRECCION','TAR_DESCRIPCION','ART_DESCRIPCION_CORTA']]
labpv_cust.head(100)
```

[5]:

	SCL_DIRECCION	TAR_DESCRIPCION	ART_DESCRIPCION_CORTA
0	CENTRO HISTÓRICO	QUIMICA CLINICA	CREATININA SERI
1	CENTRO HISTÓRICO	QUIMICA CLINICA	UREA-BUN SUERO/
2	CENTRO HISTÓRICO	QUIMICA CLINICA	PERFIL LIPIDICO
3	CENTRO HISTÓRICO	HEMATOLOGIA CLINICA	BIOMETRIA HEMAT
4	CENTRO HISTÓRICO	SEROLOGIA	VDRL-RPR
..	…	…	…
95	SAN ISIDRO DEL INCA	QUIMICA CLINICA	ACIDO URICO
96	SAN ISIDRO DEL INCA	QUIMICA CLINICA	CREATININA SERI
97	SAN ISIDRO DEL INCA	QUIMICA CLINICA	TGO (AST)
98	SAN ISIDRO DEL INCA	QUIMICA CLINICA	TGP (ALT)
99	SAN ISIDRO DEL INCA	QUIMICA CLINICA	UREA-BUN SUERO/

[100 rows x 3 columns]

Se presentan 30 registros sin embargo existen 203 diferentes tipos de examenes que se realizan en el laboratorio. Cada uno de estos examenes pueden realizarce en 66 sectores diferentes identificados en la data.

Procederemos a realizar una copia de los datos antes de iniciar el proceso

Con esto se podrá transformar las variables categoricas a numéricas

[6]:
```
labpv_cust_copy = labpv_cust.copy()
```

[7]:
```
from sklearn import preprocessing
le = preprocessing.LabelEncoder()
labpv_cust = labpv_cust.apply(le.fit_transform)
labpv_cust.head(100)
```

[7]:

	SCL_DIRECCION	TAR_DESCRIPCION	ART_DESCRIPCION_CORTA
0	12	15	60
1	12	15	193
2	12	15	143
3	12	7	27
4	12	17	196
..	…	…	…
95	54	15	4
96	54	15	60
97	54	15	180
98	54	15	181
99	54	15	193

[100 rows x 3 columns]

Pasaremos estos datos a otro dataset para poder hacer utros ensayos mas adelante

[8]:
```
labpv_cust.to_csv("E:\AMORTIZ\AAIng. Sistemas\A Maes ciencia de datos\Trabajo␣
↪de disertación\DATOS TESIS MIGUEL ORTIZ\A USAR\labpv_cust.csv", sep=",",␣
↪index=False)
```

Importamos las librerías para alplicar el algoritmo Kmodes

[9]:
```
!pip install kmodes
from kmodes.kmodes import KModes
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: kmodes in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (0.12.2)
Requirement already satisfied: numpy>=1.10.4 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from kmodes)
(1.24.1)
Requirement already satisfied: scikit-learn>=0.22.0 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from kmodes)
(1.2.0)
Requirement already satisfied: scipy>=0.13.3 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from kmodes)
(1.10.0)
Requirement already satisfied: joblib>=0.11 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from kmodes)
(1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from scikit-
learn>=0.22.0->kmodes) (3.1.0)
```

[10]:
```
km_labpv = KModes(n_clusters=30, init = "Cao", n_init = 1, verbose=1)
fitClusters_labpv = km_labpv.fit_predict(labpv_cust)
```

```
Init: initializing centroids
```

```
Init: initializing clusters
Starting iterations…
Run 1, iteration: 1/100, moves: 1001, cost: 20281.0
Run 1, iteration: 2/100, moves: 201, cost: 20250.0
Run 1, iteration: 3/100, moves: 38, cost: 20250.0
```

Identificamos los lusters previstos

[11]:
```
fitClusters_labpv
```

[11]:
```
array([1, 2, 4, …, 1, 1, 1], dtype=uint16)
```

[12]:
```
clusterCentroidsDf = pd.DataFrame(km_labpv.cluster_centroids_)
clusterCentroidsDf.columns = labpv_cust.columns
```

Identificamos el modo de los clusters

[13]:
```
clusterCentroidsDf
```

[13]:

	SCL_DIRECCION	TAR_DESCRIPCION	ART_DESCRIPCION_CORTA
0	35	15	89
1	54	15	60
2	48	15	193
3	35	8	192
4	22	15	143
5	10	15	4
6	35	7	27
7	35	19	70
8	47	15	181
9	17	15	180
10	35	4	52
11	33	15	46
12	19	15	190
13	6	15	102
14	4	15	68
15	8	15	26
16	35	13	62
17	14	15	88
18	52	15	81
19	35	9	57
20	58	15	11
21	35	17	137
22	35	12	20
23	12	15	78
24	35	6	188
25	35	14	76
26	35	17	196
27	59	17	196
28	23	15	39

29 54 3 66

Clusters previstos

[14]:
```
km_huang = KModes(n_clusters=2, init = "Huang", n_init = 1, verbose=1)
fitClusters_huang = km_huang.fit_predict(labpv_cust)
fitClusters_huang
```

```
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run 1, iteration: 1/100, moves: 0, cost: 42142.0
```

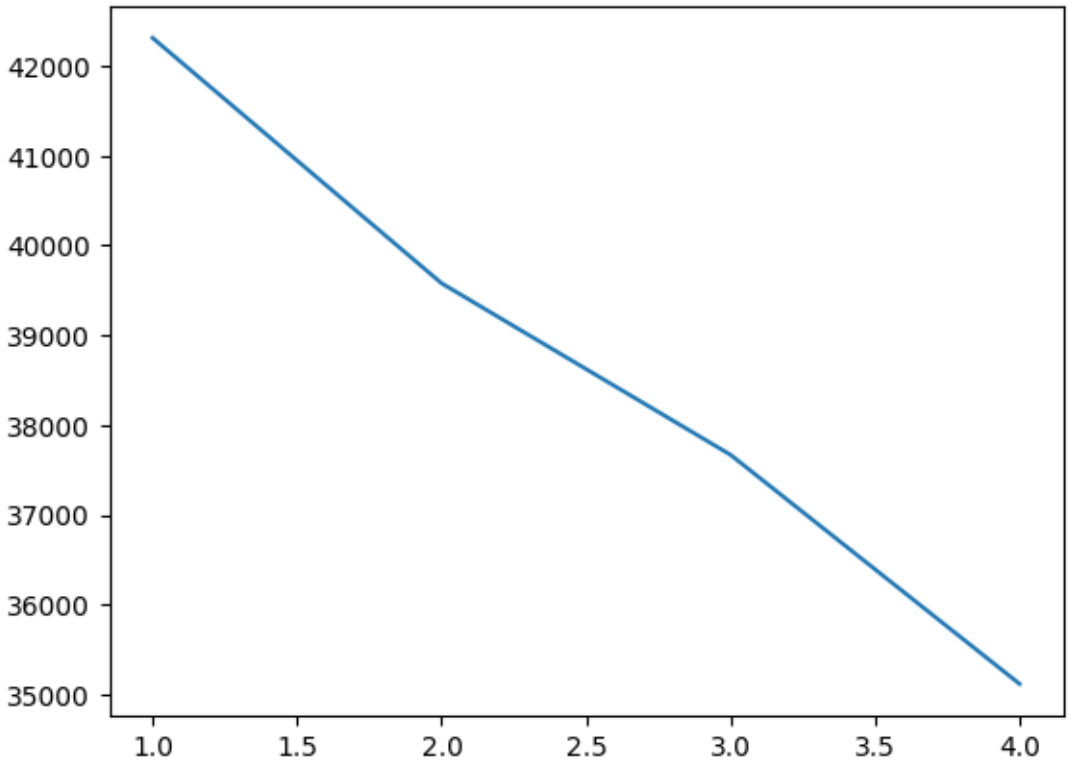
[14]:
```
array([0, 0, 0, …, 0, 0, 0], dtype=uint16)
```

[15]:
```
cost = []
for num_clusters in list(range(1,5)):
    kmode = KModes(n_clusters=num_clusters, init = "Cao", n_init = 1, verbose=1)
    kmode.fit_predict(labpv_cust)
    cost.append(kmode.cost_)

y = np.array([i for i in range(1,5,1)])
plt.plot(y,cost)
```

```
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run 1, iteration: 1/100, moves: 0, cost: 42312.0
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run 1, iteration: 1/100, moves: 243, cost: 39579.0
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run 1, iteration: 1/100, moves: 0, cost: 37668.0
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run 1, iteration: 1/100, moves: 0, cost: 35117.0
```

[15]:
```
[<matplotlib.lines.Line2D at 0x279393c4610>]
```



[16]:
```
km_cao = KModes(n_clusters=5, init = "Cao", n_init = 1, verbose=1)
fitClusters_cao = km_cao.fit_predict(labpv_cust)
```

```
Init: initializing centroids
Init: initializing clusters
Starting iterations…
Run 1, iteration: 1/100, moves: 0, cost: 33993.0
```

[17]:
```
fitClusters_cao
```

[17]:
```
array([1, 2, 4, …, 1, 1, 1], dtype=uint16)
```

[18]:
```
labpv_cust = labpv_cust_copy.reset_index()
clustersDf = pd.DataFrame(fitClusters_cao)
clustersDf.columns = ['CLUSTER_PREVISTOS']
combinedDf = pd.concat([labpv_cust, clustersDf], axis = 1).reset_index()
combinedDf = combinedDf.drop(['index', 'level_0'], axis = 1)
combinedDf.head(100)
```

[18]:

	SCL_DIRECCION	TAR_DESCRIPCION	ART_DESCRIPCION_CORTA	\
0	CENTRO HISTÓRICO	QUIMICA CLINICA	CREATININA SERI	
1	CENTRO HISTÓRICO	QUIMICA CLINICA	UREA-BUN SUERO/	
2	CENTRO HISTÓRICO	QUIMICA CLINICA	PERFIL LIPIDICO	

```
3      CENTRO HISTÓRICO  HEMATOLOGIA CLINICA       BIOMETRIA HEMAT
4      CENTRO HISTÓRICO            SEROLOGIA              VDRL-RPR
..                  …                    …                    …
95  SAN ISIDRO DEL INCA      QUIMICA CLINICA           ACIDO URICO
96  SAN ISIDRO DEL INCA      QUIMICA CLINICA       CREATININA SERI
97  SAN ISIDRO DEL INCA      QUIMICA CLINICA             TGO (AST)
98  SAN ISIDRO DEL INCA      QUIMICA CLINICA             TGP (ALT)
99  SAN ISIDRO DEL INCA      QUIMICA CLINICA       UREA-BUN SUERO/

    CLUSTER_PREVISTOS
0                   1
1                   2
2                   4
3                   0
4                   0
..                …
95                  1
96                  1
97                  1
98                  1
99                  1

[100 rows x 4 columns]
```

[19]:
```
cluster_0 = combinedDf[combinedDf['CLUSTER_PREVISTOS'] == 0]
cluster_1 = combinedDf[combinedDf['CLUSTER_PREVISTOS'] == 1]
cluster_2 = combinedDf[combinedDf['CLUSTER_PREVISTOS'] == 2]
cluster_3 = combinedDf[combinedDf['CLUSTER_PREVISTOS'] == 3]
cluster_4 = combinedDf[combinedDf['CLUSTER_PREVISTOS'] == 4]
cluster_0.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11533 entries, 3 to 19226
Data columns (total 4 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   SCL_DIRECCION         11533 non-null  object
 1   TAR_DESCRIPCION       11533 non-null  object
 2   ART_DESCRIPCION_CORTA 11533 non-null  object
 3   CLUSTER_PREVISTOS     11533 non-null  uint16
dtypes: object(3), uint16(1)
memory usage: 382.9+ KB
```

[20]:
```
cluster_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2590 entries, 0 to 19230
Data columns (total 4 columns):
```

```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   SCL_DIRECCION         2590 non-null   object
 1   TAR_DESCRIPCION       2590 non-null   object
 2   ART_DESCRIPCION_CORTA 2590 non-null   object
 3   CLUSTER_PREVISTOS     2590 non-null   uint16
dtypes: object(3), uint16(1)
memory usage: 86.0+ KB
```

[21]:
```
cluster_2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1919 entries, 1 to 19118
Data columns (total 4 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   SCL_DIRECCION         1919 non-null   object
 1   TAR_DESCRIPCION       1919 non-null   object
 2   ART_DESCRIPCION_CORTA 1919 non-null   object
 3   CLUSTER_PREVISTOS     1919 non-null   uint16
dtypes: object(3), uint16(1)
memory usage: 63.7+ KB
```

[22]:
```
cluster_3.info()
```

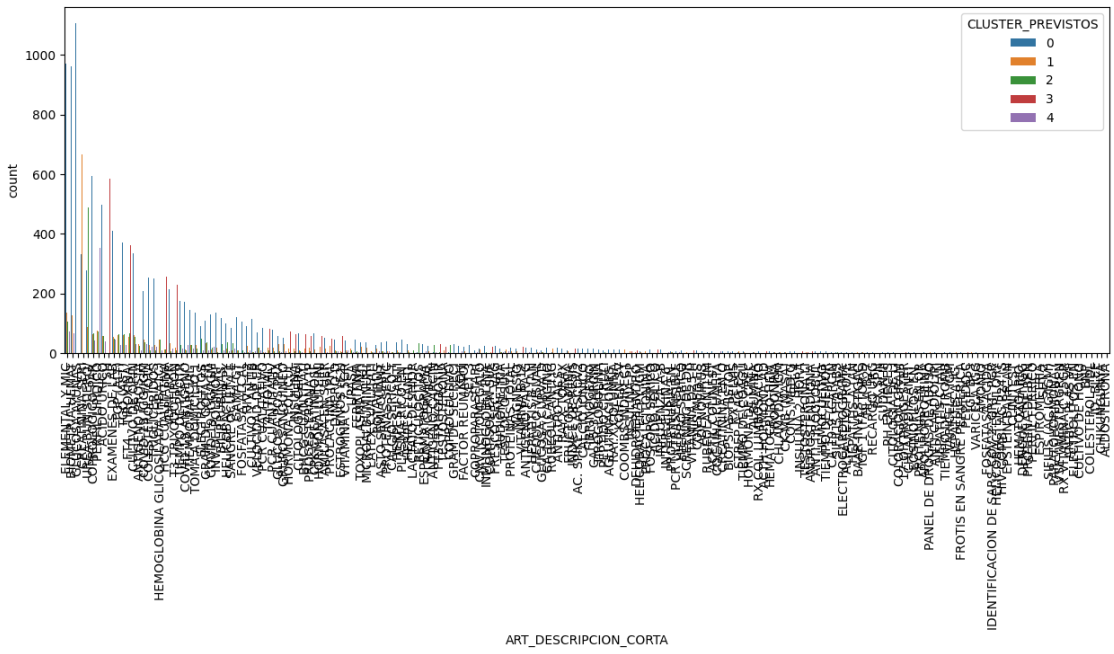
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2100 entries, 10 to 19219
Data columns (total 4 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   SCL_DIRECCION         2100 non-null   object
 1   TAR_DESCRIPCION       2100 non-null   object
 2   ART_DESCRIPCION_CORTA 2100 non-null   object
 3   CLUSTER_PREVISTOS     2100 non-null   uint16
dtypes: object(3), uint16(1)
memory usage: 69.7+ KB
```

[23]:
```
cluster_4.info()
```

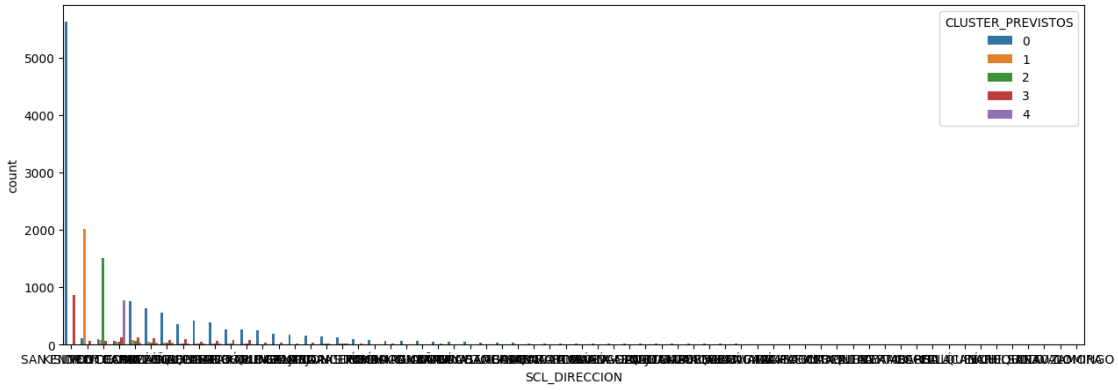
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1089 entries, 2 to 19119
Data columns (total 4 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   SCL_DIRECCION         1089 non-null   object
 1   TAR_DESCRIPCION       1089 non-null   object
 2   ART_DESCRIPCION_CORTA 1089 non-null   object
 3   CLUSTER_PREVISTOS     1089 non-null   uint16
dtypes: object(3), uint16(1)
```

