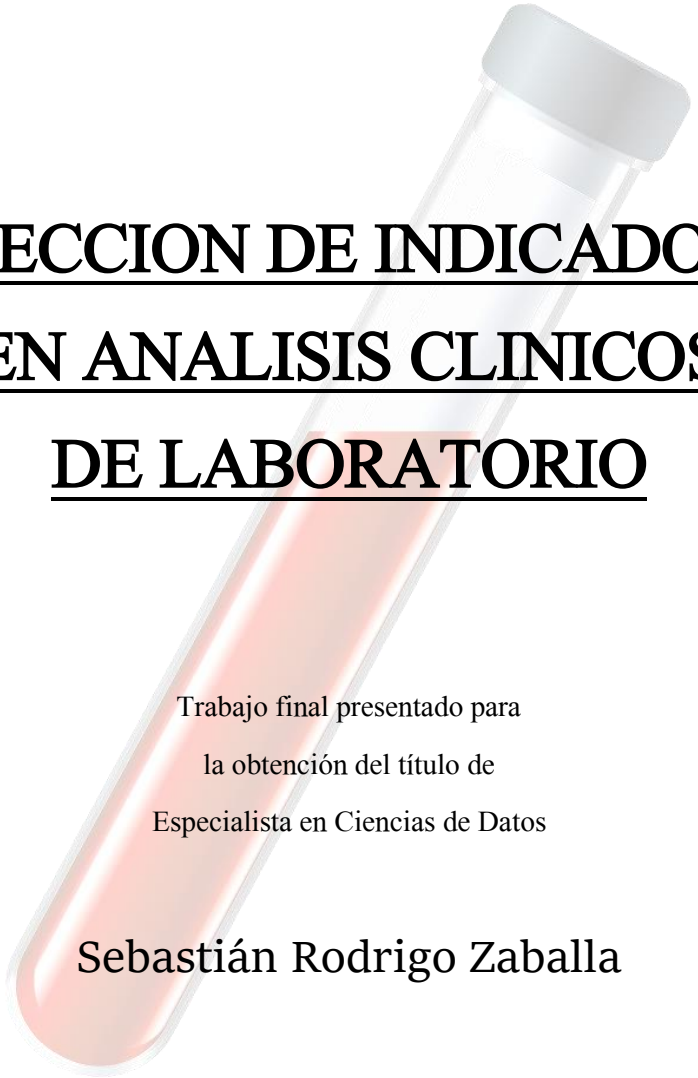
INSTITUTO TECNOLÓGICO DE BUENOS AIRES
Escuela de Postgrado
Especialización en Ciencia de Datos



DETECCION DE INDICADORES EN ANALISIS CLINICOS DE LABORATORIO

Trabajo final presentado para

la obtención del título de

Especialista en Ciencias de Datos

Sebastián Rodrigo Zaballa

Tutor del TFI: Dr. Mario Rossi
Director de carrera: Dr. Alejandro Vaisman
Ciudad Autónoma de Buenos Aires, 2020

Contenido

1. Introducción

Desde hace muchos años el sistema de salud realiza estudios de análisis clínicos de laboratorio, ya sea porque lo solicita un médico de cabecera para un chequeo anual o puntualmente por si se presenta algún problema de salud o si los pacientes se encuentran internados para saber un poco más de lo que está pasando.

La cantidad de todos los análisis que se realiza un ser humano a través de los años ocupan un espacio considerable y más aún en un hospital, donde asisten miles de pacientes por mes.

La cantidad de información generada por la masividad de los datos, resulta difícil de analizar. Tanto el análisis estadístico descriptivo y los algoritmos de clasificación o clusterización ya existen hace mucho tiempo, tanto como los análisis de laboratorio, pero no existían, en un comienzo, las herramientas tecnológicas para aplicarlas o para almacenar los datos y/o resultados. Luego, cuando empezaron a existir eran de un costo imposible de afrontar para cualquier institución pública o privada, sin una contraprestación económica que la justifique.