```
memory usage: 36.2+ KB
```

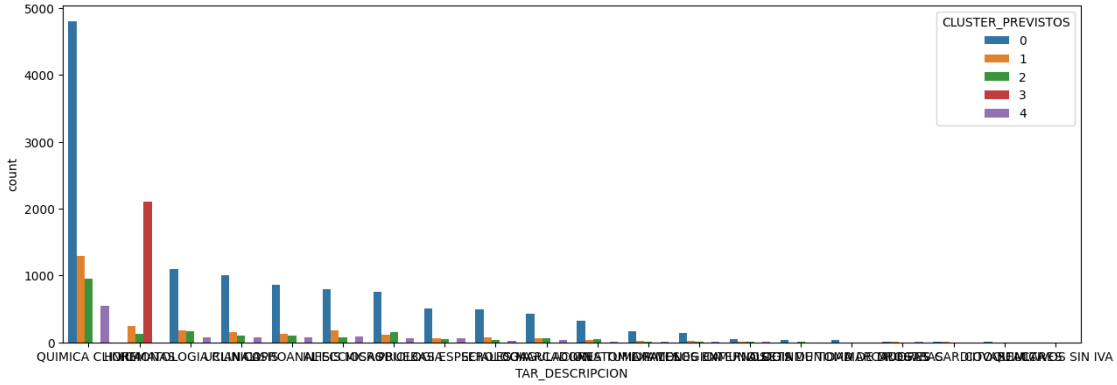
[24]:
```
plt.subplots(figsize = (15,5))
sns.
  ↪countplot(x=combinedDf['ART_DESCRIPCION_CORTA'],order=combinedDf['ART_DESCRIPCION_CORTA'].
  ↪value_counts().index,hue=combinedDf['CLUSTER_PREVISTOS'])
plt.xticks(rotation=90)
plt.show()
```



[25]:
```
plt.subplots(figsize = (15,5))
sns.countplot(x=combinedDf['SCL_DIRECCION'],order=combinedDf['SCL_DIRECCION'].
  ↪value_counts().index,hue=combinedDf['CLUSTER_PREVISTOS'])
plt.show()
```



[26]:
```
plt.subplots(figsize = (15,5))
sns.
  ↪countplot(x=combinedDf['TAR_DESCRIPCION'],order=combinedDf['TAR_DESCRIPCION'].
  ↪value_counts().index,hue=combinedDf['CLUSTER_PREVISTOS'])
plt.show()
```



[ ]:
```

```

Cluster_LAB_PV_DBSCAN

January 30, 2023

1 ANEXO 2

1.1 ALGORITMOS DE CLUSTERING K-modes

1.1.1 Para los datos del Laboratorio Pura Vida

Como las variables son categoricas se aplicará Kmodes. Se importan las librerias que se utilizarán en la clusterización

[9]:
```
import pandas as pd

dataframeLPV=pd.read_csv('datos_exa_f.csv')
dataframeLPV
```

[9]:

	DOC_PEDIDO	DIA_EXAMEN	MES_EXAMEN	AÑO_EXAMEN	CLI_CODIGO	PER_EDAD	\
0	78383	18	2	2020	9047	50	
1	78383	18	2	2020	9047	50	
2	78383	18	2	2020	9047	50	
3	78383	18	2	2020	9047	50	
4	78383	18	2	2020	9047	50	
…	…	…	…	…	…	…	
19226	66202	13	1	2018	8012	20	
19227	66202	13	1	2018	8012	20	
19228	66202	13	1	2018	8012	20	
19229	66202	13	1	2018	8012	20	
19230	66202	13	1	2018	8012	20	

	SCL_DIRECCION	TAR_CODIGO	TAR_DESCRIPCION	ART_CODIGO	\
0	CENTRO HISTÓRICO	1	QUIMICA CLINICA	194	
1	CENTRO HISTÓRICO	1	QUIMICA CLINICA	314	
2	CENTRO HISTÓRICO	1	QUIMICA CLINICA	331	
3	CENTRO HISTÓRICO	3	HEMATOLOGIA CLINICA	467	
4	CENTRO HISTÓRICO	4	SEROLOGIA	315	
…	…	…	…	…	
19226	SAN ISIDRO DEL INCA	1	QUIMICA CLINICA	103	
19227	SAN ISIDRO DEL INCA	1	QUIMICA CLINICA	105	
19228	SAN ISIDRO DEL INCA	1	QUIMICA CLINICA	181	
19229	SAN ISIDRO DEL INCA	1	QUIMICA CLINICA	194	
19230	SAN ISIDRO DEL INCA	1	QUIMICA CLINICA	299	

```
      ART_DESCRIPCION_CORTA
0            CREATININA SERI
1            UREA-BUN SUERO/
2            PERFIL LIPIDICO
3            BIOMETRIA HEMAT
4                   VDRL-RPR
…                         …
19226         GLUCOSA AYUNAS
19227            ACIDO URICO
19228        COLESTEROL TOTA
19229        CREATININA SERI
19230              TGO (AST)

[19231 rows x 11 columns]
```

[10]:
```
!pip install gower

import gower
distance_matrix = gower.gower_matrix(dataframeLPV)
distance_matrix
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: gower in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (0.1.2)
Requirement already satisfied: numpy in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from gower)
(1.24.1)
Requirement already satisfied: scipy in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (from gower)
(1.10.0)
```

[10]:
```
array([[0.        , 0.1017964 , 0.10333878, …, 0.32282555, 0.230737  ,
        0.3311725 ],
       [0.1017964 , 0.        , 0.09245146, …, 0.33371288, 0.33253342,
        0.32300702],
       [0.10333878, 0.09245146, 0.        , …, 0.33525524, 0.33407578,
        0.32454938],
       …,
       [0.32282555, 0.33371288, 0.33525524, …, 0.        , 0.09208855,
        0.10161495],
       [0.230737  , 0.33253342, 0.33407578, …, 0.09208855, 0.        ,
        0.1004355 ],
       [0.3311725 , 0.32300702, 0.32454938, …, 0.10161495, 0.1004355 ,
        0.        ]], dtype=float32)
```

[11]:
```
from sklearn.cluster import DBSCAN

dbscan_cluster = DBSCAN(eps=0.3,
                        min_samples=2,
                        metric="precomputed")

dbscan_cluster.fit(distance_matrix)
```

[11]:
```
DBSCAN(eps=0.3, metric='precomputed', min_samples=2)
```

[12]:
```
dataframe["cluster"] = dbscan_cluster.labels_
dataframe
```

[12]:

	DOC_PEDIDO	DIA_EXAMEN	MES_EXAMEN	AÑO_EXAMEN	CLI_CODIGO	PER_EDAD	\
0	78383	18	2	2020	9047	50	
1	78383	18	2	2020	9047	50	
2	78383	18	2	2020	9047	50	
3	78383	18	2	2020	9047	50	
4	78383	18	2	2020	9047	50	
…	…	…	…	…	…	…	
19226	66202	13	1	2018	8012	20	
19227	66202	13	1	2018	8012	20	
19228	66202	13	1	2018	8012	20	
19229	66202	13	1	2018	8012	20	
19230	66202	13	1	2018	8012	20	

	SCL_DIRECCION	TAR_CODIGO	TAR_DESCRIPCION	ART_CODIGO	\
0	CENTRO HISTÓRICO	1	QUIMICA CLINICA	194	
1	CENTRO HISTÓRICO	1	QUIMICA CLINICA	314	
2	CENTRO HISTÓRICO	1	QUIMICA CLINICA	331	
3	CENTRO HISTÓRICO	3	HEMATOLOGIA CLINICA	467	
4	CENTRO HISTÓRICO	4	SEROLOGIA	315	
…	…	…	…	…	
19226	SAN ISIDRO DEL INCA	1	QUIMICA CLINICA	103	
19227	SAN ISIDRO DEL INCA	1	QUIMICA CLINICA	105	
19228	SAN ISIDRO DEL INCA	1	QUIMICA CLINICA	181	
19229	SAN ISIDRO DEL INCA	1	QUIMICA CLINICA	194	
19230	SAN ISIDRO DEL INCA	1	QUIMICA CLINICA	299	

	ART_DESCRIPCION_CORTA	cluster
0	CREATININA SERI	0
1	UREA-BUN SUERO/	0
2	PERFIL LIPIDICO	0
3	BIOMETRIA HEMAT	0
4	VDRL-RPR	0
…	…	…
19226	GLUCOSA AYUNAS	0

```
19227           ACIDO URICO        0
19228       COLESTEROL TOTA        0
19229       CREATININA SERI        0
19230             TGO (AST)        0

[19231 rows x 12 columns]
```

[ ]:

Clustering_LAB_PV_K-means

January 31, 2023

1 ANEXO 3

2 ALGORITMOS DE CLUSTERING

[ ]:
```
!pip instal numpy
```

[2]:
```
# Importamos las librerias que vamos a utilizar
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import math
import seaborn as ans
```

[3]:
```
# Cargamos el archivo csv
df=pd.read_csv('datos_exa_f.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19231 entries, 0 to 19230
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   DOC_PEDIDO             19231 non-null  int64
 1   DIA_EXAMEN             19231 non-null  int64
 2   MES_EXAMEN             19231 non-null  int64
 3   AÑO_EXAMEN             19231 non-null  int64
 4   CLI_CODIGO             19231 non-null  int64
 5   PER_EDAD               19231 non-null  int64
 6   SCL_DIRECCION          19231 non-null  object
 7   TAR_CODIGO             19231 non-null  int64
 8   TAR_DESCRIPCION        19231 non-null  object
 9   ART_CODIGO             19231 non-null  int64
 10  ART_DESCRIPCION_CORTA  19231 non-null  object
dtypes: int64(8), object(3)
memory usage: 1.6+ MB
```

[4]:
```
df.head()
```

[4]:
```
   DOC_PEDIDO  DIA_EXAMEN  MES_EXAMEN  AÑO_EXAMEN  CLI_CODIGO  PER_EDAD  \
0       78383          18           2        2020        9047        50
1       78383          18           2        2020        9047        50
2       78383          18           2        2020        9047        50
3       78383          18           2        2020        9047        50
4       78383          18           2        2020        9047        50

      SCL_DIRECCION  TAR_CODIGO      TAR_DESCRIPCION  ART_CODIGO  \
0  CENTRO HISTÓRICO           1      QUIMICA CLINICA         194
1  CENTRO HISTÓRICO           1      QUIMICA CLINICA         314
2  CENTRO HISTÓRICO           1      QUIMICA CLINICA         331
3  CENTRO HISTÓRICO           3  HEMATOLOGIA CLINICA         467
4  CENTRO HISTÓRICO           4            SEROLOGIA         315

  ART_DESCRIPCION_CORTA
0       CREATININA SERI
1       UREA-BUN SUERO/
2       PERFIL LIPIDICO
3       BIOMETRIA HEMAT
4              VDRL-RPR
```

[5]:
```
# Vamos a utilizar las dimenciones Annual Income(Ingresos anuales) y Spending␣
 ↪Score(Por centaje de gasto)
X=df.iloc[:,[5,9]].values
```

3 K-means

[6]:
```
X
```

[6]:
```
array([[ 50, 194],
       [ 50, 314],
       [ 50, 331],
       …,
       [ 20, 181],
       [ 20, 194],
       [ 20, 299]], dtype=int64)
```

[7]:
```
# Algoritmo K-means
from sklearn.cluster import KMeans
# Método del Codo
# Implementación de la función objetivo que va calular la distancia entre el␣
 ↪centroide y el punto X del conjunto de datos
objective_function=[]
for i in range(1,11):
    clustering=KMeans(n_clusters=i, init='k-means++')
    clustering.fit(X)
    objective_function.append(clustering.inertia_)
```

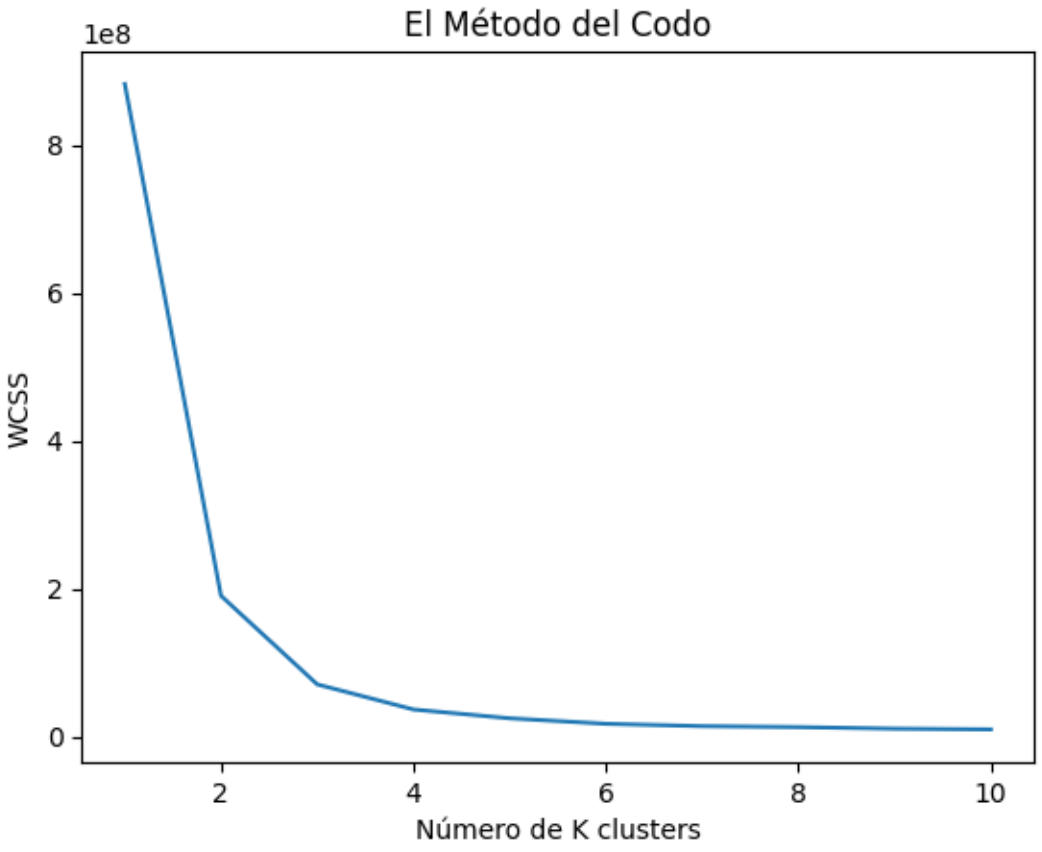
```
# se calculan las distancias y almacenamos los valores
objective_function
```

```
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
```

```
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
```

[7]:
```
[882339502.2349331,
 190073376.3258168,
 70254472.34010571,
 36281630.59484829,
 24356010.41231152,
 16924410.966697052,
 13710046.225524593,
 12430611.30866526,
 10305787.431861486,
 9360983.918626321]
```

[8]:
```
# Graficamos el número de clusters y la distancia
plt.plot(range(1,11),objective_function)
plt.title('El Método del Codo')
plt.xlabel('Número de K clusters' )
plt.ylabel('WCSS')
plt.show()
```

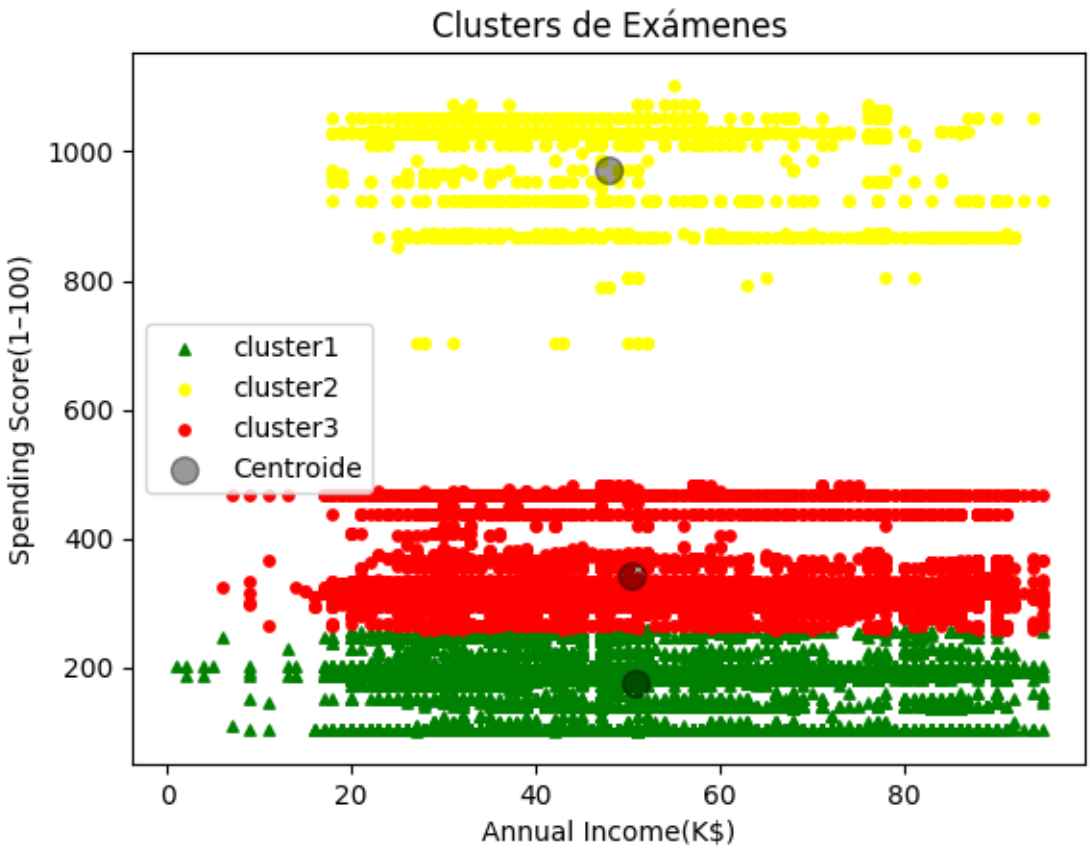


[9]:
```
# Entrenamos al modelo con el número óptimo de clusters, en este caso es 5
tuned_clustering=KMeans(n_clusters=3,init='k-means++',random_state=0)
labels=tuned_clustering.fit_predict(X)
# Los centroides calculados son
tuned_clustering.cluster_centers_[:]
```

```
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of
`n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init`
explicitly to suppress the warning
  warnings.warn(
```

[9]:
```
array([[ 50.88796366, 175.67411052],
       [ 47.85006605, 971.5673712 ],
       [ 50.46175059, 341.06659177]])
```

[18]:
```
# Graficamos los Clusters
plt.scatter(X[labels==0,0],X[labels==0,1],s=15,c='green',label='cluster1',␣
 ↪marker='^')
plt.scatter(X[labels==1,0],X[labels==1,1],s=15,c='yellow',label='cluster2')
plt.scatter(X[labels==2,0],X[labels==2,1],s=15,c='red',label='cluster3')
#plt.scatter(X[labels==3,0],X[labels==3,1],s=15,c='orange',label='cluster4')
#plt.scatter(X[labels==4,0],X[labels==4,1],s=15,c='blue',label='cluster5')
#plt.scatter(X[labels==5,0],X[labels==5,1],s=15,c='purple',label='cluster6')
plt.scatter(tuned_clustering.cluster_centers_[:,0],tuned_clustering.
 ↪cluster_centers_[:,
1],s=100,c='black',label='Centroide',alpha=0.4)
plt.title('Clusters de Exámenes')
plt.xlabel('Annual Income(K$)')
plt.ylabel('Spending Score(1-100)')
plt.legend()
plt.show()
```



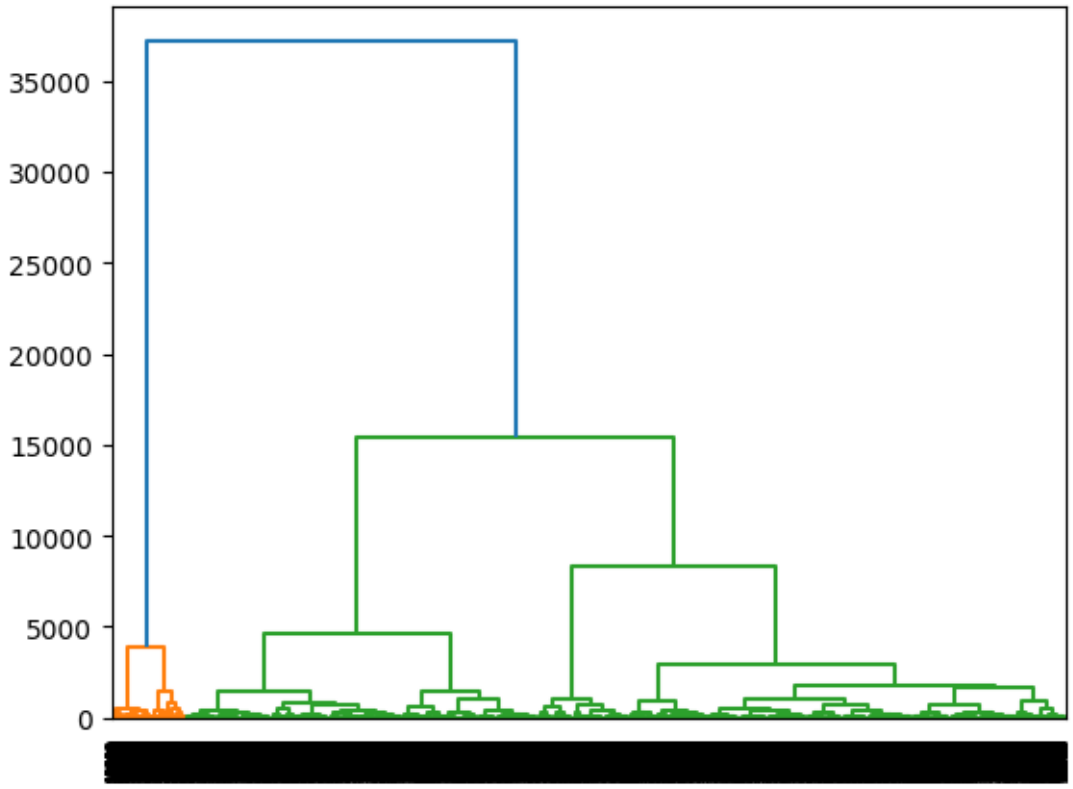
[11]:
```
# Evaluación del algoritmo
from sklearn import metrics
metrics.silhouette_score(X, tuned_clustering.labels_,metric='euclidean')
```

[11]: 0.6067419655091678

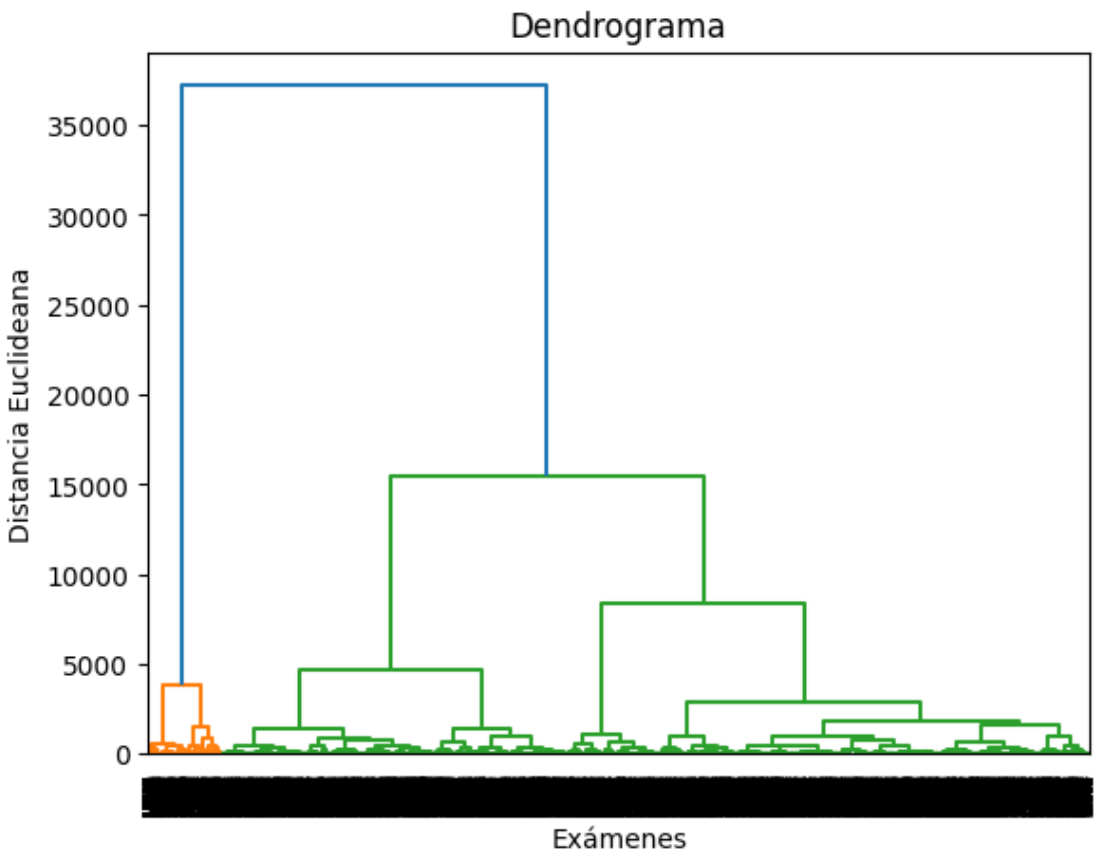
4 Agrupamiento Jerárquico HCA

[12]:
```
import scipy.cluster.hierarchy as sch
```

[13]:
```
cluster_visualising=sch.dendrogram(sch.linkage(df.iloc[:,[5,9]].
↪values,method='ward'))
```



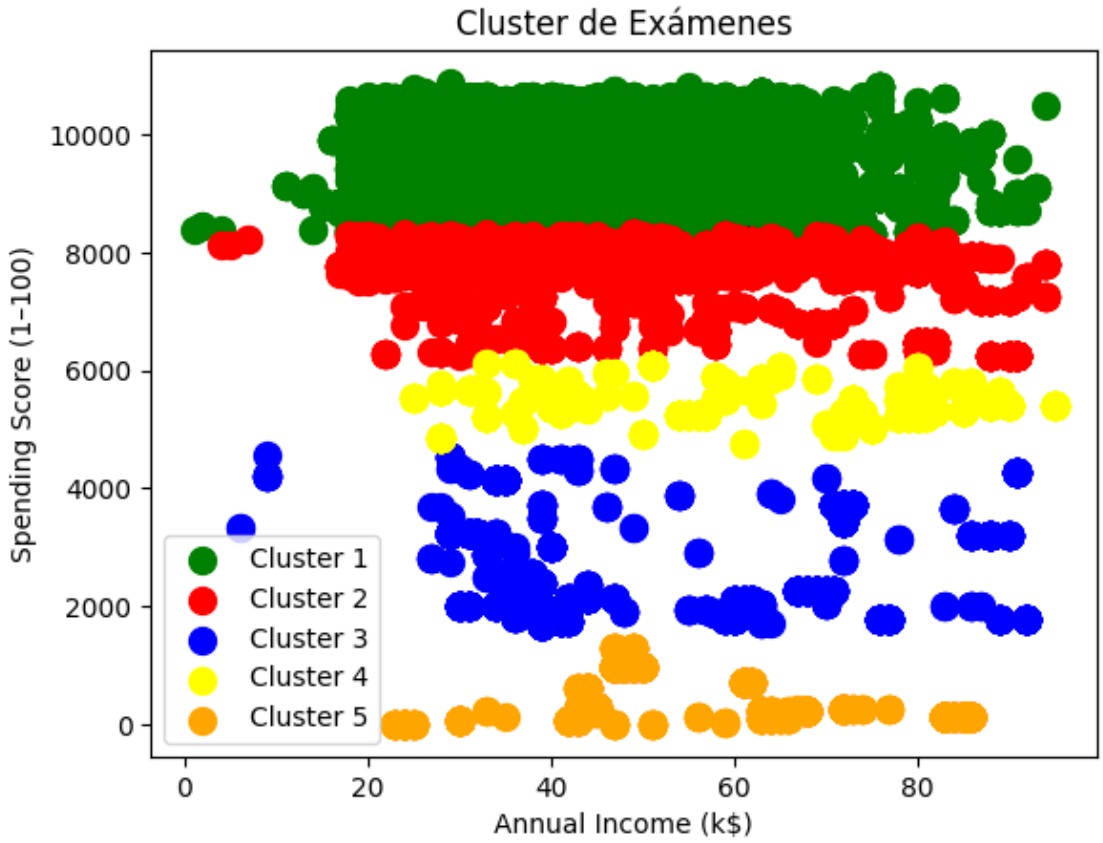
[15]:
```
# Algoritmo HCA
import scipy.cluster.hierarchy as sch
# Dendograma
cluster_visualising=sch.dendrogram(sch.linkage(df.iloc[:,[5,9]].
  ↪values,method='ward'))
plt.title('Dendrograma')
plt.xlabel('Exámenes')
plt.ylabel('Distancia Euclideana')
plt.show()
```



[16]:
```
# Inicialización del algortimo AgglomerativeClustering
# De acuerdo al dendograma el número de clusters debe ser igual a 5
from sklearn.cluster import AgglomerativeClustering
clustering_model=AgglomerativeClustering(n_clusters = 5, affinity =␣
 ↪'euclidean', linkage = 'ward')
clustering_model.fit(df.iloc[:,[5,9]].values)
#Predicción de clusters
clustering_prediction=clustering_model.fit_predict(df.iloc[:,[3,4]])
```

```
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_agglomerative.py:983: FutureWarning: Attribute
`affinity` was deprecated in version 1.2 and will be removed in 1.4. Use
`metric` instead
  warnings.warn(
C:\Users\MORTIZ\AppData\Roaming\Python\Python311\site-
packages\sklearn\cluster\_agglomerative.py:983: FutureWarning: Attribute
`affinity` was deprecated in version 1.2 and will be removed in 1.4. Use
`metric` instead
  warnings.warn(
```

[19]:
```
# Grafiación de los clusters
plt.scatter(df.iloc[:,[5,9]].values[clustering_prediction == 0, 0], df.iloc[:
 ↪,[3,4]].values[clustering_prediction == 0, 1], s = 100, c = 'green', label =␣
 ↪'Cluster 1')
plt.scatter(df.iloc[:,[5,9]].values[clustering_prediction == 1, 0], df.iloc[:
 ↪,[3,4]].values[clustering_prediction == 1, 1], s = 100, c = 'red', label =␣
 ↪'Cluster 2')
plt.scatter(df.iloc[:,[5,9]].values[clustering_prediction == 2, 0], df.iloc[:
 ↪,[3,4]].values[clustering_prediction == 2, 1], s = 100, c = 'blue', label =␣
 ↪'Cluster 3')
plt.scatter(df.iloc[:,[5,9]].values[clustering_prediction == 3, 0], df.iloc[:
 ↪,[3,4]].values[clustering_prediction == 3, 1], s = 100, c = 'yellow', label␣
 ↪= 'Cluster 4')
plt.scatter(df.iloc[:,[5,9]].values[clustering_prediction == 4, 0], df.iloc[:
 ↪,[3,4]].values[clustering_prediction == 4, 1], s = 100, c = 'orange', label␣
 ↪= 'Cluster 5')
plt.title('Cluster de Exámenes')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```

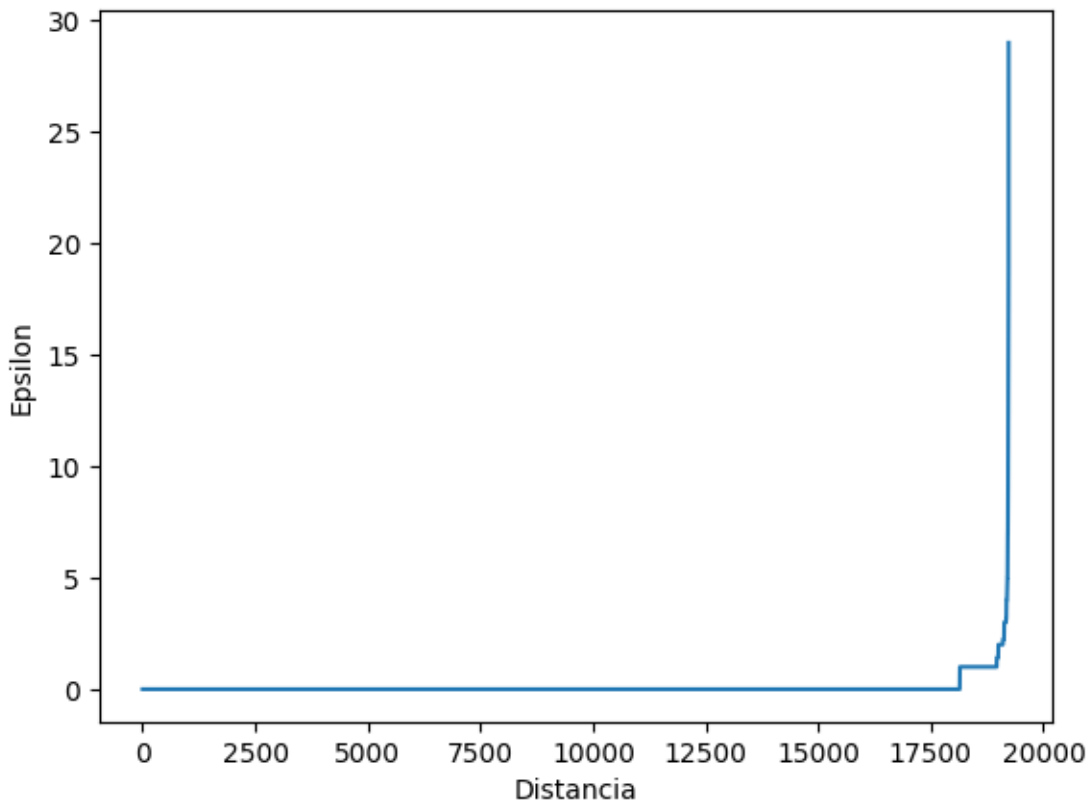


[21]:
```
#Evaluación del modelo HCA
from sklearn import metrics
metrics.silhouette_score(df.iloc[:,[5,9]].values, clustering_prediction ,␣
 ↪metric='euclidean')
```

[21]:
```
-0.12013070377666583
```

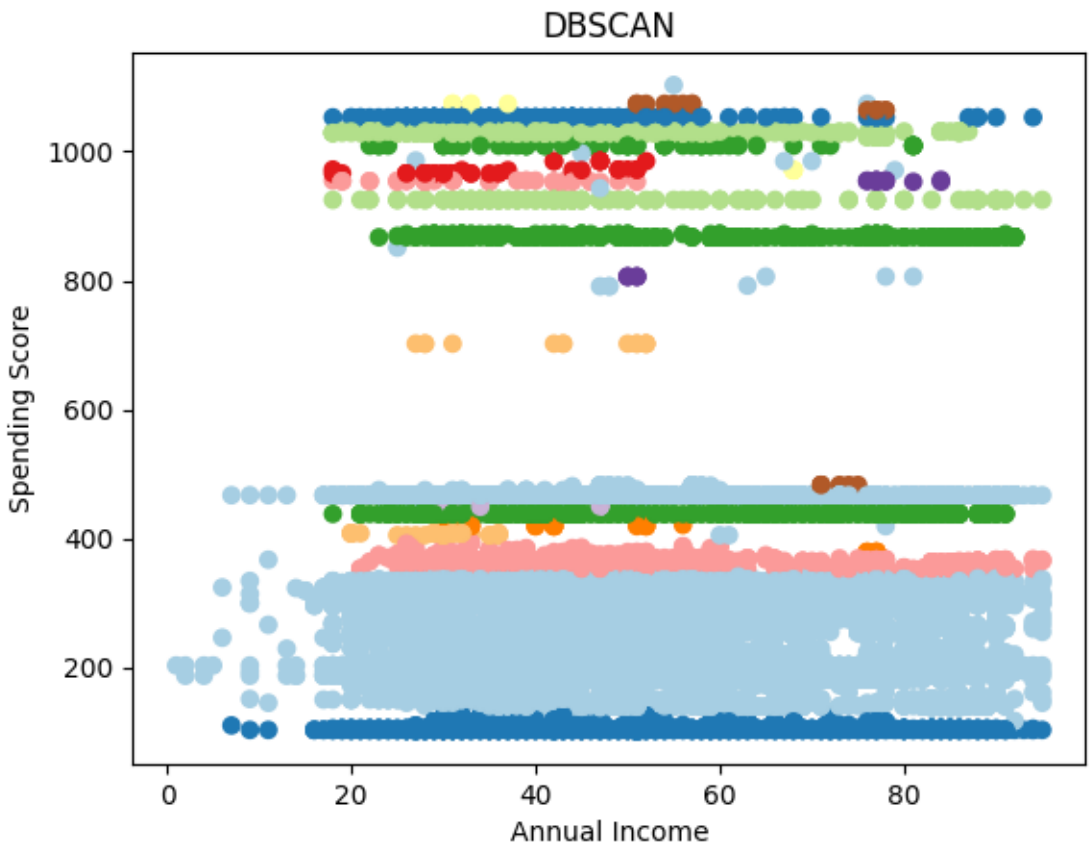
5 Agrupamiento por densidad DBSCAN

[22]:
```
#Algoritmo DBSCAN
# Calculamos la distancia entre puntos usando el algortimo NearestNeighbors
from sklearn.neighbors import NearestNeighbors
# calculando la distancia
neigh=NearestNeighbors(n_neighbors=2)
distance=neigh.fit(X)
# indices y valores de distancia
distances,indices=distance.kneighbors(X)
# Se deben ordenar las distancias en orden incremental
sorting_distances=np.sort(distances,axis=0)
# ordenando las distancias
sorted_distances=sorting_distances[:,1]
# gráfico entre la distancia vs epsilon
plt.plot(sorted_distances)
plt.xlabel('Distancia')
plt.ylabel('Epsilon')
plt.show()
```



[23]:
```
# inicializando DBSCAN
from sklearn.cluster import DBSCAN
clustering_model=DBSCAN(eps=9,min_samples=4)
# fit the model to X
clustering_model.fit(X)
predicted_labels=clustering_model.labels_
# visualizando los clusters
plt.scatter(X[:,0], X[:,1],c=predicted_labels, cmap='Paired')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.title('DBSCAN')
```

[23]:
```
Text(0.5, 1.0, 'DBSCAN')
```



[24]:
```
#Evaluando el modelo
from sklearn import metrics
metrics.silhouette_score(X, predicted_labels)
```

[24]: 0.09421249854353589

[ ]:

Asociacion_Lab_PV

January 31, 2023

1 ANEXO 4

1.1 ALGORITMOS DE ASOCIACIÓN APRIORI ECLAT

1.1.1 Para los datos del Laboratorio Pura Vida

[1]:
```
##Algoritmo Apriori
#Previamente debe instalar desde el algoritmo Apriori con el siguiente comando:
!pip install apyori
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: apyori in
c:\users\mortiz\appdata\roaming\python\python311\site-packages (1.1.2)


[notice] A new release of pip available: 22.3.1 -> 23.0
[notice] To update, run: python.exe -m pip install --upgrade pip
```

[37]:
```
#Importando las Bibliotecas a utilizar
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

[38]:
```
#Prepocesamiento de datos
dataset=pd.read_csv('datos_exa_f.csv')
dataset.head()
dataset.shape
```

[38]:
```
(19231, 11)
```

[39]:
```
transactions=[]
for i in range (0,19231):
  transactions.append([str(dataset.values[i,j]) for j in range(5,10)])
```

[40]:
```
dataset.head(5)
```

[40]:

	DOC_PEDIDO	DIA_EXAMEN	MES_EXAMEN	AÑO_EXAMEN	CLI_CODIGO	PER_EDAD	\
0	78383	18	2	2020	9047	50	
1	78383	18	2	2020	9047	50	
2	78383	18	2	2020	9047	50	