Actualmente existen diferentes técnicas tanto para el almacenaje de datos como en los algoritmos de programación y la capacidad de procesamiento es posible analizar todos los análisis de laboratorio juntos, permitiéndonos encontrar patrones y generar alertas para los profesionales de la salud y por ende mejorar la calidad de atención de los pacientes. También nos abre una ventana a la detección temprana de posibles problemas de salud, tomando como base toda la información disponible y en forma automática.

2. Antecedentes/Estado del Arte/Marco Teórico/Conceptual

Los análisis de laboratorio son una parte fundamental en el cuidado de la salud y la seguridad del paciente.

Estos análisis son una herramienta clave de ayuda para la decisión clínica, ya sea para un diagnóstico o una acción inmediata.

Un buen diagnóstico médico depende de un análisis clínico con calidad, las pruebas diagnósticas de laboratorio vienen cobrando una marcada relevancia, a la hora de establecer un diagnóstico certero. Así, uno de los actores del Sistema de Salud es el laboratorio clínico, que es una herramienta primordial para el área médica, ya que por medio de este se diagnostican diferentes patologías y además se realizan estudios para establecer el tipo de tratamiento que se debe administrar al paciente, al igual que el seguimiento del mismo. [1]

La información que aporta el laboratorio al clínico es de gran importancia. En un elevado porcentaje de casos la decisión tomada por el médico clínico respecto a la actuación sobre el paciente está basada en esta información. Por este motivo, la calidad de los resultados del informe del laboratorio clínico es esencial. Todo el proceso debe estar controlado, desde la

solicitud de las determinaciones hasta la interpretación de los resultados, ya que cualquier error podría potencialmente tender consecuencias negativas sobre los pacientes. [2]

Un laboratorio clínico debe establecer reglas de aceptación o rechazo de los ítems solicitados por los profesionales de la salud en base a lo que este esté buscado, así como el profesional de la salud debe saber que pedir en el análisis de laboratorio en base a los síntomas, diagnósticos anteriores o actuales que el paciente manifieste.

Para disminuir los efectos, se elaboró entonces en el 2010 una guía denominada "Guía para garantizar la correcta identificación del paciente y las muestras de laboratorio" que explica todo lo referente a la toma y el marcaje de una muestra aplicado a diferentes profesionales de la salud involucrados en este proceso.

En otros casos los errores pueden no tener repercusiones sobre el paciente, pero si conllevan a repeticiones innecesarias de mediciones y exámenes in vitro, dando lugar a un aumento del costo y trato inadecuado del paciente. En la situación actual la optimización de los recursos, tanto humanos como económicos, es esencial. [2]

3. Definición del Problema

3.1. Planteamiento del Problema

Los análisis de laboratorio se realizan en forma individual y aislada uno de otro para cada paciente.

Por lo general se realizan esporádicamente, cuando los pacientes son jóvenes y con cierta periodicidad a medida que pasan los años y en forma diaria si el paciente se encuentra internado o si se está realizando algún tratamiento, pueden ser que se realicen una vez por semana.

Los resultados de cada uno son analizados en forma independiente y diagnosticados en la misma forma. Estos resultados pueden indicar un problema de salud manifestado o pronto a manifestarse.

Algunos estudios de investigación demuestran que un 30% de las variables solicitados son innecesarios.

No se toman en cuenta las cantidades de sangre extraídas cuando los pacientes están internados y se les realizan extracciones para su análisis, lo que puede generar para el paciente un perjuicio.

Esto genera para cada hospital una falta global del análisis y de la evolución de los diferentes indicadores que se miden, que pueden derivar en problemas de salud, así como la imposibilidad de la búsqueda de patrones o indicadores que puedan anticiparse al problema, la solicitud de estudios innecesarios la falta de alertas en la Historia Clínica Electrónica (HCE).

4. Justificación del estudio

La gran cantidad de estudios clínicos de laboratorio y las herramientas de análisis de grandes volúmenes de información, junto con el avance de la tecnología que permite estos análisis, produce en el Hospital Italiano de Buenos Aires la necesidad de realizar un estudio de investigación de los mismos.

Con el análisis descriptivo de los estudios, nos permitirá generar una base de conocimiento con los conceptos de los estudios de laboratorio por las diferentes variables de estudio. Esto brindara múltiples fuentes de análisis para realizar otros estudios de investigación.

Fundamentalmente, este estudio, beneficia principalmente la salud del paciente. Dado que el estudio busca resolver los problemas planteados mediante análisis estadísticos descriptivos y algoritmos de inteligencia artificial.

5. Alcances del trabajo y limitaciones

5.1. Alcance

Este estudio está orientado a investigadores, médicos y pacientes. A los investigadores, porque se dispondrá de una base de información estructurada, limpia y confiable para su análisis. Para los médicos porque les bridaran alertas basadas en patrones de la información. Y para los pacientes, porque ellos son los principales beneficiarios del estudio.

Los resultados de este estudio quedaran disponibles para el área de investigación y para el área de ingeniería de software del Hospital Italiano de Buenos Aires para desarrollo de aplicaciones o funcionalidades nuevas a aplicaciones existentes.

El alcance del estudio está dado por el análisis de 10 años (01/01/2010 al 31/12/2019) de estudios de laboratorio de pacientes internados y cuyo financiador sea el mismo Hospital Italiano de Buenos Aires (prepaga Plan de Salud) y por los problemas detectados a los pacientes en el mismo lapso de tiempo.

No se tendrá en cuenta en este estudio, las cirugías que hayan tenido los pacientes, otros estudios de los diferentes servicios del hospital, como ser radiografías (Diagnóstico por Imagen), video colonoscopías (Gastroenterología), biopsias (Anatomía Patológica), etc.

5.2. Limitaciones

Para el estudio de investigación se dispone en el equipo de un profesional de la salud (médico), como soporte de consulta.

Como se mencionó en el punto anterior se tomarán solo 10 años (01/01/2010 al 31/12/2019) de análisis clínicos de laboratorio y por un presupuesto limitado para el procesamiento de dicha información, se tomarán solamente las variables más representativas, que presenten más de 80% de ocurrencia dentro de la muestra tomada.

6. Hipótesis

Si tomamos 10 años de historia de análisis clínicos de laboratorio junto con los diagnósticos de los pacientes:

- Es posible detectar patrones en las variables que permitan diagnosticar o guiar al profesional en forma automática.
- Se puede reducir el error en solicitudes en un 10% en un comienzo, hasta alcanzar solamente un 5% de error en ellas.
- Para cada ítem de un análisis de laboratorio se puede realizar un análisis estadístico descriptivo a fin de poder determinar limites, valores atípicos, errores, según diferentes variables como ser sexo y edad.

6.1. Variables

- Sexo
  Genero de nacimiento del paciente.

- Edad
  Cantidad de años que tiene el paciente al momento de realizar el análisis de laboratorio o se le manifiesta un problema registrado en su historia clínica.

- Series de Tiempo
  Para la medicina, el tiempo es un factor muy importante en la investigación. Por ej. el año se divide en semanas epidemiológicas, así también como son las estaciones del año y los problemas que pueden producirse en cada uno.

- Variables de Sangre
  Como se plantea en la hipótesis, si de los 10 años de historia, tomamos los pacientes internados en la institución y que correspondan a la obra social de dicha institución, a fin de tener una misma población de muestra a lo largo de los 10 años. Y de ellos seleccionamos los análisis de laboratorio y variables que cumplen con un porcentaje mayor o igual al 80% de la población, analizaremos 34 variables, las cuales se describen a continuación:

    1. GLUCOSA – Bioquímica básica
       Mide la cantidad (concentración) de glucosa presente en la sangre.

La glucosa es un azúcar que es utilizado por los tejidos como forma de energía al combinarlo con el oxígeno de la respiración. Cuando comemos el azúcar en la sangre se eleva, lo que se consume desaparece de la sangre, para ello hay una hormona reguladora que es la insulina producida por el páncreas (islotes pancreáticos). Esta hormona hace que la glucosa de la sangre entre en los tejidos y sea utilizada en forma de glucógeno, aminoácidos, y ácidos grasos. Cuando la glucosa en sangre está muy baja, en condiciones normales por el ayuno, se secreta otra hormona llamada glucagón que hace lo contrario y mantiene los niveles de glucosa en sangre.