```
3       78383          18           2        2020        9047        50
4       78383          18           2        2020        9047        50

      SCL_DIRECCION  TAR_CODIGO      TAR_DESCRIPCION  ART_CODIGO  \
0  CENTRO HISTÓRICO           1      QUIMICA CLINICA         194
1  CENTRO HISTÓRICO           1      QUIMICA CLINICA         314
2  CENTRO HISTÓRICO           1      QUIMICA CLINICA         331
3  CENTRO HISTÓRICO           3  HEMATOLOGIA CLINICA         467
4  CENTRO HISTÓRICO           4             SEROLOGIA         315

  ART_DESCRIPCION_CORTA
0       CREATININA SERI
1       UREA-BUN SUERO/
2       PERFIL LIPIDICO
3       BIOMETRIA HEMAT
4              VDRL-RPR
```

[41]:
```
#Entrenando el algoritmo Apriori
from apyori import apriori

#los parámetros son:
#records: lista de listas
#min_support: valor de probabilidad para seleccionar los elementos con valores␣
 ↪de soporte superiores al valor especificado por el parámetro.
#min_confidence: valor de probabilidad para filtrar reglas con mayor confianza␣
 ↪que el umbral especificado
#min_lift: valor mínimo de lift para preseleccionar la lista de reglas
#min_length: número mínimo de elementos que desea en sus reglas

rules = apriori(transactions,min_support=0.0056,min_confidence=0.
 ↪2,min_lift=3,min_length=3)
```

[42]:
```
#Visualizando los resultados
rslt =list(rules)
print(len(rslt))
print(rslt)
```

```
128
[RelationRecord(items=frozenset({'MARCADORES TUMORALES', '10'}),
support=0.021735739171129947,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'10'}),
items_add=frozenset({'MARCADORES TUMORALES'}), confidence=1.0,
lift=46.00717703349282), OrderedStatistic(items_base=frozenset({'MARCADORES
TUMORALES'}), items_add=frozenset({'10'}), confidence=1.0,
lift=46.00717703349282)]), RelationRecord(items=frozenset({'INFECCIOSAS',
'11'}), support=0.059071291144506266,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'11'}),
items_add=frozenset({'INFECCIOSAS'}), confidence=0.9956178790534619,
```

lift=16.854513584574935),
OrderedStatistic(items_base=frozenset({'INFECCIOSAS'}),
items_add=frozenset({'11'}), confidence=1.0, lift=16.854513584574935)]),
RelationRecord(items=frozenset({'83', '128'}), support=0.012583848993812074,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'128'}),
items_add=frozenset({'83'}), confidence=0.39285714285714285,
lift=26.232762896825395), OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'128'}), confidence=0.8402777777777778,
lift=26.232762896825395)]), RelationRecord(items=frozenset({'85', '128'}),
support=0.009151890177317872,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'128'}),
items_add=frozenset({'85'}), confidence=0.2857142857142857,
lift=17.170535714285716), OrderedStatistic(items_base=frozenset({'85'}),
items_add=frozenset({'128'}), confidence=0.55, lift=17.170535714285716)]),
RelationRecord(items=frozenset({'128', '86'}), support=0.0068639176329884045,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'128'}),
items_add=frozenset({'86'}), confidence=0.2142857142857143,
lift=19.530467163168584), OrderedStatistic(items_base=frozenset({'86'}),
items_add=frozenset({'128'}), confidence=0.6255924170616114,
lift=19.530467163168584)]), RelationRecord(items=frozenset({'PONCEANO', '128'}),
support=0.032031615620612554,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'128'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629),
OrderedStatistic(items_base=frozenset({'PONCEANO'}),
items_add=frozenset({'128'}), confidence=0.3526044647967945,
lift=11.008013737836292)]), RelationRecord(items=frozenset({'13',
'COAGULACION'}), support=0.03109562685247777,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'13'}),
items_add=frozenset({'COAGULACION'}), confidence=0.9933554817275748,
lift=31.945182724252494),
OrderedStatistic(items_base=frozenset({'COAGULACION'}),
items_add=frozenset({'13'}), confidence=1.0, lift=31.945182724252494)]),
RelationRecord(items=frozenset({'PRUEBAS ESPECIALES', '14'}),
support=0.0361395663252041,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'14'}),
items_add=frozenset({'PRUEBAS ESPECIALES'}), confidence=0.9156785243741765,
lift=25.337285902503293), OrderedStatistic(items_base=frozenset({'PRUEBAS
ESPECIALES'}), items_add=frozenset({'14'}), confidence=1.0,
lift=25.337285902503293)]), RelationRecord(items=frozenset({'16', 'ANATOMIA
PATOLOGICA'}), support=0.011231865217617388,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'16'}),
items_add=frozenset({'ANATOMIA PATOLOGICA'}), confidence=0.96,
lift=85.47111111111111), OrderedStatistic(items_base=frozenset({'ANATOMIA
PATOLOGICA'}), items_add=frozenset({'16'}), confidence=1.0,
lift=85.47111111111111)]), RelationRecord(items=frozenset({'1687', '39'}),
support=0.008891893297280433,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'1687'}),
items_add=frozenset({'39'}), confidence=1.0, lift=41.35698924731183),

OrderedStatistic(items_base=frozenset({'39'}), items_add=frozenset({'1687'}),
confidence=0.36774193548387096, lift=41.35698924731183)]),
RelationRecord(items=frozenset({'SAN ISIDRO DEL INCA', '1687'}),
support=0.008891893297280433,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'1687'}),
items_add=frozenset({'SAN ISIDRO DEL INCA'}), confidence=1.0,
lift=8.741363636363637)]), RelationRecord(items=frozenset({'HORMONAS', '2'}),
support=0.12843845873849513,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2'}),
items_add=frozenset({'HORMONAS'}), confidence=0.9991909385113269,
lift=7.77953074433657), OrderedStatistic(items_base=frozenset({'HORMONAS'}),
items_add=frozenset({'2'}), confidence=1.0, lift=7.77953074433657)]),
RelationRecord(items=frozenset({'2012', 'PONCEANO'}),
support=0.006239925120898549,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2012'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629)]),
RelationRecord(items=frozenset({'2013', 'PONCEANO'}),
support=0.007903905153138163,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2013'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629)]),
RelationRecord(items=frozenset({'68', '218'}), support=0.007279912641048308,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'218'}),
items_add=frozenset({'68'}), confidence=1.0, lift=65.41156462585033),
OrderedStatistic(items_base=frozenset({'68'}), items_add=frozenset({'218'}),
confidence=0.47619047619047616, lift=65.41156462585033)]),
RelationRecord(items=frozenset({'COMITÉ DEL PUEBLO', '218'}),
support=0.007279912641048308,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'218'}),
items_add=frozenset({'COMITÉ DEL PUEBLO'}), confidence=1.0,
lift=25.641333333333332)]), RelationRecord(items=frozenset({'CALDERÓN',
'2405'}), support=0.0066559201289584525,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2405'}),
items_add=frozenset({'CALDERÓN'}), confidence=1.0, lift=38.85050505050505),
OrderedStatistic(items_base=frozenset({'CALDERÓN'}),
items_add=frozenset({'2405'}), confidence=0.2585858585858586,
lift=38.85050505050505)]), RelationRecord(items=frozenset({'36', '2563'}),
support=0.005927928864853622,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2563'}),
items_add=frozenset({'36'}), confidence=0.6000000000000001,
lift=27.342654028436023), OrderedStatistic(items_base=frozenset({'36'}),
items_add=frozenset({'2563'}), confidence=0.27014218009478674,
lift=27.342654028436023)]), RelationRecord(items=frozenset({'IÑAQUITO',
'2563'}), support=0.009879881441422702,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2563'}),
items_add=frozenset({'IÑAQUITO'}), confidence=1.0, lift=36.63047619047619),
OrderedStatistic(items_base=frozenset({'IÑAQUITO'}),
items_add=frozenset({'2563'}), confidence=0.3619047619047619,
lift=36.630476190476195)]), RelationRecord(items=frozenset({'3226', '29'}),

support=0.011231865217617388,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'29'}),
items_add=frozenset({'3226'}), confidence=0.4114285714285714,
lift=19.98025974025974), OrderedStatistic(items_base=frozenset({'3226'}),
items_add=frozenset({'29'}), confidence=0.5454545454545454,
lift=19.98025974025974)]), RelationRecord(items=frozenset({'SAN ISIDRO DEL
INCA', '29'}), support=0.013987832146014248,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'29'}),
items_add=frozenset({'SAN ISIDRO DEL INCA'}), confidence=0.5123809523809524,
lift=4.478908225108225)]), RelationRecord(items=frozenset({'3', 'HEMATOLOGIA
CLINICA'}), support=0.07940304716343403,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'3'}),
items_add=frozenset({'HEMATOLOGIA CLINICA'}), confidence=1.0,
lift=12.593975114603799), OrderedStatistic(items_base=frozenset({'HEMATOLOGIA
CLINICA'}), items_add=frozenset({'3'}), confidence=1.0,
lift=12.593975114603799)]), RelationRecord(items=frozenset({'3226', '31'}),
support=0.005615932608808694,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'3226'}),
items_add=frozenset({'31'}), confidence=0.2727272727272727,
lift=7.1358070500927635)]), RelationRecord(items=frozenset({'EXAMENES EXTERNOS',
'31'}), support=0.010035879569445167,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'31'}),
items_add=frozenset({'EXAMENES EXTERNOS'}), confidence=0.26258503401360545,
lift=26.164625850340133), OrderedStatistic(items_base=frozenset({'EXAMENES
EXTERNOS'}), items_add=frozenset({'31'}), confidence=1.0,
lift=26.164625850340133)]), RelationRecord(items=frozenset({'CARCELÉN', '32'}),
support=0.006551921376943477,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'32'}),
items_add=frozenset({'CARCELÉN'}), confidence=0.302158273381295,
lift=5.435739715056767)]), RelationRecord(items=frozenset({'3226', 'SAN ISIDRO
DEL INCA'}), support=0.020591752898965213,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'3226'}),
items_add=frozenset({'SAN ISIDRO DEL INCA'}), confidence=1.0,
lift=8.741363636363637)]), RelationRecord(items=frozenset({'36', 'IÑAQUITO'}),
support=0.005927928864853622,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'36'}),
items_add=frozenset({'IÑAQUITO'}), confidence=0.27014218009478674,
lift=9.895436696005417), OrderedStatistic(items_base=frozenset({'IÑAQUITO'}),
items_add=frozenset({'36'}), confidence=0.21714285714285714,
lift=9.895436696005415)]), RelationRecord(items=frozenset({'SAN ISIDRO DEL
INCA', '39'}), support=0.013467838385939368,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'39'}),
items_add=frozenset({'SAN ISIDRO DEL INCA'}), confidence=0.556989247311828,
lift=4.868845552297166)]), RelationRecord(items=frozenset({'SEROLOGIA', '4'}),
support=0.03260360875669492,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'4'}),
items_add=frozenset({'SEROLOGIA'}), confidence=0.9968203497615263,
lift=30.573926868044516), OrderedStatistic(items_base=frozenset({'SEROLOGIA'}),

items_add=frozenset({'4'}), confidence=1.0, lift=30.573926868044516)]),
RelationRecord(items=frozenset({'42', '5372'}), support=0.007279912641048308,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'42'}),
items_add=frozenset({'5372'}), confidence=0.3218390804597701,
lift=29.472796934865897), OrderedStatistic(items_base=frozenset({'5372'}),
items_add=frozenset({'42'}), confidence=0.6666666666666666,
lift=29.472796934865897)]), RelationRecord(items=frozenset({'BELISARIO QUEVEDO',
'42'}), support=0.007279912641048308,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'42'}),
items_add=frozenset({'BELISARIO QUEVEDO'}), confidence=0.3218390804597701,
lift=15.748822789623), OrderedStatistic(items_base=frozenset({'BELISARIO
QUEVEDO'}), items_add=frozenset({'42'}), confidence=0.356234096692112,
lift=15.748822789623)]), RelationRecord(items=frozenset({'45', 'COMITÉ DEL
PUEBLO'}), support=0.007383911393063283,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'45'}),
items_add=frozenset({'COMITÉ DEL PUEBLO'}), confidence=0.3183856502242152,
lift=8.163832585949176)]), RelationRecord(items=frozenset({'47', '9713'}),
support=0.0066559201289584525,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'47'}),
items_add=frozenset({'9713'}), confidence=0.2277580071174377,
lift=34.21886120996441), OrderedStatistic(items_base=frozenset({'9713'}),
items_add=frozenset({'47'}), confidence=1.0, lift=34.21886120996441)]),
RelationRecord(items=frozenset({'981', '47'}), support=0.009359887681347824,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'47'}),
items_add=frozenset({'981'}), confidence=0.3202846975088968,
lift=14.665226232841892), OrderedStatistic(items_base=frozenset({'981'}),
items_add=frozenset({'47'}), confidence=0.4285714285714286,
lift=14.665226232841892)]), RelationRecord(items=frozenset({'47', 'CARAPUNGO'}),
support=0.0066559201289584525,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'47'}),
items_add=frozenset({'CARAPUNGO'}), confidence=0.2277580071174377,
lift=11.202082442136687), OrderedStatistic(items_base=frozenset({'CARAPUNGO'}),
items_add=frozenset({'47'}), confidence=0.3273657289002558,
lift=11.202082442136689)]), RelationRecord(items=frozenset({'981', '48'}),
support=0.006239925120898549,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'48'}),
items_add=frozenset({'981'}), confidence=0.38961038961038963,
lift=17.83951762523191), OrderedStatistic(items_base=frozenset({'981'}),
items_add=frozenset({'48'}), confidence=0.2857142857142857,
lift=17.83951762523191)]), RelationRecord(items=frozenset({'URIANALISIS', '5'}),
support=0.06936716759398887,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'5'}),
items_add=frozenset({'URIANALISIS'}), confidence=1.0, lift=14.416041979010496),
OrderedStatistic(items_base=frozenset({'URIANALISIS'}),
items_add=frozenset({'5'}), confidence=1.0, lift=14.416041979010496)]),
RelationRecord(items=frozenset({'51', 'CONDADO'}), support=0.006031927616868597,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'51'}),
items_add=frozenset({'CONDADO'}), confidence=0.21928166351606804,

lift=8.140937588952712), OrderedStatistic(items_base=frozenset({'CONDADO'}),
items_add=frozenset({'51'}), confidence=0.22393822393822393,
lift=8.140937588952712)]), RelationRecord(items=frozenset({'5251', 'JIPIJAPA'}),
support=0.006759918880973428,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'5251'}),
items_add=frozenset({'JIPIJAPA'}), confidence=1.0, lift=90.28638497652582),
OrderedStatistic(items_base=frozenset({'JIPIJAPA'}),
items_add=frozenset({'5251'}), confidence=0.6103286384976525,
lift=90.28638497652581)]), RelationRecord(items=frozenset({'BELISARIO QUEVEDO',
'5372'}), support=0.010919868961572462,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'5372'}),
items_add=frozenset({'BELISARIO QUEVEDO'}), confidence=1.0,
lift=48.93384223918575), OrderedStatistic(items_base=frozenset({'BELISARIO
QUEVEDO'}), items_add=frozenset({'5372'}), confidence=0.5343511450381679,
lift=48.93384223918575)]), RelationRecord(items=frozenset({'CARCELÉN', '5658'}),
support=0.008111902657168114,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'5658'}),
items_add=frozenset({'CARCELÉN'}), confidence=1.0, lift=17.989710009354535)]),
RelationRecord(items=frozenset({'6', 'COPROANALISIS'}),
support=0.06094326868077583,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'6'}),
items_add=frozenset({'COPROANALISIS'}), confidence=0.979933110367893,
lift=16.07943143812709),
OrderedStatistic(items_base=frozenset({'COPROANALISIS'}),
items_add=frozenset({'6'}), confidence=1.0, lift=16.07943143812709)]),
RelationRecord(items=frozenset({'7684', '60'}), support=0.006551921376943477,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'60'}),
items_add=frozenset({'7684'}), confidence=0.4131147540983607,
lift=36.78060109289618), OrderedStatistic(items_base=frozenset({'7684'}),
items_add=frozenset({'60'}), confidence=0.5833333333333334,
lift=36.78060109289618)]), RelationRecord(items=frozenset({'6250', 'SAN ISIDRO
DEL INCA'}), support=0.007799906401123186,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'6250'}),
items_add=frozenset({'SAN ISIDRO DEL INCA'}), confidence=1.0,
lift=8.741363636363637)]), RelationRecord(items=frozenset({'66', '7629'}),
support=0.009359887681347824,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'66'}),
items_add=frozenset({'7629'}), confidence=0.5901639344262296,
lift=63.05245901639345), OrderedStatistic(items_base=frozenset({'7629'}),
items_add=frozenset({'66'}), confidence=1.0, lift=63.05245901639345)]),
RelationRecord(items=frozenset({'66', 'RUMIPAMBA'}),
support=0.009359887681347824,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'66'}),
items_add=frozenset({'RUMIPAMBA'}), confidence=0.5901639344262296,
lift=43.6517023959647), OrderedStatistic(items_base=frozenset({'RUMIPAMBA'}),
items_add=frozenset({'66'}), confidence=0.6923076923076924,
lift=43.6517023959647)]), RelationRecord(items=frozenset({'COMITÉ DEL PUEBLO',
'68'}), support=0.0073319120170557955,

ordered_statistics=[OrderedStatistic(items_base=frozenset({'68'}),
items_add=frozenset({'COMITÉ DEL PUEBLO'}), confidence=0.47959183673469385,
lift=12.297374149659863)]), RelationRecord(items=frozenset({'RUMIPAMBA',
'7629'}), support=0.009359887681347824,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'7629'}),
items_add=frozenset({'RUMIPAMBA'}), confidence=1.0, lift=73.96538461538462),
OrderedStatistic(items_base=frozenset({'RUMIPAMBA'}),
items_add=frozenset({'7629'}), confidence=0.6923076923076924,
lift=73.96538461538462)]), RelationRecord(items=frozenset({'77', '8919'}),
support=0.008527897665228018,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'77'}),
items_add=frozenset({'8919'}), confidence=0.5173501577287066,
lift=34.666065795403334), OrderedStatistic(items_base=frozenset({'8919'}),
items_add=frozenset({'77'}), confidence=0.5714285714285715,
lift=34.666065795403334)]), RelationRecord(items=frozenset({'77',
'CHILLOGALLO'}), support=0.008527897665228018,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'77'}),
items_add=frozenset({'CHILLOGALLO'}), confidence=0.5173501577287066,
lift=31.385365562399862),
OrderedStatistic(items_base=frozenset({'CHILLOGALLO'}),
items_add=frozenset({'77'}), confidence=0.5173501577287066,
lift=31.385365562399862)]), RelationRecord(items=frozenset({'8',
'MICROBIOLOGIA'}), support=0.05647132234413187,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8'}),
items_add=frozenset({'MICROBIOLOGIA'}), confidence=1.0,
lift=17.708103130755063),
OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA'}),
items_add=frozenset({'8'}), confidence=1.0, lift=17.708103130755063)]),
RelationRecord(items=frozenset({'POMASQUI', '8200'}),
support=0.007487910145078259,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8200'}),
items_add=frozenset({'POMASQUI'}), confidence=1.0, lift=22.053899082568808)]),
RelationRecord(items=frozenset({'83', 'PONCEANO'}),
support=0.012583848993812074,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'PONCEANO'}), confidence=0.8402777777777778,
lift=9.249789321376328)]), RelationRecord(items=frozenset({'PONCEANO', '85'}),
support=0.009151890177317872,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'85'}),
items_add=frozenset({'PONCEANO'}), confidence=0.55, lift=6.05440755580996)]),
RelationRecord(items=frozenset({'PONCEANO', '86'}),
support=0.007071915137018356,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'86'}),
items_add=frozenset({'PONCEANO'}), confidence=0.6445497630331753,
lift=7.09521264618832)]), RelationRecord(items=frozenset({'CHILLOGALLO',
'8919'}), support=0.01492382091414903,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8919'}),
items_add=frozenset({'CHILLOGALLO'}), confidence=1.0, lift=60.66561514195583),

OrderedStatistic(items_base=frozenset({'CHILLOGALLO'}),
items_add=frozenset({'8919'}), confidence=0.9053627760252365,
lift=60.66561514195583)]), RelationRecord(items=frozenset({'9426', 'PONCEANO'}),
support=0.005615932608808694,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'9426'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629)]),
RelationRecord(items=frozenset({'CARAPUNGO', '9713'}),
support=0.0066559201289584525,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'9713'}),
items_add=frozenset({'CARAPUNGO'}), confidence=1.0, lift=49.18414322250639),
OrderedStatistic(items_base=frozenset({'CARAPUNGO'}),
items_add=frozenset({'9713'}), confidence=0.3273657289002558,
lift=49.1841432225064)]), RelationRecord(items=frozenset({'1', '83', '128'}),
support=0.00571993136082367,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'1', '128'}), confidence=0.3819444444444444,
lift=26.232762896825395), OrderedStatistic(items_base=frozenset({'1', '128'}),
items_add=frozenset({'83'}), confidence=0.3928571428571428,
lift=26.232762896825392), OrderedStatistic(items_base=frozenset({'1', '83'}),
items_add=frozenset({'128'}), confidence=0.8461538461538461,
lift=26.41620879120879)]), RelationRecord(items=frozenset({'1', '128',
'PONCEANO'}), support=0.014559825282096615,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'128'}),
items_add=frozenset({'1', 'PONCEANO'}), confidence=0.45454545454545453,
lift=12.346558808423216), OrderedStatistic(items_base=frozenset({'1', '128'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629),
OrderedStatistic(items_base=frozenset({'1', 'PONCEANO'}),
items_add=frozenset({'128'}), confidence=0.39548022598870064,
lift=12.346558808423216)]), RelationRecord(items=frozenset({'1', '2256',
'KENNEDY'}), support=0.006239925120898549,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2256'}),
items_add=frozenset({'1', 'KENNEDY'}), confidence=0.6,
lift=4.609908110267678)]), RelationRecord(items=frozenset({'3226', 'SAN ISIDRO
DEL INCA', '1'}), support=0.0068639176329884045,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'3226'}),
items_add=frozenset({'SAN ISIDRO DEL INCA', '1'}),
confidence=0.33333333333333337, lift=7.613222486144102),
OrderedStatistic(items_base=frozenset({'3226', '1'}), items_add=frozenset({'SAN
ISIDRO DEL INCA'}), confidence=1.0, lift=8.741363636363637)]),
RelationRecord(items=frozenset({'1', '7684', 'KENNEDY'}),
support=0.009359887681347824,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'7684'}),
items_add=frozenset({'1', 'KENNEDY'}), confidence=0.8333333333333334,
lift=6.402650153149554)]), RelationRecord(items=frozenset({'1', '83',
'PONCEANO'}), support=0.00571993136082367,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'1', 'PONCEANO'}), confidence=0.3819444444444444,
lift=10.374538998744507), OrderedStatistic(items_base=frozenset({'1', '83'}),

items_add=frozenset({'PONCEANO'}), confidence=0.8461538461538461,
lift=9.314473162784553)]), RelationRecord(items=frozenset({'1', 'CHILLOGALLO',
'8919'}), support=0.005823930112838646,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8919'}),
items_add=frozenset({'1', 'CHILLOGALLO'}), confidence=0.3902439024390244,
lift=58.17659292871999), OrderedStatistic(items_base=frozenset({'CHILLOGALLO'}),
items_add=frozenset({'1', '8919'}), confidence=0.3533123028391167,
lift=60.665615141955826), OrderedStatistic(items_base=frozenset({'1', '8919'}),
items_add=frozenset({'CHILLOGALLO'}), confidence=1.0, lift=60.66561514195583),
OrderedStatistic(items_base=frozenset({'1', 'CHILLOGALLO'}),
items_add=frozenset({'8919'}), confidence=0.868217054263566,
lift=58.17659292871999)]), RelationRecord(items=frozenset({'MARCADORES
TUMORALES', 'KENNEDY', '10'}), support=0.007851905777130675,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'10'}),
items_add=frozenset({'MARCADORES TUMORALES', 'KENNEDY'}),
confidence=0.361244019138756, lift=46.007177033492816),
OrderedStatistic(items_base=frozenset({'MARCADORES TUMORALES'}),
items_add=frozenset({'KENNEDY', '10'}), confidence=0.361244019138756,
lift=46.007177033492816), OrderedStatistic(items_base=frozenset({'KENNEDY',
'10'}), items_add=frozenset({'MARCADORES TUMORALES'}), confidence=1.0,
lift=46.00717703349282), OrderedStatistic(items_base=frozenset({'MARCADORES
TUMORALES', 'KENNEDY'}), items_add=frozenset({'10'}), confidence=1.0,
lift=46.00717703349282)]), RelationRecord(items=frozenset({'INFECCIOSAS', '11',
'KENNEDY'}), support=0.01570381155426135,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'11'}),
items_add=frozenset({'INFECCIOSAS', 'KENNEDY'}), confidence=0.26468010517090274,
lift=16.854513584574935),
OrderedStatistic(items_base=frozenset({'INFECCIOSAS'}),
items_add=frozenset({'11', 'KENNEDY'}), confidence=0.26584507042253525,
lift=16.652985502592102), OrderedStatistic(items_base=frozenset({'11',
'KENNEDY'}), items_add=frozenset({'INFECCIOSAS'}), confidence=0.98371335504886,
lift=16.652985502592102), OrderedStatistic(items_base=frozenset({'INFECCIOSAS',
'KENNEDY'}), items_add=frozenset({'11'}), confidence=1.0,
lift=16.854513584574935)]), RelationRecord(items=frozenset({'INFECCIOSAS', 'SAN
ISIDRO DEL INCA', '11'}), support=0.009255888929332849,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL
INCA', '11'}), items_add=frozenset({'INFECCIOSAS'}), confidence=1.0,
lift=16.928697183098592), OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL
INCA', 'INFECCIOSAS'}), items_add=frozenset({'11'}), confidence=1.0,
lift=16.854513584574935)]), RelationRecord(items=frozenset({'8', '128',
'MICROBIOLOGIA'}), support=0.005823930112838646,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8', '128'}),
items_add=frozenset({'MICROBIOLOGIA'}), confidence=1.0,
lift=17.708103130755063),
OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA', '128'}),
items_add=frozenset({'8'}), confidence=1.0, lift=17.708103130755063)]),
RelationRecord(items=frozenset({'8', 'PONCEANO', '128'}),
support=0.005823930112838646,

ordered_statistics=[OrderedStatistic(items_base=frozenset({'8', '128'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629),
OrderedStatistic(items_base=frozenset({'8', 'PONCEANO'}),
items_add=frozenset({'128'}), confidence=0.717948717948718,
lift=22.413752913752912)]), RelationRecord(items=frozenset({'83', '128',
'PONCEANO'}), support=0.012583848993812074,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'128'}),
items_add=frozenset({'83', 'PONCEANO'}), confidence=0.39285714285714285,
lift=31.219155844155843), OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'PONCEANO', '128'}), confidence=0.8402777777777778,
lift=26.232762896825395), OrderedStatistic(items_base=frozenset({'83', '128'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629),
OrderedStatistic(items_base=frozenset({'PONCEANO', '128'}),
items_add=frozenset({'83'}), confidence=0.39285714285714285,
lift=26.232762896825395), OrderedStatistic(items_base=frozenset({'83',
'PONCEANO'}), items_add=frozenset({'128'}), confidence=1.0,
lift=31.219155844155843)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'83', '128'}), support=0.00571993136082367,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'QUIMICA CLINICA', '128'}), confidence=0.3819444444444444,
lift=26.232762896825395), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '128'}), items_add=frozenset({'83'}), confidence=0.3928571428571428,
lift=26.232762896825392), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '83'}), items_add=frozenset({'128'}), confidence=0.8461538461538461,
lift=26.41620879120879)]), RelationRecord(items=frozenset({'PONCEANO', '85',
'128'}), support=0.009151890177317872,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'128'}),
items_add=frozenset({'PONCEANO', '85'}), confidence=0.2857142857142857,
lift=31.219155844155846), OrderedStatistic(items_base=frozenset({'85'}),
items_add=frozenset({'PONCEANO', '128'}), confidence=0.55,
lift=17.170535714285716), OrderedStatistic(items_base=frozenset({'85', '128'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629),
OrderedStatistic(items_base=frozenset({'PONCEANO', '128'}),
items_add=frozenset({'85'}), confidence=0.2857142857142857,
lift=17.170535714285716), OrderedStatistic(items_base=frozenset({'PONCEANO',
'85'}), items_add=frozenset({'128'}), confidence=1.0,
lift=31.219155844155843)]), RelationRecord(items=frozenset({'PONCEANO', '128',
'86'}), support=0.0068639176329884045,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'128'}),
items_add=frozenset({'PONCEANO', '86'}), confidence=0.2142857142857143,
lift=30.300945378151262), OrderedStatistic(items_base=frozenset({'86'}),
items_add=frozenset({'PONCEANO', '128'}), confidence=0.6255924170616114,
lift=19.530467163168584), OrderedStatistic(items_base=frozenset({'128', '86'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629),
OrderedStatistic(items_base=frozenset({'PONCEANO', '128'}),
items_add=frozenset({'86'}), confidence=0.2142857142857143,
lift=19.530467163168584), OrderedStatistic(items_base=frozenset({'PONCEANO',
'86'}), items_add=frozenset({'128'}), confidence=0.9705882352941178,

lift=30.300945378151262)]), RelationRecord(items=frozenset({'MICROBIOLOGIA',
'PONCEANO', '128'}), support=0.005823930112838646,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA',
'128'}), items_add=frozenset({'PONCEANO'}), confidence=1.0,
lift=11.00801373783629), OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA',
'PONCEANO'}), items_add=frozenset({'128'}), confidence=0.717948717948718,
lift=22.413752913752912)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'PONCEANO', '128'}), support=0.014559825282096615,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'128'}),
items_add=frozenset({'QUIMICA CLINICA', 'PONCEANO'}),
confidence=0.45454545454545453, lift=12.346558808423216),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '128'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', 'PONCEANO'}),
items_add=frozenset({'128'}), confidence=0.39548022598870064,
lift=12.346558808423216)]), RelationRecord(items=frozenset({'13', 'KENNEDY',
'COAGULACION'}), support=0.01221985336175966,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'13'}),
items_add=frozenset({'KENNEDY', 'COAGULACION'}), confidence=0.39036544850498345,
lift=31.945182724252494),
OrderedStatistic(items_base=frozenset({'COAGULACION'}),
items_add=frozenset({'13', 'KENNEDY'}), confidence=0.39297658862876256,
lift=31.620639229789678), OrderedStatistic(items_base=frozenset({'13',
'KENNEDY'}), items_add=frozenset({'COAGULACION'}),
confidence=0.9832635983263599, lift=31.620639229789678),
OrderedStatistic(items_base=frozenset({'KENNEDY', 'COAGULACION'}),
items_add=frozenset({'13'}), confidence=1.0, lift=31.945182724252494)]),
RelationRecord(items=frozenset({'KENNEDY', 'PRUEBAS ESPECIALES', '14'}),
support=0.006915917008995892,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'KENNEDY', '14'}),
items_add=frozenset({'PRUEBAS ESPECIALES'}), confidence=1.0,
lift=27.670503597122302), OrderedStatistic(items_base=frozenset({'PRUEBAS
ESPECIALES', 'KENNEDY'}), items_add=frozenset({'14'}), confidence=1.0,
lift=25.337285902503293)]), RelationRecord(items=frozenset({'SAN ISIDRO DEL
INCA', '1687', '39'}), support=0.008891893297280433,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'1687'}),
items_add=frozenset({'SAN ISIDRO DEL INCA', '39'}), confidence=1.0,
lift=74.25096525096525), OrderedStatistic(items_base=frozenset({'39'}),
items_add=frozenset({'SAN ISIDRO DEL INCA', '1687'}),
confidence=0.36774193548387096, lift=41.35698924731183),
OrderedStatistic(items_base=frozenset({'1687', '39'}), items_add=frozenset({'SAN
ISIDRO DEL INCA'}), confidence=1.0, lift=8.741363636363637),
OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL INCA', '1687'}),
items_add=frozenset({'39'}), confidence=1.0, lift=41.35698924731183),
OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL INCA', '39'}),
items_add=frozenset({'1687'}), confidence=0.6602316602316602,
lift=74.25096525096525)]), RelationRecord(items=frozenset({'HORMONAS', '47',
'2'}), support=0.0061879257448910615,

ordered_statistics=[OrderedStatistic(items_base=frozenset({'47', '2'}),
items_add=frozenset({'HORMONAS'}), confidence=1.0, lift=7.78582995951417),
OrderedStatistic(items_base=frozenset({'HORMONAS', '47'}),
items_add=frozenset({'2'}), confidence=1.0, lift=7.77953074433657)]),
RelationRecord(items=frozenset({'981', 'HORMONAS', '2'}),
support=0.006551921376943477,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'981', '2'}),
items_add=frozenset({'HORMONAS'}), confidence=1.0, lift=7.78582995951417),
OrderedStatistic(items_base=frozenset({'981', 'HORMONAS'}),
items_add=frozenset({'2'}), confidence=1.0, lift=7.77953074433657)]),
RelationRecord(items=frozenset({'981', 'KENNEDY', '2'}),
support=0.006551921376943477,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'981'}),
items_add=frozenset({'KENNEDY', '2'}), confidence=0.3,
lift=6.623765786452354)]), RelationRecord(items=frozenset({'HORMONAS',
'CARCELÉN', '2'}), support=0.006759918880973428,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'CARCELÉN', '2'}),
items_add=frozenset({'HORMONAS'}), confidence=1.0, lift=7.78582995951417),
OrderedStatistic(items_base=frozenset({'HORMONAS', 'CARCELÉN'}),
items_add=frozenset({'2'}), confidence=1.0, lift=7.77953074433657)]),
RelationRecord(items=frozenset({'HORMONAS', 'COTOCOLLAO', '2'}),
support=0.006759918880973428,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'COTOCOLLAO', '2'}),
items_add=frozenset({'HORMONAS'}), confidence=1.0, lift=7.78582995951417),
OrderedStatistic(items_base=frozenset({'HORMONAS', 'COTOCOLLAO'}),
items_add=frozenset({'2'}), confidence=1.0, lift=7.77953074433657)]),
RelationRecord(items=frozenset({'HORMONAS', 'KENNEDY', '2'}),
support=0.045187457750507,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2'}),
items_add=frozenset({'HORMONAS', 'KENNEDY'}), confidence=0.351537216828479,
lift=7.7795307443365695), OrderedStatistic(items_base=frozenset({'HORMONAS'}),
items_add=frozenset({'KENNEDY', '2'}), confidence=0.3518218623481782,
lift=7.7679520491593745), OrderedStatistic(items_base=frozenset({'KENNEDY',
'2'}), items_add=frozenset({'HORMONAS'}), confidence=0.9977037887485649,
lift=7.767952049159374), OrderedStatistic(items_base=frozenset({'HORMONAS',
'KENNEDY'}), items_add=frozenset({'2'}), confidence=1.0,
lift=7.77953074433657)]), RelationRecord(items=frozenset({'PONCEANO',
'HORMONAS', '2'}), support=0.009879881441422702,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'PONCEANO', '2'}),
items_add=frozenset({'HORMONAS'}), confidence=1.0, lift=7.78582995951417),
OrderedStatistic(items_base=frozenset({'PONCEANO', 'HORMONAS'}),
items_add=frozenset({'2'}), confidence=1.0, lift=7.77953074433657)]),
RelationRecord(items=frozenset({'SAN ISIDRO DEL INCA', 'HORMONAS', '2'}),
support=0.016379803442358693,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL
INCA', '2'}), items_add=frozenset({'HORMONAS'}), confidence=1.0,
lift=7.78582995951417), OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL
INCA', 'HORMONAS'}), items_add=frozenset({'2'}), confidence=1.0,

lift=7.77953074433657)]), RelationRecord(items=frozenset({'68', 'COMITÉ DEL
PUEBLO', '218'}), support=0.007279912641048308,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'218'}),
items_add=frozenset({'COMITÉ DEL PUEBLO', '68'}), confidence=1.0,
lift=136.39007092198582), OrderedStatistic(items_base=frozenset({'68'}),
items_add=frozenset({'COMITÉ DEL PUEBLO', '218'}),
confidence=0.47619047619047616, lift=65.41156462585033),
OrderedStatistic(items_base=frozenset({'68', '218'}),
items_add=frozenset({'COMITÉ DEL PUEBLO'}), confidence=1.0,
lift=25.641333333333332), OrderedStatistic(items_base=frozenset({'COMITÉ DEL
PUEBLO', '218'}), items_add=frozenset({'68'}), confidence=1.0,
lift=65.41156462585033), OrderedStatistic(items_base=frozenset({'COMITÉ DEL
PUEBLO', '68'}), items_add=frozenset({'218'}), confidence=0.9929078014184397,
lift=136.3900709219858)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'2256', 'KENNEDY'}), support=0.006239925120898549,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2256'}),
items_add=frozenset({'QUIMICA CLINICA', 'KENNEDY'}), confidence=0.6,
lift=4.611750599520384)]), RelationRecord(items=frozenset({'36', '2563',
'IÑAQUITO'}), support=0.005927928864853622,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2563'}),
items_add=frozenset({'36', 'IÑAQUITO'}), confidence=0.6000000000000001,
lift=101.21578947368423), OrderedStatistic(items_base=frozenset({'36'}),
items_add=frozenset({'IÑAQUITO', '2563'}), confidence=0.27014218009478674,
lift=27.342654028436023), OrderedStatistic(items_base=frozenset({'IÑAQUITO'}),
items_add=frozenset({'36', '2563'}), confidence=0.21714285714285714,
lift=36.63047619047619), OrderedStatistic(items_base=frozenset({'36', '2563'}),
items_add=frozenset({'IÑAQUITO'}), confidence=1.0, lift=36.63047619047619),
OrderedStatistic(items_base=frozenset({'IÑAQUITO', '2563'}),
items_add=frozenset({'36'}), confidence=0.6000000000000001,
lift=27.342654028436023), OrderedStatistic(items_base=frozenset({'36',
'IÑAQUITO'}), items_add=frozenset({'2563'}), confidence=1.0,
lift=101.21578947368421)]), RelationRecord(items=frozenset({'3226', 'SAN ISIDRO
DEL INCA', '29'}), support=0.011231865217617388,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'29'}),
items_add=frozenset({'3226', 'SAN ISIDRO DEL INCA'}),
confidence=0.4114285714285714, lift=19.98025974025974),
OrderedStatistic(items_base=frozenset({'3226'}), items_add=frozenset({'SAN
ISIDRO DEL INCA', '29'}), confidence=0.5454545454545454,
lift=38.99493071983778), OrderedStatistic(items_base=frozenset({'3226', '29'}),
items_add=frozenset({'SAN ISIDRO DEL INCA'}), confidence=1.0,
lift=8.741363636363637), OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL
INCA', '29'}), items_add=frozenset({'3226'}), confidence=0.8029739776951672,
lift=38.994930719837775), OrderedStatistic(items_base=frozenset({'3226', 'SAN
ISIDRO DEL INCA'}), items_add=frozenset({'29'}), confidence=0.5454545454545454,
lift=19.98025974025974)]), RelationRecord(items=frozenset({'3', 'KENNEDY',
'HEMATOLOGIA CLINICA'}), support=0.02511569861161666,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'3'}),
items_add=frozenset({'KENNEDY', 'HEMATOLOGIA CLINICA'}),

confidence=0.3163064833005894, lift=12.5939751146038),
OrderedStatistic(items_base=frozenset({'HEMATOLOGIA CLINICA'}),
items_add=frozenset({'3', 'KENNEDY'}), confidence=0.3163064833005894,
lift=12.5939751146038), OrderedStatistic(items_base=frozenset({'3', 'KENNEDY'}),
items_add=frozenset({'HEMATOLOGIA CLINICA'}), confidence=1.0,
lift=12.593975114603799), OrderedStatistic(items_base=frozenset({'KENNEDY',
'HEMATOLOGIA CLINICA'}), items_add=frozenset({'3'}), confidence=1.0,
lift=12.593975114603799)]), RelationRecord(items=frozenset({'3', 'PONCEANO',
'HEMATOLOGIA CLINICA'}), support=0.008839893921272945,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'3', 'PONCEANO'}),
items_add=frozenset({'HEMATOLOGIA CLINICA'}), confidence=1.0,
lift=12.593975114603799), OrderedStatistic(items_base=frozenset({'PONCEANO',
'HEMATOLOGIA CLINICA'}), items_add=frozenset({'3'}), confidence=1.0,
lift=12.593975114603799)]), RelationRecord(items=frozenset({'3', 'SAN ISIDRO DEL
INCA', 'HEMATOLOGIA CLINICA'}), support=0.009567885185377776,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'3', 'SAN ISIDRO DEL
INCA'}), items_add=frozenset({'HEMATOLOGIA CLINICA'}), confidence=1.0,
lift=12.593975114603799), OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL
INCA', 'HEMATOLOGIA CLINICA'}), items_add=frozenset({'3'}), confidence=1.0,
lift=12.593975114603799)]), RelationRecord(items=frozenset({'3226', 'SAN ISIDRO
DEL INCA', '31'}), support=0.005615932608808694,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'3226'}),
items_add=frozenset({'SAN ISIDRO DEL INCA', '31'}),
confidence=0.2727272727272727, lift=22.223805855161785),
OrderedStatistic(items_base=frozenset({'3226', '31'}), items_add=frozenset({'SAN
ISIDRO DEL INCA'}), confidence=1.0, lift=8.741363636363637),
OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL INCA', '31'}),
items_add=frozenset({'3226'}), confidence=0.45762711864406774,
lift=22.223805855161785), OrderedStatistic(items_base=frozenset({'3226', 'SAN
ISIDRO DEL INCA'}), items_add=frozenset({'31'}), confidence=0.2727272727272727,
lift=7.1358070500927635)]), RelationRecord(items=frozenset({'3226', 'QUIMICA
CLINICA', 'SAN ISIDRO DEL INCA'}), support=0.0068639176329884045,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'3226'}),
items_add=frozenset({'QUIMICA CLINICA', 'SAN ISIDRO DEL INCA'}),
confidence=0.33333333333333337, lift=7.613222486144102),
OrderedStatistic(items_base=frozenset({'3226', 'QUIMICA CLINICA'}),
items_add=frozenset({'SAN ISIDRO DEL INCA'}), confidence=1.0,
lift=8.741363636363637)]), RelationRecord(items=frozenset({'SEROLOGIA', '4',
'KENNEDY'}), support=0.007955904529145 65,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'4'}),
items_add=frozenset({'SEROLOGIA', 'KENNEDY'}), confidence=0.24324324324324328,
lift=30.57392686804452), OrderedStatistic(items_base=frozenset({'SEROLOGIA'}),
items_add=frozenset({'4', 'KENNEDY'}), confidence=0.24401913875598086,
lift=30.472286087118622), OrderedStatistic(items_base=frozenset({'4',
'KENNEDY'}), items_add=frozenset({'SEROLOGIA'}), confidence=0.9935064935064936,
lift=30.472286087118622), OrderedStatistic(items_base=frozenset({'SEROLOGIA',
'KENNEDY'}), items_add=frozenset({'4'}), confidence=1.0,
lift=30.573926868044516)]), RelationRecord(items=frozenset({'BELISARIO QUEVEDO',

'42', '5372'}), support=0.007279912641048308,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'42'}),
items_add=frozenset({'BELISARIO QUEVEDO', '5372'}),
confidence=0.3218390804597701, lift=29.472796934865897),
OrderedStatistic(items_base=frozenset({'5372'}), items_add=frozenset({'BELISARIO
QUEVEDO', '42'}), confidence=0.6666666666666666, lift=91.57619047619046),
OrderedStatistic(items_base=frozenset({'BELISARIO QUEVEDO'}),
items_add=frozenset({'42', '5372'}), confidence=0.356234096692112,
lift=48.93384223918575), OrderedStatistic(items_base=frozenset({'42', '5372'}),
items_add=frozenset({'BELISARIO QUEVEDO'}), confidence=1.0,
lift=48.93384223918575), OrderedStatistic(items_base=frozenset({'BELISARIO
QUEVEDO', '42'}), items_add=frozenset({'5372'}), confidence=1.0,
lift=91.57619047619048), OrderedStatistic(items_base=frozenset({'BELISARIO
QUEVEDO', '5372'}), items_add=frozenset({'42'}), confidence=0.6666666666666666,
lift=29.472796934865897)]), RelationRecord(items=frozenset({'CARAPUNGO', '47',
'9713'}), support=0.0066559201289584525,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'47'}),
items_add=frozenset({'9713', 'CARAPUNGO'}), confidence=0.2277580071174377,
lift=34.21886120996441), OrderedStatistic(items_base=frozenset({'9713'}),
items_add=frozenset({'47', 'CARAPUNGO'}), confidence=1.0, lift=150.2421875),
OrderedStatistic(items_base=frozenset({'CARAPUNGO'}), items_add=frozenset({'47',
'9713'}), confidence=0.3273657289002558, lift=49.1841432225064),
OrderedStatistic(items_base=frozenset({'47', '9713'}),
items_add=frozenset({'CARAPUNGO'}), confidence=1.0, lift=49.18414322250639),
OrderedStatistic(items_base=frozenset({'47', 'CARAPUNGO'}),
items_add=frozenset({'9713'}), confidence=1.0, lift=150.2421875),
OrderedStatistic(items_base=frozenset({'CARAPUNGO', '9713'}),
items_add=frozenset({'47'}), confidence=1.0, lift=34.21886120996441)]),
RelationRecord(items=frozenset({'981', '47', 'KENNEDY'}),
support=0.009359887681347824,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'47'}),
items_add=frozenset({'981', 'KENNEDY'}), confidence=0.3202846975088968,
lift=14.665226232841892), OrderedStatistic(items_base=frozenset({'981'}),
items_add=frozenset({'KENNEDY', '47'}), confidence=0.4285714285714286,
lift=31.10134770889488), OrderedStatistic(items_base=frozenset({'KENNEDY',
'47'}), items_add=frozenset({'981'}), confidence=0.6792452830188679,
lift=31.101347708894878), OrderedStatistic(items_base=frozenset({'981',
'KENNEDY'}), items_add=frozenset({'47'}), confidence=0.4285714285714286,
lift=14.665226232841892)]), RelationRecord(items=frozenset({'981', 'KENNEDY',
'48'}), support=0.006239925120898549,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'48'}),
items_add=frozenset({'981', 'KENNEDY'}), confidence=0.38961038961038963,
lift=17.83951762523191), OrderedStatistic(items_base=frozenset({'981'}),
items_add=frozenset({'KENNEDY', '48'}), confidence=0.2857142857142857,
lift=31.57799671592775), OrderedStatistic(items_base=frozenset({'KENNEDY',
'48'}), items_add=frozenset({'981'}), confidence=0.6896551724137931,
lift=31.57799671592775), OrderedStatistic(items_base=frozenset({'981',
'KENNEDY'}), items_add=frozenset({'48'}), confidence=0.2857142857142857,

lift=17.83951762523191)]), RelationRecord(items=frozenset({'KENNEDY',
'URIANALISIS', '5'}), support=0.027091674899901202,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'5'}),
items_add=frozenset({'KENNEDY', 'URIANALISIS'}), confidence=0.3905547226386807,
lift=14.416041979010496),
OrderedStatistic(items_base=frozenset({'URIANALISIS'}),
items_add=frozenset({'KENNEDY', '5'}), confidence=0.3905547226386807,
lift=14.416041979010496), OrderedStatistic(items_base=frozenset({'KENNEDY',
'5'}), items_add=frozenset({'URIANALISIS'}), confidence=1.0,
lift=14.416041979010496), OrderedStatistic(items_base=frozenset({'KENNEDY',
'URIANALISIS'}), items_add=frozenset({'5'}), confidence=1.0,
lift=14.416041979010496)]), RelationRecord(items=frozenset({'SAN ISIDRO DEL
INCA', 'URIANALISIS', '5'}), support=0.007851905777130675,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL
INCA', '5'}), items_add=frozenset({'URIANALISIS'}), confidence=1.0,
lift=14.416041979010496), OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL
INCA', 'URIANALISIS'}), items_add=frozenset({'5'}), confidence=1.0,
lift=14.416041979010496)]), RelationRecord(items=frozenset({'KENNEDY', '6',
'COPROANALISIS'}), support=0.026467682387811345,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'6'}),
items_add=frozenset({'KENNEDY', 'COPROANALISIS'}),
confidence=0.42558528428093645, lift=16.07943143812709),
OrderedStatistic(items_base=frozenset({'COPROANALISIS'}),
items_add=frozenset({'KENNEDY', '6'}), confidence=0.4343003412969283,
lift=16.344481141841936), OrderedStatistic(items_base=frozenset({'KENNEDY',
'6'}), items_add=frozenset({'COPROANALISIS'}), confidence=0.9960861056751468,
lift=16.344481141841936), OrderedStatistic(items_base=frozenset({'KENNEDY',
'COPROANALISIS'}), items_add=frozenset({'6'}), confidence=1.0,
lift=16.07943143812709)]), RelationRecord(items=frozenset({'SAN ISIDRO DEL
INCA', '6', 'COPROANALISIS'}), support=0.006395923248921013,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL
INCA', '6'}), items_add=frozenset({'COPROANALISIS'}), confidence=1.0,
lift=16.408703071672353), OrderedStatistic(items_base=frozenset({'SAN ISIDRO DEL
INCA', 'COPROANALISIS'}), items_add=frozenset({'6'}), confidence=1.0,
lift=16.07943143812709)]), RelationRecord(items=frozenset({'7684', 'KENNEDY',
'60'}), support=0.006551921376943477,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'60'}),
items_add=frozenset({'7684', 'KENNEDY'}), confidence=0.4131147540983607,
lift=36.78060109289618), OrderedStatistic(items_base=frozenset({'7684'}),
items_add=frozenset({'KENNEDY', '60'}), confidence=0.5833333333333334,
lift=73.32080610021787), OrderedStatistic(items_base=frozenset({'KENNEDY',
'60'}), items_add=frozenset({'7684'}), confidence=0.8235294117647058,
lift=73.32080610021787), OrderedStatistic(items_base=frozenset({'7684',
'KENNEDY'}), items_add=frozenset({'60'}), confidence=0.5833333333333334,
lift=36.78060109289618)]), RelationRecord(items=frozenset({'66', 'RUMIPAMBA',
'7629'}), support=0.009359887681347824,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'66'}),
items_add=frozenset({'RUMIPAMBA', '7629'}), confidence=0.5901639344262296,

lift=63.05245901639345), OrderedStatistic(items_base=frozenset({'7629'}),
items_add=frozenset({'66', 'RUMIPAMBA'}), confidence=1.0,
lift=106.83888888888889), OrderedStatistic(items_base=frozenset({'RUMIPAMBA'}),
items_add=frozenset({'66', '7629'}), confidence=0.6923076923076924,
lift=73.96538461538462), OrderedStatistic(items_base=frozenset({'66', '7629'}),
items_add=frozenset({'RUMIPAMBA'}), confidence=1.0, lift=73.96538461538462),
OrderedStatistic(items_base=frozenset({'66', 'RUMIPAMBA'}),
items_add=frozenset({'7629'}), confidence=1.0, lift=106.83888888888889),
OrderedStatistic(items_base=frozenset({'RUMIPAMBA', '7629'}),
items_add=frozenset({'66'}), confidence=1.0, lift=63.05245901639345)]),
RelationRecord(items=frozenset({'QUIMICA CLINICA', '7684', 'KENNEDY'}),
support=0.009359887681347824,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'7684'}),
items_add=frozenset({'QUIMICA CLINICA', 'KENNEDY'}),
confidence=0.8333333333333334, lift=6.4052091660005335)]),
RelationRecord(items=frozenset({'77', 'CHILLOGALLO', '8919'}),
support=0.008527897665228018,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'77'}),
items_add=frozenset({'CHILLOGALLO', '8919'}), confidence=0.5173501577287066,
lift=34.666065795403334), OrderedStatistic(items_base=frozenset({'8919'}),
items_add=frozenset({'77', 'CHILLOGALLO'}), confidence=0.5714285714285715,
lift=67.00696864111498), OrderedStatistic(items_base=frozenset({'CHILLOGALLO'}),
items_add=frozenset({'77', '8919'}), confidence=0.5173501577287066,
lift=60.66561514195583), OrderedStatistic(items_base=frozenset({'77', '8919'}),
items_add=frozenset({'CHILLOGALLO'}), confidence=1.0, lift=60.66561514195583),
OrderedStatistic(items_base=frozenset({'77', 'CHILLOGALLO'}),
items_add=frozenset({'8919'}), confidence=1.0, lift=67.00696864111498),
OrderedStatistic(items_base=frozenset({'CHILLOGALLO', '8919'}),
items_add=frozenset({'77'}), confidence=0.5714285714285715,
lift=34.666065795403334)]), RelationRecord(items=frozenset({'8', 'KENNEDY',
'MICROBIOLOGIA'}), support=0.027195673651916177,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8'}),
items_add=frozenset({'MICROBIOLOGIA', 'KENNEDY'}),
confidence=0.48158379373848986, lift=17.708103130755063),
OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA'}),
items_add=frozenset({'8', 'KENNEDY'}), confidence=0.48158379373848986,
lift=17.708103130755063), OrderedStatistic(items_base=frozenset({'8',
'KENNEDY'}), items_add=frozenset({'MICROBIOLOGIA'}), confidence=1.0,
lift=17.708103130755063),
OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA', 'KENNEDY'}),
items_add=frozenset({'8'}), confidence=1.0, lift=17.708103130755063)]),
RelationRecord(items=frozenset({'8', 'PONCEANO', 'MICROBIOLOGIA'}),
support=0.008111902657168114,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8', 'PONCEANO'}),
items_add=frozenset({'MICROBIOLOGIA'}), confidence=1.0,
lift=17.708103130755063),
OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA', 'PONCEANO'}),
items_add=frozenset({'8'}), confidence=1.0, lift=17.708103130755063)]),

RelationRecord(items=frozenset({'8', 'SAN ISIDRO DEL INCA', 'MICROBIOLOGIA'}),
support=0.005927928864853622,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8', 'SAN ISIDRO DEL
INCA'}), items_add=frozenset({'MICROBIOLOGIA'}), confidence=1.0,
lift=17.708103130755063),
OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA', 'SAN ISIDRO DEL INCA'}),
items_add=frozenset({'8'}), confidence=1.0, lift=17.708103130755063)]),
RelationRecord(items=frozenset({'QUIMICA CLINICA', '83', 'PONCEANO'}),
support=0.00571993136082367,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'QUIMICA CLINICA', 'PONCEANO'}),
confidence=0.3819444444444444, lift=10.374538998744507),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '83'}),
items_add=frozenset({'PONCEANO'}), confidence=0.8461538461538461,
lift=9.314473162784553)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'CHILLOGALLO', '8919'}), support=0.005823930112838646,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8919'}),
items_add=frozenset({'QUIMICA CLINICA', 'CHILLOGALLO'}),
confidence=0.3902439024390244, lift=58.17659292871999),
OrderedStatistic(items_base=frozenset({'CHILLOGALLO'}),
items_add=frozenset({'QUIMICA CLINICA', '8919'}), confidence=0.3533123028391167,
lift=60.665615141955826), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '8919'}), items_add=frozenset({'CHILLOGALLO'}), confidence=1.0,
lift=60.66561514195583), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', 'CHILLOGALLO'}), items_add=frozenset({'8919'}),
confidence=0.868217054263566, lift=58.17659292871999)]),
RelationRecord(items=frozenset({'981', 'HORMONAS', 'KENNEDY'}),
support=0.006551921376943477,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'981'}),
items_add=frozenset({'HORMONAS', 'KENNEDY'}), confidence=0.3,
lift=6.639010356731875)]), RelationRecord(items=frozenset({'1', '83', '128',
'PONCEANO'}), support=0.00571993136082367,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'1', '128', 'PONCEANO'}), confidence=0.3819444444444444,
lift=26.232762896825395), OrderedStatistic(items_base=frozenset({'1', '128'}),
items_add=frozenset({'83', 'PONCEANO'}), confidence=0.3928571428571428,
lift=31.21915584415584), OrderedStatistic(items_base=frozenset({'1', '83'}),
items_add=frozenset({'PONCEANO', '128'}), confidence=0.8461538461538461,
lift=26.41620879120879), OrderedStatistic(items_base=frozenset({'83', '128'}),
items_add=frozenset({'1', 'PONCEANO'}), confidence=0.45454545454545453,
lift=12.346558808423216), OrderedStatistic(items_base=frozenset({'83',
'PONCEANO'}), items_add=frozenset({'1', '128'}), confidence=0.45454545454545453,
lift=31.219155844155843), OrderedStatistic(items_base=frozenset({'1', '83',
'128'}), items_add=frozenset({'PONCEANO'}), confidence=1.0,
lift=11.00801373783629), OrderedStatistic(items_base=frozenset({'1', '128',
'PONCEANO'}), items_add=frozenset({'83'}), confidence=0.3928571428571428,
lift=26.232762896825392), OrderedStatistic(items_base=frozenset({'1', '83',
'PONCEANO'}), items_add=frozenset({'128'}), confidence=1.0,

lift=31.219155844155843)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'1', '83', '128'}), support=0.00571993136082367,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'QUIMICA CLINICA', '1', '128'}),
confidence=0.3819444444444444, lift=26.232762896825395),
OrderedStatistic(items_base=frozenset({'1', '128'}),
items_add=frozenset({'QUIMICA CLINICA', '83'}), confidence=0.3928571428571428,
lift=58.115659340659334), OrderedStatistic(items_base=frozenset({'1', '83'}),
items_add=frozenset({'QUIMICA CLINICA', '128'}), confidence=0.8461538461538461,
lift=58.115659340659334), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '128'}), items_add=frozenset({'1', '83'}),
confidence=0.3928571428571428, lift=58.115659340659334),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '83'}),
items_add=frozenset({'1', '128'}), confidence=0.8461538461538461,
lift=58.115659340659334), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '1', '128'}), items_add=frozenset({'83'}),
confidence=0.3928571428571428, lift=26.232762896825392),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '1', '83'}),
items_add=frozenset({'128'}), confidence=0.8461538461538461,
lift=26.41620879120879)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'1', '128', 'PONCEANO'}), support=0.014559825282096615,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'128'}),
items_add=frozenset({'QUIMICA CLINICA', '1', 'PONCEANO'}),
confidence=0.45454545454545453, lift=12.346558808423216),
OrderedStatistic(items_base=frozenset({'1', '128'}),
items_add=frozenset({'QUIMICA CLINICA', 'PONCEANO'}), confidence=1.0,
lift=27.162429378531076), OrderedStatistic(items_base=frozenset({'1',
'PONCEANO'}), items_add=frozenset({'QUIMICA CLINICA', '128'}),
confidence=0.39548022598870064, lift=27.162429378531076),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '128'}),
items_add=frozenset({'1', 'PONCEANO'}), confidence=1.0,
lift=27.162429378531076), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', 'PONCEANO'}), items_add=frozenset({'1', '128'}),
confidence=0.39548022598870064, lift=27.162429378531076),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '1', '128'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '1', 'PONCEANO'}),
items_add=frozenset({'128'}), confidence=0.39548022598870064,
lift=12.346558808423216)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'1', '2256', 'KENNEDY'}), support=0.006239925120898549,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'2256'}),
items_add=frozenset({'QUIMICA CLINICA', '1', 'KENNEDY'}), confidence=0.6,
lift=4.611750599520384), OrderedStatistic(items_base=frozenset({'1', '2256'}),
items_add=frozenset({'QUIMICA CLINICA', 'KENNEDY'}), confidence=1.0,
lift=7.68625099920064), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '2256'}), items_add=frozenset({'1', 'KENNEDY'}), confidence=1.0,
lift=7.683180183779465)]), RelationRecord(items=frozenset({'3226', 'QUIMICA
CLINICA', '1', 'SAN ISIDRO DEL INCA'}), support=0.0068639176329884045,

ordered_statistics=[OrderedStatistic(items_base=frozenset({'3226'}),
items_add=frozenset({'QUIMICA CLINICA', '1', 'SAN ISIDRO DEL INCA'}),
confidence=0.33333333333333337, lift=7.613222486144102),
OrderedStatistic(items_base=frozenset({'3226', '1'}),
items_add=frozenset({'QUIMICA CLINICA', 'SAN ISIDRO DEL INCA'}), confidence=1.0,
lift=22.839667458432306), OrderedStatistic(items_base=frozenset({'3226',
'QUIMICA CLINICA'}), items_add=frozenset({'SAN ISIDRO DEL INCA', '1'}),
confidence=1.0, lift=22.839667458432306),
OrderedStatistic(items_base=frozenset({'3226', 'QUIMICA CLINICA', '1'}),
items_add=frozenset({'SAN ISIDRO DEL INCA'}), confidence=1.0,
lift=8.741363636363637)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'1', 'KENNEDY', '60'}), support=0.005927928864853622,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'1', '60'}),
items_add=frozenset({'QUIMICA CLINICA', 'KENNEDY'}),
confidence=0.6514285714285715, lift=5.007043508050703),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '60'}),
items_add=frozenset({'1', 'KENNEDY'}), confidence=0.6514285714285715,
lift=5.0050430911477655)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'1', '7684', 'KENNEDY'}), support=0.009359887681347824,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'7684'}),
items_add=frozenset({'QUIMICA CLINICA', '1', 'KENNEDY'}),
confidence=0.8333333333333334, lift=6.4052091660005335),
OrderedStatistic(items_base=frozenset({'1', '7684'}),
items_add=frozenset({'QUIMICA CLINICA', 'KENNEDY'}), confidence=1.0,
lift=7.68625099920064), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '7684'}), items_add=frozenset({'1', 'KENNEDY'}), confidence=1.0,
lift=7.683180183779465)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'1', '83', 'PONCEANO'}), support=0.00571993136082367,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'QUIMICA CLINICA', '1', 'PONCEANO'}),
confidence=0.3819444444444444, lift=10.374538998744507),
OrderedStatistic(items_base=frozenset({'1', '83'}),
items_add=frozenset({'QUIMICA CLINICA', 'PONCEANO'}),
confidence=0.8461538461538461, lift=22.983594089526296),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '83'}),
items_add=frozenset({'1', 'PONCEANO'}), confidence=0.8461538461538461,
lift=22.983594089526296), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '1', '83'}), items_add=frozenset({'PONCEANO'}),
confidence=0.8461538461538461, lift=9.314473162784553)]),
RelationRecord(items=frozenset({'QUIMICA CLINICA', '1', 'CHILLOGALLO', '8919'}),
support=0.005823930112838646,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8919'}),
items_add=frozenset({'QUIMICA CLINICA', '1', 'CHILLOGALLO'}),
confidence=0.3902439024390244, lift=58.17659292871999),
OrderedStatistic(items_base=frozenset({'CHILLOGALLO'}),
items_add=frozenset({'QUIMICA CLINICA', '1', '8919'}),
confidence=0.3533123028391167, lift=60.665615141955826),
OrderedStatistic(items_base=frozenset({'1', '8919'}),

items_add=frozenset({'QUIMICA CLINICA', 'CHILLOGALLO'}), confidence=1.0,
lift=149.07751937984497), OrderedStatistic(items_base=frozenset({'1',
'CHILLOGALLO'}), items_add=frozenset({'QUIMICA CLINICA', '8919'}),
confidence=0.868217054263566, lift=149.07751937984497),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '8919'}),
items_add=frozenset({'1', 'CHILLOGALLO'}), confidence=1.0,
lift=149.07751937984497), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', 'CHILLOGALLO'}), items_add=frozenset({'1', '8919'}),
confidence=0.868217054263566, lift=149.07751937984497),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '1', '8919'}),
items_add=frozenset({'CHILLOGALLO'}), confidence=1.0, lift=60.66561514195583),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '1', 'CHILLOGALLO'}),
items_add=frozenset({'8919'}), confidence=0.868217054263566,
lift=58.17659292871999)]), RelationRecord(items=frozenset({'8', 'PONCEANO',
'128', 'MICROBIOLOGIA'}), support=0.005823930112838646,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'8', '128'}),
items_add=frozenset({'MICROBIOLOGIA', 'PONCEANO'}), confidence=1.0,
lift=123.27564102564102),
OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA', '128'}),
items_add=frozenset({'8', 'PONCEANO'}), confidence=1.0,
lift=123.27564102564102), OrderedStatistic(items_base=frozenset({'8',
'PONCEANO'}), items_add=frozenset({'MICROBIOLOGIA', '128'}),
confidence=0.717948717948718, lift=123.27564102564102),
OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA', 'PONCEANO'}),
items_add=frozenset({'8', '128'}), confidence=0.717948717948718,
lift=123.27564102564102), OrderedStatistic(items_base=frozenset({'8', '128',
'MICROBIOLOGIA'}), items_add=frozenset({'PONCEANO'}), confidence=1.0,
lift=11.00801373783629), OrderedStatistic(items_base=frozenset({'8', 'PONCEANO',
'128'}), items_add=frozenset({'MICROBIOLOGIA'}), confidence=1.0,
lift=17.708103130755063),
OrderedStatistic(items_base=frozenset({'MICROBIOLOGIA', 'PONCEANO', '128'}),
items_add=frozenset({'8'}), confidence=1.0, lift=17.708103130755063),
OrderedStatistic(items_base=frozenset({'8', 'PONCEANO', 'MICROBIOLOGIA'}),
items_add=frozenset({'128'}), confidence=0.717948717948718,
lift=22.413752913752912)]), RelationRecord(items=frozenset({'QUIMICA CLINICA',
'83', '128', 'PONCEANO'}), support=0.00571993136082367,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'QUIMICA CLINICA', 'PONCEANO', '128'}),
confidence=0.3819444444444444, lift=26.232762896825395),
OrderedStatistic(items_base=frozenset({'83', '128'}),
items_add=frozenset({'QUIMICA CLINICA', 'PONCEANO'}),
confidence=0.45454545454545453, lift=12.346558808423216),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '128'}),
items_add=frozenset({'83', 'PONCEANO'}), confidence=0.3928571428571428,
lift=31.21915584415584), OrderedStatistic(items_base=frozenset({'83',
'PONCEANO'}), items_add=frozenset({'QUIMICA CLINICA', '128'}),
confidence=0.45454545454545453, lift=31.219155844155843),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '83'}),

items_add=frozenset({'PONCEANO', '128'}), confidence=0.8461538461538461,
lift=26.41620879120879), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '83', '128'}), items_add=frozenset({'PONCEANO'}), confidence=1.0,
lift=11.00801373783629), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', 'PONCEANO', '128'}), items_add=frozenset({'83'}),
confidence=0.3928571428571428, lift=26.232762896825392),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '83', 'PONCEANO'}),
items_add=frozenset({'128'}), confidence=1.0, lift=31.219155844155843)]),
RelationRecord(items=frozenset({'981', 'HORMONAS', 'KENNEDY', '2'}),
support=0.006551921376943477,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'981'}),
items_add=frozenset({'HORMONAS', 'KENNEDY', '2'}), confidence=0.3,
lift=6.639010356731875), OrderedStatistic(items_base=frozenset({'981', '2'}),
items_add=frozenset({'HORMONAS', 'KENNEDY'}), confidence=1.0,
lift=22.130034522439583), OrderedStatistic(items_base=frozenset({'981',
'HORMONAS'}), items_add=frozenset({'KENNEDY', '2'}), confidence=1.0,
lift=22.079219288174514), OrderedStatistic(items_base=frozenset({'981',
'KENNEDY', '2'}), items_add=frozenset({'HORMONAS'}), confidence=1.0,
lift=7.78582995951417), OrderedStatistic(items_base=frozenset({'981',
'HORMONAS', 'KENNEDY'}), items_add=frozenset({'2'}), confidence=1.0,
lift=7.77953074433657)]), RelationRecord(items=frozenset({'83', 'PONCEANO', '1',
'QUIMICA CLINICA', '128'}), support=0.00571993136082367,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'83'}),
items_add=frozenset({'QUIMICA CLINICA', 'PONCEANO', '1', '128'}),
confidence=0.3819444444444444, lift=26.232762896825395),
OrderedStatistic(items_base=frozenset({'1', '128'}),
items_add=frozenset({'QUIMICA CLINICA', '83', 'PONCEANO'}),
confidence=0.3928571428571428, lift=68.68214285714285),
OrderedStatistic(items_base=frozenset({'1', '83'}),
items_add=frozenset({'QUIMICA CLINICA', 'PONCEANO', '128'}),
confidence=0.8461538461538461, lift=58.115659340659334),
OrderedStatistic(items_base=frozenset({'83', '128'}),
items_add=frozenset({'QUIMICA CLINICA', 'PONCEANO', '1'}),
confidence=0.45454545454545453, lift=12.346558808423216),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '128'}),
items_add=frozenset({'83', '1', 'PONCEANO'}), confidence=0.3928571428571428,
lift=68.68214285714285), OrderedStatistic(items_base=frozenset({'83',
'PONCEANO'}), items_add=frozenset({'QUIMICA CLINICA', '1', '128'}),
confidence=0.45454545454545453, lift=31.219155844155843),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '83'}),
items_add=frozenset({'PONCEANO', '1', '128'}), confidence=0.8461538461538461,
lift=58.115659340659334), OrderedStatistic(items_base=frozenset({'1', '83',
'128'}), items_add=frozenset({'QUIMICA CLINICA', 'PONCEANO'}), confidence=1.0,
lift=27.162429378531076), OrderedStatistic(items_base=frozenset({'1', '128',
'PONCEANO'}), items_add=frozenset({'QUIMICA CLINICA', '83'}),
confidence=0.3928571428571428, lift=58.115659340659334),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '1', '128'}),
items_add=frozenset({'83', 'PONCEANO'}), confidence=0.3928571428571428,