El tejido más sensible a los cambios de la glucemia es el cerebro, en concentraciones muy bajas o muy altas aparecen síntomas de confusión mental e inconsciencia.

2. HEMATIES RECUENTO – Serie roja

Los hematíes, también llamados eritrocitos o simplemente glóbulos rojos son las células sanguíneas encargadas de llevar el oxígeno a las células y los tejidos.

El recuento de hematíes en sangre es adecuado para conocer el estado general de salud, la existencia de una anemia, de enfermedades generales o diferentes tipos de cáncer.

3. HEMATOCRITO – Serie roja

Tras una centrifugación de la sangre total se pueden apreciar dos niveles, uno con el depósito de los glóbulos rojos, principalmente, y otro nivel del plasma total. La relación porcentual entre ambos es lo que describe el **hematocrito** y describe el porcentaje de células transportadoras de oxígeno con respecto al volumen total de sangre.

4. HEMOGLOBINA SANGRE TOTAL – Serie roja

La **hemoglobina** es una proteína que contiene hierro y que le otorga el color rojo a la sangre. Se encuentra en los glóbulos rojos y es la encargada del transporte de oxígeno por la sangre desde los pulmones a los tejidos.

La hemoglobina también transporta el dióxido de carbono, que es el producto de desecho del proceso de producción de energía, lo lleva a los pulmones desde donde es exhalado al aire.

5. LEUCOCITOS RECUENTO – Serie blanca

Los **leucocitos** o **glóbulos blancos** son células que están principalmente en la sangre y circulan por ella con la función de combatir las infecciones o cuerpos extraños; pero en ocasiones pueden atacar los tejidos normales del propio cuerpo. Es una parte de las defensas inmunitarias del cuerpo humano.

La modificación de la cantidad de leucocitos puede orientar al diagnóstico de enfermedades infecciosas, inflamatorias, cáncer y leucemias, y otros procesos. Por ello el recuento es muy orientativo en diferentes enfermedades.

Además, el porcentaje de cada grupo de leucocitos nos ofrecerá una mayor información para precisar un diagnóstico.

6. NEUTRÓFILOS MIELOCITOS – Serie blanca

Los neutrófilos son uno de los tipos más numerosos de glóbulos blancos (o leucocitos) que participan en la respuesta del sistema inmunitario fagocitando los gérmenes extraños.

Son, por tanto, células responsables de atacar a los antígenos agresores (bacterias, virus, hongos o células tumorales) siendo el primer tipo de célula que responde a la infección por lo que se consideran la primera línea de defensa.

Existen dos tipos de neutrófilos:

Segmentados: Son neutrófilos maduros. Son los más numerosos en el torrente sanguíneo y se trasladan a los tejidos para combatir a los gérmenes.

Cayados o en banda: Son neutrófilos inmaduros de reserva que se encuentran en la médula ósea. Suelen aumentar su presencia en la sangre en respuesta a una infección bacteriana.

7. NEUTRÓFILOS SEGMENTADOS – Serie blanca

Idem anterior, parte del mismo origen.

8. VOLUMEN CORPUSCULAR MEDIO (VCM) – Serie roja

Permite discernir entre diferentes tipos de anemia.

9. HEMOGLOBINA CORPUSCULAR MEDIA (HCM) – Serie roja

La hemoglobina es una proteína que contienen los glóbulos rojos (también llamados eritrocitos o hematíes) encargada de transportar el oxígeno desde los pulmones a los tejidos y recoger a su vez el dióxido de carbono sobrante para expulsarlo.

La HCM se reduce cuando hay problemas en la síntesis de hemoglobina. Es un parámetro que cobra sentido cuando la hemoglobina se encuentra por debajo de lo normal, es decir, cuando existe anemia.

10. CONCENTRACIÓN DE HCM (CHCM) – Serie roja

La hemoglobina es la proteína presente en los glóbulos rojos encargada de suministrar oxígeno a las células mientras que el hematocrito es el porcentaje de glóbulos rojos con respecto al volumen total de la sangre.