```
lift=31.21915584415584), OrderedStatistic(items_base=frozenset({'1', '83',
'PONCEANO'}), items_add=frozenset({'QUIMICA CLINICA', '128'}), confidence=1.0,
lift=68.68214285714285), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '1', '83'}), items_add=frozenset({'PONCEANO', '128'}),
confidence=0.8461538461538461, lift=26.41620879120879),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '83', '128'}),
items_add=frozenset({'PONCEANO', '1'}), confidence=1.0,
lift=27.162429378531076), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', 'PONCEANO', '128'}), items_add=frozenset({'83', '1'}),
confidence=0.3928571428571428, lift=58.115659340659334),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '83', 'PONCEANO'}),
items_add=frozenset({'1', '128'}), confidence=1.0, lift=68.68214285714285),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '1', '83', '128'}),
items_add=frozenset({'PONCEANO'}), confidence=1.0, lift=11.00801373783629),
OrderedStatistic(items_base=frozenset({'QUIMICA CLINICA', '1', '128',
'PONCEANO'}), items_add=frozenset({'83'}), confidence=0.3928571428571428,
lift=26.232762896825392), OrderedStatistic(items_base=frozenset({'QUIMICA
CLINICA', '1', '83', 'PONCEANO'}), items_add=frozenset({'128'}), confidence=1.0,
lift=31.219155844155843)])]
```

[43]:
```
for item in rslt:
  pair=item[0]
  items=[x for x in pair]
  print("Regla: "+items[0]+" -> "+items[1])
  print("Support: "+str(item[1]))
  print("Confidence: "+str(item[2][0][2]))
  print("Lift: "+str(item[2][0][3]))
  print("====================================")
```