El CHCM es por tanto la masa de hemoglobina presente en un volumen concreto de glóbulos rojos.

Es un parámetro con una utilidad diagnóstica limitada ya que presenta poca variación, pero puede ser útil para detectar determinadas anemias.

11. RDW – Serie roja

El parámetro **RDW** expresa la variación o dispersión del tamaño de los glóbulos rojos en la sangre.

La RDW de forma aislada no tiene utilidad. Sirve de ayuda cuando se padece anemia (hemoglobina en sangre está por debajo de lo normal) porque se eleva solamente en ciertos tipos de anemias.

12. NEUTRÓFILOS METAMIELOCITOS – Serie blanca

    Idem anterior, parte de la variable nro. 17

13. NEUTRÓFILOS EN CAYADO – Serie blanca

    Idem anterior, parte de la variable nro. 17

14. BASÓFILOS – Serie blanca

    Los **basófilos** son un tipo de glóbulos blancos (=leucocitos) que forman parte del sistema inmunitario liberando enzimas que ayudan al cuerpo a protegerse frente a parásitos e invasores externos.

15. EOSINÓFILOS – Serie blanca

    Los eosinófilos son un tipo de glóbulos blancos (=leucocitos) que participan en la respuesta inmunitaria del organismo principalmente frente a los parásitos.

    Son células mieloides que derivan de las mismas células precursoras que los neutrófilos, basófilos y monocitos y se crean en la médula ósea.

16. LINFOCITOS – Serie blanca

    Los linfocitos son un tipo de glóbulos blancos (también llamados leucocitos) cuya principal misión es luchar contra las infecciones provocadas por bacterias y virus. Todos los tipos de linfocitos trabajan conjuntamente para combatir las infecciones.

17. MONOCITOS – Serie blanca

    Los monocitos son un tipo de glóbulos blancos (=leucocitos) que tienen la función de “comer” o eliminar las sustancias extrañas que pueden atacar a nuestro organismo (virus, bacterias, hongos).

    Los monocitos son fundamentales en la defensa del organismo ya que son las células precursoras de los macrófagos que se trasladan a los tejidos para defender al organismo de los cuerpos extraños mediante la fagocitación.

18. CÉLULAS DE DOWNEY

    Corresponden a los LINFOCITOS Tipo T. Variable nro. 16.

La utilidad de estas variables, fueron extraídas de un portal web denominado “tuotromedico”.[3]

7. Objetivos

7.1. Objetivo General

Elaborar una base de datos con los estudios de laboratorio realizados en el Hospital Italiano de Buenos Aires, parametrizada y libre de errores.

7.2. Objetivos Específicos

- Se puede reducir el error en solicitudes en un 10% en un comienzo, hasta alcanzar solamente un 5% de error en ellas.
- Realizar un análisis descriptivo de los diferentes ítems que se solicitan en un análisis de laboratorio.
- Generar patrones entre resultados de laboratorio y diagnósticos, basado en modelos de clusterización y componentes principales.
- Generar alertas en la HCE basados en los resultados de los análisis de laboratorio y diagnósticos, basados en modelos de clasificación.

8. Metodología

8.1. Técnicas

Primero, se generará una interface para extraer los datos que necesitamos para el análisis. Dándole un pre formato para poder realizar nuestro objetivo, por eso se dejará en un repositorio destino diferente.

Luego se realizará un análisis estadístico descriptivo de cada variable de los valores de laboratorio. Con esos resultados se descartarán los errores detectados y/o se aplicarán métodos de corrección a fin de poder ser utilizados en los algoritmos que planteamos.

Ya con ello tendremos nuestro primer objetivo logrado y con su resultado se armará el dataset necesario para aplicarles los diferentes modelos de aprendizaje.

Se aplicarán modelos de componentes principales y de clusterización para alcanzar los objetivos planteados.

8.2. Herramientas

Como herramientas para el estudio se seleccionaron aquellas que son de código abierto, quiere decir que son de libre acceso y costo.

La muestra se extrajo de a base de datos Oracle del Hospital mediante sentencias de SQL, realizando un análisis de desidentificación de cualquier dato que permita identificar al paciente.

Para el reservorio de información se utilizó el motor de base de datos PostgresSQL, la cual es de muy buena performance y funcionalidades.

El análisis estadístico descriptivo se realizó con el programa R.

Para los algoritmos de componentes principales y de clusterización se utilizaron los ya desarrollados que se encuentran en las bibliotecas para el programa R.

Por último, como herramienta complementaria a R, para generar otros tipos de gráficos estadísticos se utilizó Tableau y Excel, utilizada durante el posgrado bajo licencia educativa.

8.3. Obtención de datos

Los datos recabados para la muestra se iniciaron con las solicitudes al servicio de Laboratorio, que se hayan completados en forma exitosa, donde las variables solicitadas sean numéricas y descartados los registros de prueba.

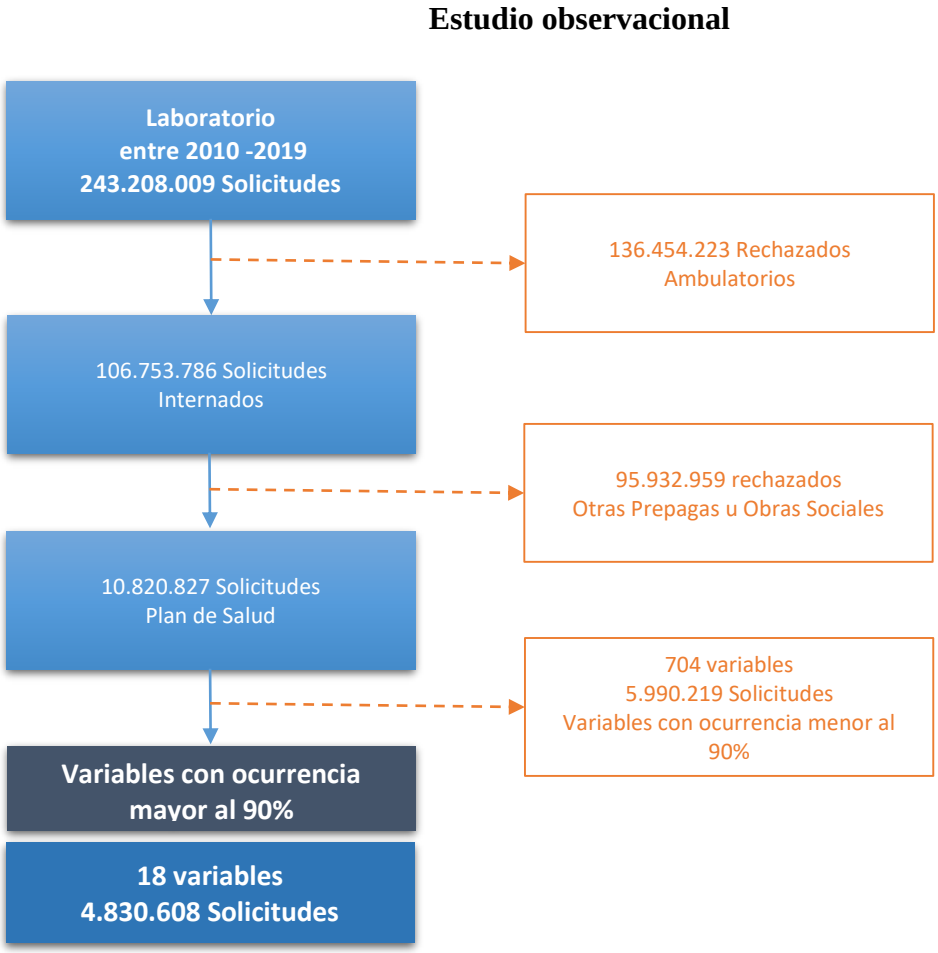


Figura 1 – Proceso de obtención del dataset

De esta forma nos quedamos con 18 variables de laboratorio para realizar el análisis, descripto en el ítem 6.1.

8.4. Data Cleaning

Se realizó el análisis de los datos obtenidos, el cual nos permitió definir la estrategia para el armado final del dataset que nos permita aplicar los diferentes algoritmos planteados.