```
Regla: MARCADORES TUMORALES -> 10
Support: 0.021735739171129947
Confidence: 1.0
Lift: 46.00717703349282
====================================
Regla: INFECCIOSAS -> 11
Support: 0.059071291144506266
Confidence: 0.9956178790534619
Lift: 16.854513584574935
====================================
Regla: 83 -> 128
Support: 0.012583848993812074
Confidence: 0.39285714285714285
Lift: 26.232762896825395
====================================
Regla: 85 -> 128
Support: 0.009151890177317872
Confidence: 0.2857142857142857
Lift: 17.170535714285716
```

```
====================================
Regla: 128 -> 86
Support: 0.0068639176329884045
Confidence: 0.2142857142857143
Lift: 19.530467163168584
====================================
Regla: PONCEANO -> 128
Support: 0.032031615620612554
Confidence: 1.0
Lift: 11.00801373783629
====================================
Regla: 13 -> COAGULACION
Support: 0.03109562685247777
Confidence: 0.9933554817275748
Lift: 31.945182724252494
====================================
Regla: PRUEBAS ESPECIALES -> 14
Support: 0.0361395663252041
Confidence: 0.9156785243741765
Lift: 25.337285902503293
====================================
Regla: 16 -> ANATOMIA PATOLOGICA
Support: 0.011231865217617388
Confidence: 0.96
Lift: 85.47111111111111
====================================
Regla: 1687 -> 39
Support: 0.008891893297280433
Confidence: 1.0
Lift: 41.35698924731183
====================================
Regla: SAN ISIDRO DEL INCA -> 1687
Support: 0.008891893297280433
Confidence: 1.0
Lift: 8.741363636363637
====================================
Regla: HORMONAS -> 2
Support: 0.12843845873849513
Confidence: 0.9991909385113269
Lift: 7.77953074433657
====================================
Regla: 2012 -> PONCEANO
Support: 0.006239925120898549
Confidence: 1.0
Lift: 11.00801373783629
====================================
Regla: 2013 -> PONCEANO
Support: 0.007903905153138163
```

```
Confidence: 1.0
Lift: 11.00801373783629
====================================
Regla: 68 -> 218
Support: 0.007279912641048308
Confidence: 1.0
Lift: 65.41156462585033
====================================
Regla: COMITÉ DEL PUEBLO -> 218
Support: 0.007279912641048308
Confidence: 1.0
Lift: 25.641333333333332
====================================
Regla: CALDERÓN -> 2405
Support: 0.0066559201289584525
Confidence: 1.0
Lift: 38.85050505050505
====================================
Regla: 36 -> 2563
Support: 0.005927928864853622
Confidence: 0.6000000000000001
Lift: 27.342654028436023
====================================
Regla: IÑAQUITO -> 2563
Support: 0.009879881441422702
Confidence: 1.0
Lift: 36.63047619047619
====================================
Regla: 3226 -> 29
Support: 0.011231865217617388
Confidence: 0.4114285714285714
Lift: 19.98025974025974
====================================
Regla: SAN ISIDRO DEL INCA -> 29
Support: 0.013987832146014248
Confidence: 0.5123809523809524
Lift: 4.478908225108225
====================================
Regla: 3 -> HEMATOLOGIA CLINICA
Support: 0.07940304716343403
Confidence: 1.0
Lift: 12.593975114603799
====================================
Regla: 3226 -> 31
Support: 0.005615932608808694
Confidence: 0.2727272727272727
Lift: 7.1358070500927635
====================================
```

```
Regla: EXAMENES EXTERNOS -> 31
Support: 0.010035879569445167
Confidence: 0.26258503401360545
Lift: 26.164625850340133
====================================
Regla: CARCELÉN -> 32
Support: 0.006551921376943477
Confidence: 0.302158273381295
Lift: 5.435739715056767
====================================
Regla: 3226 -> SAN ISIDRO DEL INCA
Support: 0.020591752898965213
Confidence: 1.0
Lift: 8.741363636363637
====================================
Regla: 36 -> IÑAQUITO
Support: 0.005927928864853622
Confidence: 0.27014218009478674
Lift: 9.895436696005417
====================================
Regla: SAN ISIDRO DEL INCA -> 39
Support: 0.013467838385939368
Confidence: 0.556989247311828
Lift: 4.868845552297166
====================================
Regla: SEROLOGIA -> 4
Support: 0.03260360875669492
Confidence: 0.9968203497615263
Lift: 30.573926868044516
====================================
Regla: 42 -> 5372
Support: 0.007279912641048308
Confidence: 0.3218390804597701
Lift: 29.472796934865897
====================================
Regla: BELISARIO QUEVEDO -> 42
Support: 0.007279912641048308
Confidence: 0.3218390804597701
Lift: 15.748822789623
====================================
Regla: 45 -> COMITÉ DEL PUEBLO
Support: 0.007383911393063283
Confidence: 0.3183856502242152
Lift: 8.163832585949176
====================================
Regla: 47 -> 9713
Support: 0.0066559201289584525
Confidence: 0.2277580071174377
```

```
Lift: 34.21886120996441
====================================
Regla: 981 -> 47
Support: 0.009359887681347824
Confidence: 0.3202846975088968
Lift: 14.665226232841892
====================================
Regla: 47 -> CARAPUNGO
Support: 0.0066559201289584525
Confidence: 0.2277580071174377
Lift: 11.202082442136687
====================================
Regla: 981 -> 48
Support: 0.006239925120898549
Confidence: 0.38961038961038963
Lift: 17.83951762523191
====================================
Regla: URIANALISIS -> 5
Support: 0.06936716759398887
Confidence: 1.0
Lift: 14.416041979010496
====================================
Regla: 51 -> CONDADO
Support: 0.006031927616868597
Confidence: 0.21928166351606804
Lift: 8.140937588952712
====================================
Regla: 5251 -> JIPIJAPA
Support: 0.006759918880973428
Confidence: 1.0
Lift: 90.28638497652582
====================================
Regla: BELISARIO QUEVEDO -> 5372
Support: 0.010919868961572462
Confidence: 1.0
Lift: 48.93384223918575
====================================
Regla: CARCELÉN -> 5658
Support: 0.008111902657168114
Confidence: 1.0
Lift: 17.989710009354535
====================================
Regla: 6 -> COPROANALISIS
Support: 0.06094326868077583
Confidence: 0.979933110367893
Lift: 16.07943143812709
====================================
Regla: 7684 -> 60
```

```
Support: 0.006551921376943477
Confidence: 0.4131147540983607
Lift: 36.78060109289618
====================================
Regla: 6250 -> SAN ISIDRO DEL INCA
Support: 0.007799906401123186
Confidence: 1.0
Lift: 8.741363636363637
====================================
Regla: 66 -> 7629
Support: 0.009359887681347824
Confidence: 0.5901639344262296
Lift: 63.05245901639345
====================================
Regla: 66 -> RUMIPAMBA
Support: 0.009359887681347824
Confidence: 0.5901639344262296
Lift: 43.6517023959647
====================================
Regla: COMITÉ DEL PUEBLO -> 68
Support: 0.0073319120170557955
Confidence: 0.47959183673469385
Lift: 12.297374149659863
====================================
Regla: RUMIPAMBA -> 7629
Support: 0.009359887681347824
Confidence: 1.0
Lift: 73.96538461538462
====================================
Regla: 77 -> 8919
Support: 0.008527897665228018
Confidence: 0.5173501577287066
Lift: 34.666065795403334
====================================
Regla: 77 -> CHILLOGALLO
Support: 0.008527897665228018
Confidence: 0.5173501577287066
Lift: 31.385365562399862
====================================
Regla: 8 -> MICROBIOLOGIA
Support: 0.05647132234413187
Confidence: 1.0
Lift: 17.708103130755063
====================================
Regla: POMASQUI -> 8200
Support: 0.007487910145078259
Confidence: 1.0
Lift: 22.053899082568808
```

```
====================================
Regla: 83 -> PONCEANO
Support: 0.012583848993812074
Confidence: 0.8402777777777778
Lift: 9.249789321376328
====================================
Regla: PONCEANO -> 85
Support: 0.009151890177317872
Confidence: 0.55
Lift: 6.05440755580996
====================================
Regla: PONCEANO -> 86
Support: 0.007071915137018356
Confidence: 0.6445497630331753
Lift: 7.09521264618832
====================================
Regla: CHILLOGALLO -> 8919
Support: 0.01492382091414903
Confidence: 1.0
Lift: 60.66561514195583
====================================
Regla: 9426 -> PONCEANO
Support: 0.005615932608808694
Confidence: 1.0
Lift: 11.00801373783629
====================================
Regla: CARAPUNGO -> 9713
Support: 0.0066559201289584525
Confidence: 1.0
Lift: 49.18414322250639
====================================
Regla: 1 -> 83
Support: 0.00571993136082367
Confidence: 0.3819444444444444
Lift: 26.232762896825395
====================================
Regla: 1 -> 128
Support: 0.014559825282096615
Confidence: 0.45454545454545453
Lift: 12.346558808423216
====================================
Regla: 1 -> 2256
Support: 0.006239925120898549
Confidence: 0.6
Lift: 4.609908110267678
====================================
Regla: 3226 -> SAN ISIDRO DEL INCA
Support: 0.0068639176329884045
```

```
Confidence: 0.33333333333333337
Lift: 7.613222486144102
====================================
Regla: 1 -> 7684
Support: 0.009359887681347824
Confidence: 0.8333333333333334
Lift: 6.402650153149554
====================================
Regla: 1 -> 83
Support: 0.00571993136082367
Confidence: 0.3819444444444444
Lift: 10.374538998744507
====================================
Regla: 1 -> CHILLOGALLO
Support: 0.005823930112838646
Confidence: 0.3902439024390244
Lift: 58.17659292871999
====================================
Regla: MARCADORES TUMORALES -> KENNEDY
Support: 0.007851905777130675
Confidence: 0.361244019138756
Lift: 46.007177033492816
====================================
Regla: INFECCIOSAS -> 11
Support: 0.01570381155426135
Confidence: 0.26468010517090274
Lift: 16.854513584574935
====================================
Regla: INFECCIOSAS -> SAN ISIDRO DEL INCA
Support: 0.009255888929332849
Confidence: 1.0
Lift: 16.928697183098592
====================================
Regla: 8 -> 128
Support: 0.005823930112838646
Confidence: 1.0
Lift: 17.708103130755063
====================================
Regla: 8 -> PONCEANO
Support: 0.005823930112838646
Confidence: 1.0
Lift: 11.00801373783629
====================================
Regla: 83 -> 128
Support: 0.012583848993812074
Confidence: 0.39285714285714285
Lift: 31.219155844155843
====================================
```

```
Regla: QUIMICA CLINICA -> 83
Support: 0.00571993136082367
Confidence: 0.3819444444444444
Lift: 26.232762896825395
====================================
Regla: PONCEANO -> 85
Support: 0.009151890177317872
Confidence: 0.2857142857142857
Lift: 31.219155844155846
====================================
Regla: PONCEANO -> 128
Support: 0.0068639176329884045
Confidence: 0.2142857142857143
Lift: 30.300945378151262
====================================
Regla: MICROBIOLOGIA -> PONCEANO
Support: 0.005823930112838646
Confidence: 1.0
Lift: 11.00801373783629
====================================
Regla: QUIMICA CLINICA -> PONCEANO
Support: 0.014559825282096615
Confidence: 0.45454545454545453
Lift: 12.346558808423216
====================================
Regla: 13 -> KENNEDY
Support: 0.01221985336175966
Confidence: 0.39036544850498345
Lift: 31.945182724252494
====================================
Regla: KENNEDY -> PRUEBAS ESPECIALES
Support: 0.006915917008995892
Confidence: 1.0
Lift: 27.670503597122302
====================================
Regla: SAN ISIDRO DEL INCA -> 1687
Support: 0.008891893297280433
Confidence: 1.0
Lift: 74.25096525096525
====================================
Regla: HORMONAS -> 47
Support: 0.0061879257448910615
Confidence: 1.0
Lift: 7.78582995951417
====================================
Regla: 981 -> HORMONAS
Support: 0.006551921376943477
Confidence: 1.0
```

```
Lift: 7.78582995951417
====================================
Regla: 981 -> KENNEDY
Support: 0.006551921376943477
Confidence: 0.3
Lift: 6.623765786452354
====================================
Regla: HORMONAS -> CARCELÉN
Support: 0.006759918880973428
Confidence: 1.0
Lift: 7.78582995951417
====================================
Regla: HORMONAS -> COTOCOLLAO
Support: 0.006759918880973428
Confidence: 1.0
Lift: 7.78582995951417
====================================
Regla: HORMONAS -> KENNEDY
Support: 0.045187457750507
Confidence: 0.351537216828479
Lift: 7.7795307443365695
====================================
Regla: PONCEANO -> HORMONAS
Support: 0.009879881441422702
Confidence: 1.0
Lift: 7.78582995951417
====================================
Regla: SAN ISIDRO DEL INCA -> HORMONAS
Support: 0.016379803442358693
Confidence: 1.0
Lift: 7.78582995951417
====================================
Regla: 68 -> COMITÉ DEL PUEBLO
Support: 0.007279912641048308
Confidence: 1.0
Lift: 136.39007092198582
====================================
Regla: QUIMICA CLINICA -> 2256
Support: 0.006239925120898549
Confidence: 0.6
Lift: 4.611750599520384
====================================
Regla: 36 -> 2563
Support: 0.005927928864853622
Confidence: 0.6000000000000001
Lift: 101.21578947368423
====================================
Regla: 3226 -> SAN ISIDRO DEL INCA
```

```
Support: 0.011231865217617388
Confidence: 0.4114285714285714
Lift: 19.98025974025974
====================================
Regla: 3 -> KENNEDY
Support: 0.02511569861161666
Confidence: 0.3163064833005894
Lift: 12.5939751146038
====================================
Regla: 3 -> PONCEANO
Support: 0.008839893921272945
Confidence: 1.0
Lift: 12.593975114603799
====================================
Regla: 3 -> SAN ISIDRO DEL INCA
Support: 0.009567885185377776
Confidence: 1.0
Lift: 12.593975114603799
====================================
Regla: 3226 -> SAN ISIDRO DEL INCA
Support: 0.005615932608808694
Confidence: 0.2727272727272727
Lift: 22.223805855161785
====================================
Regla: 3226 -> QUIMICA CLINICA
Support: 0.0068639176329884045
Confidence: 0.33333333333333337
Lift: 7.613222486144102
====================================
Regla: SEROLOGIA -> 4
Support: 0.00795590452914565
Confidence: 0.24324324324324328
Lift: 30.57392686804452
====================================
Regla: BELISARIO QUEVEDO -> 42
Support: 0.007279912641048308
Confidence: 0.3218390804597701
Lift: 29.472796934865897
====================================
Regla: CARAPUNGO -> 47
Support: 0.0066559201289584525
Confidence: 0.2277580071174377
Lift: 34.21886120996441
====================================
Regla: 981 -> 47
Support: 0.009359887681347824
Confidence: 0.3202846975088968
Lift: 14.665226232841892
```

```
====================================
Regla: 981 -> KENNEDY
Support: 0.006239925120898549
Confidence: 0.38961038961038963
Lift: 17.83951762523191
====================================
Regla: KENNEDY -> URIANALISIS
Support: 0.027091674899901202
Confidence: 0.3905547226386807
Lift: 14.416041979010496
====================================
Regla: SAN ISIDRO DEL INCA -> URIANALISIS
Support: 0.007851905777130675
Confidence: 1.0
Lift: 14.416041979010496
====================================
Regla: KENNEDY -> 6
Support: 0.026467682387811345
Confidence: 0.42558528428093645
Lift: 16.07943143812709
====================================
Regla: SAN ISIDRO DEL INCA -> 6
Support: 0.006395923248921013
Confidence: 1.0
Lift: 16.408703071672353
====================================
Regla: 7684 -> KENNEDY
Support: 0.006551921376943477
Confidence: 0.4131147540983607
Lift: 36.78060109289618
====================================
Regla: 66 -> RUMIPAMBA
Support: 0.009359887681347824
Confidence: 0.5901639344262296
Lift: 63.05245901639345
====================================
Regla: QUIMICA CLINICA -> 7684
Support: 0.009359887681347824
Confidence: 0.8333333333333334
Lift: 6.4052091660005335
====================================
Regla: 77 -> CHILLOGALLO
Support: 0.008527897665228018
Confidence: 0.5173501577287066
Lift: 34.666065795403334
====================================
Regla: 8 -> KENNEDY
Support: 0.027195673651916177
```

```
Confidence: 0.48158379373848986
Lift: 17.708103130755063
====================================
Regla: 8 -> PONCEANO
Support: 0.008111902657168114
Confidence: 1.0
Lift: 17.708103130755063
====================================
Regla: 8 -> SAN ISIDRO DEL INCA
Support: 0.005927928864853622
Confidence: 1.0
Lift: 17.708103130755063
====================================
Regla: QUIMICA CLINICA -> 83
Support: 0.00571993136082367
Confidence: 0.3819444444444444
Lift: 10.374538998744507
====================================
Regla: QUIMICA CLINICA -> CHILLOGALLO
Support: 0.005823930112838646
Confidence: 0.3902439024390244
Lift: 58.17659292871999
====================================
Regla: 981 -> HORMONAS
Support: 0.006551921376943477
Confidence: 0.3
Lift: 6.639010356731875
====================================
Regla: 1 -> 83
Support: 0.00571993136082367
Confidence: 0.3819444444444444
Lift: 26.232762896825395
====================================
Regla: QUIMICA CLINICA -> 1
Support: 0.00571993136082367
Confidence: 0.3819444444444444
Lift: 26.232762896825395
====================================
Regla: QUIMICA CLINICA -> 1
Support: 0.014559825282096615
Confidence: 0.45454545454545453
Lift: 12.346558808423216
====================================
Regla: QUIMICA CLINICA -> 1
Support: 0.006239925120898549
Confidence: 0.6
Lift: 4.611750599520384
====================================
```

```
Regla: 3226 -> QUIMICA CLINICA
Support: 0.0068639176329884045
Confidence: 0.33333333333333337
Lift: 7.613222486144102
====================================
Regla: QUIMICA CLINICA -> 1
Support: 0.005927928864853622
Confidence: 0.6514285714285715
Lift: 5.007043508050703
====================================
Regla: QUIMICA CLINICA -> 1
Support: 0.009359887681347824
Confidence: 0.8333333333333334
Lift: 6.4052091660005335
====================================
Regla: QUIMICA CLINICA -> 1
Support: 0.00571993136082367
Confidence: 0.3819444444444444
Lift: 10.374538998744507
====================================
Regla: QUIMICA CLINICA -> 1
Support: 0.005823930112838646
Confidence: 0.3902439024390244
Lift: 58.17659292871999
====================================
Regla: 8 -> PONCEANO
Support: 0.005823930112838646
Confidence: 1.0
Lift: 123.27564102564102
====================================
Regla: QUIMICA CLINICA -> 83
Support: 0.00571993136082367
Confidence: 0.3819444444444444
Lift: 26.232762896825395
====================================
Regla: 981 -> HORMONAS
Support: 0.006551921376943477
Confidence: 0.3
Lift: 6.639010356731875
====================================
Regla: 83 -> PONCEANO
Support: 0.00571993136082367
Confidence: 0.3819444444444444
Lift: 26.232762896825395
====================================
```

[24]:
```
lift = []
association = []
for i in range (0, len(rslt)):
    lift.append(rslt[:len(rslt)][i][2][0][3])
    association.append(list(rslt[:len(rslt)][i][0]))
```

[25]:
```
rank = pd.DataFrame([association, lift]).T
rank.columns = ['Association', 'Lift']
```

[26]:
```
# Show top 10 higher lift scores
rank.sort_values('Lift', ascending=False).head(100)
```

[26]:

	Association	Lift
751	[3, 365, VELOCIDAD DE ER]	173.252252
1344	[3, 365, VELOCIDAD DE ER, HEMATOLOGIA CLINICA]	173.252252
822	[365, VELOCIDAD DE ER, HEMATOLOGIA CLINICA]	173.252252
134	[365, VELOCIDAD DE ER]	173.252252
51	[CULTIVO Y ANTIB, 195]	152.626984
…	…	…
807	[VDRL-RPR, 4, 315]	71.757463
1378	[VDRL-RPR, SEROLOGIA, 4, 315]	71.757463
809	[HORMONAS, HCG CUALITATIVA, 319]	68.928315
505	[HCG CUALITATIVA, 319, 2]	68.928315
1198	[KENNEDY, HCG CUALITATIVA, 319, 2]	68.928315

[100 rows x 2 columns]

[28]:
```
def inspect(rslt): # function to organize the output
    left_handSide         = [tuple(result[2][0][0])[0] for result in rslt] #␣
 ↪get the left hand side of the rules
    right_handSide        = [tuple(result[2][0][1])[0] for result in rslt] #␣
 ↪get the right hand side of the rules
    supports    = [result[1] for result in rslt] # get the supports
    return list(zip(left_handSide,right_handSide, supports)) # zip the above␣
 ↪three lists together
rslt_DataFrame = pd.DataFrame(inspect(rslt), columns = ['Product 1', 'Product␣
 ↪2', 'Support']) # create a pandas dataframe
```

[30]:
```
rslt_DataFrame.nlargest(n = 100, columns = 'Support') # printing the first 7␣
 ↪supports
```

[30]:

	Product 1	Product 2	Support
59	2	HORMONAS	0.128438
100	3	HEMATOLOGIA CLINICA	0.079403
151	5	URIANALISIS	0.069367
71	203	5	0.066559
72	203	ELEMENTAL Y MIC	0.066559

```
…                 …                       …        …
725                 203                   KENNEDY  0.026572
864   ELEMENTAL Y MIC                     KENNEDY  0.026572
916   ELEMENTAL Y MIC                     KENNEDY  0.026572
1328                203          ELEMENTAL Y MIC  0.026572
1331                203                   KENNEDY  0.026572

[100 rows x 3 columns]
```

2 Implementación de ECLAT

[33]:
```
# Putting all transactions in a single list
itens = []
for i in range(0, len(transactions)):
    itens.extend(transactions[i])

# Finding unique items from transactions and removing nan
uniqueItems = list(set(itens))
#uniqueItems.remove('nan')
```

[34]:
```
pair = []
for j in range(0, len(uniqueItems)):
    k = 1;
    while k <= len(uniqueItems):
        try:
            pair.append([uniqueItems[j], uniqueItems[j+k]])
        except IndexError:
            pass
        k = k + 1;
```

[35]:
```
score = []
for i in pair:
    cond = []
    for item in i:
        cond.append('("%s") in s' %item)
    mycode = ('[s for s in transactions if ' + ' and '.join(cond) + ']')
    #mycode = "print 'hello world'"
    score.append(len(eval(mycode))/19231.)
```

```
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
Cell In[35], line 8
      6 mycode = ('[s for s in transactions if ' + ' and '.join(cond) + ']')
      7 #mycode = "print 'hello world'"
----> 8 score.append(len(eval(mycode))/19231.)
```

```
File <string>:1

File <string>:1, in <listcomp>(.0)

KeyboardInterrupt:
```

[ ]:
```
ranking_ECLAT = pd.DataFrame([pair, score]).T
ranking_ECLAT.columns = ['Pair', 'Score']
```

[ ]:
```
ranking_ECLAT.sort_values('Score', ascending=False).head(100)
```

[ ]:
```

```