Primero, ingresamos los datos (tabla ancha) obtenidos del estudio observacional en Postgres, a fin de poder manipularlos libremente.

Mediante consultas de SQL, se dividió la información en 18 tablas, cada una correspondiente a una variable de las identificadas para su análisis.

A cada tabla (variable), se aplicó técnicas de SQL a fin de detectar duplicidad de datos, dando el resultado observado en la Tabla 1.

Variable	Descripción	Total	Duplicados	%
348	GLUCOSA	263.643	280	0,11%
368	HEMATIES RECUENTO	263.474	247	0,09%
369	HEMATOCRITO	307.156	3.538	1,15%
379	HEMOGLOBINA	270.448	886	0,33%
463	LEUCOCITOS RECUENTO	284.713	2.160	0,76%
6082	NEUTROFILOS MIELOCITOS	264.200	343	0,13%
6083	NEUTROFILOS SEGMENTADOS	263.907	341	0,13%
6098	VCM	263.518	246	0,09%
6099	HCM	263.366	245	0,09%
6100	CHCM	263.512	245	0,09%
6101	RDW	263.527	245	0,09%
6520	NEUTROFILOS METAMIELOCITOS	264.197	343	0,13%
6521	NEUTROFILOS EN CAYADO	264.197	343	0,13%
7911	BASOFILOS	263.906	341	0,13%
7912	EOSINOFILOS	263.902	341	0,13%
7913	LINFOCITOS	263.900	341	0,13%
7914	MONOCITOS	263.840	241	0,09%
7915	CELULAS DE DOWNEY	264.133	343	0,13%
	Total	4.819.539	11.069	0,23%

Tabla 1 – Cantidad de solicitudes duplicadas por variable.

8.5. Análisis de las solicitudes

Continuando con el análisis de los datos, se realizó sobre cada variable (tabla) un análisis enfocado a la periodicidad de extracción de la muestra de sangre sobre los pacientes.

Para ello, se continuo con técnicas de SQL, particularmente con las **Windows Función**, dónde para cada internación se calculó el tiempo entre cada extracción mientras estuvo internado y se clasifico ese tiempo en horas. Luego se representan la cantidad de internaciones que se realizaron extracciones en el intervalo, por ej. 10 internaciones se realizaron extracciones entre la primera y la segunda en una hora de diferencia. Lo que buscamos representar y analizar es la cantidad de extracciones que se hacen en un periodo de

tiempo, el cual, según un análisis médico, tiene sentido de realizar y evitar una práctica invasiva al paciente y de costos quizás innecesarios.

En el análisis, se tomaron todas las variables, en un rango de 0 (cero), que representa el periodo de tiempo entre 0,01 y 0,5 decimal de 1 que representa entre 1 (un) minuto y 30 (treinta) minutos, y 75 horas que representa 3 días y 3 horas dado que en ese lapso se concentran la mayor cantidad de internaciones.

También se diferenció si el valor de resultado de la variable es igual al resultado de la extracción anterior o si es diferente, a fin de medir el porcentaje sobre el total de valores que no varían y sobre ellas, analizarlas y definir acciones para cumplir con el objetivo planteado de reducir el error de solicitudes y variables de análisis.

GLUCOSA – En la Figura 2 podemos visualizar, en primero lugar que luego de la primera extracción, durante la primera hora los valores son iguales a la primera extracción en un 9% y luego se mantiene hasta las 75 hs en un promedio del 2%, observando que a las 24 hs, 48 hs y 72 hs se producen los picos de extracciones, donde su pico mayor es a las 24 hs, luego a las 48 y por ultimo a las 72 hs, esto nos indica que durante la internación la frecuencia más importante es cada 24 hs. Observamos que en esos picos, los valores que permanecen iguales van descendiendo, en el primer pico al 7%, en el segundo al 4% y en el último al 3%. Por último, es importante observar que en las primeras 24 hs hay un promedio del 10% de extracciones cada una hora en referencia al total del pico de las 24 hs, en esas observaciones se pueden trabajar con los especialistas de laboratorio a fin de determinar si son necesarias.

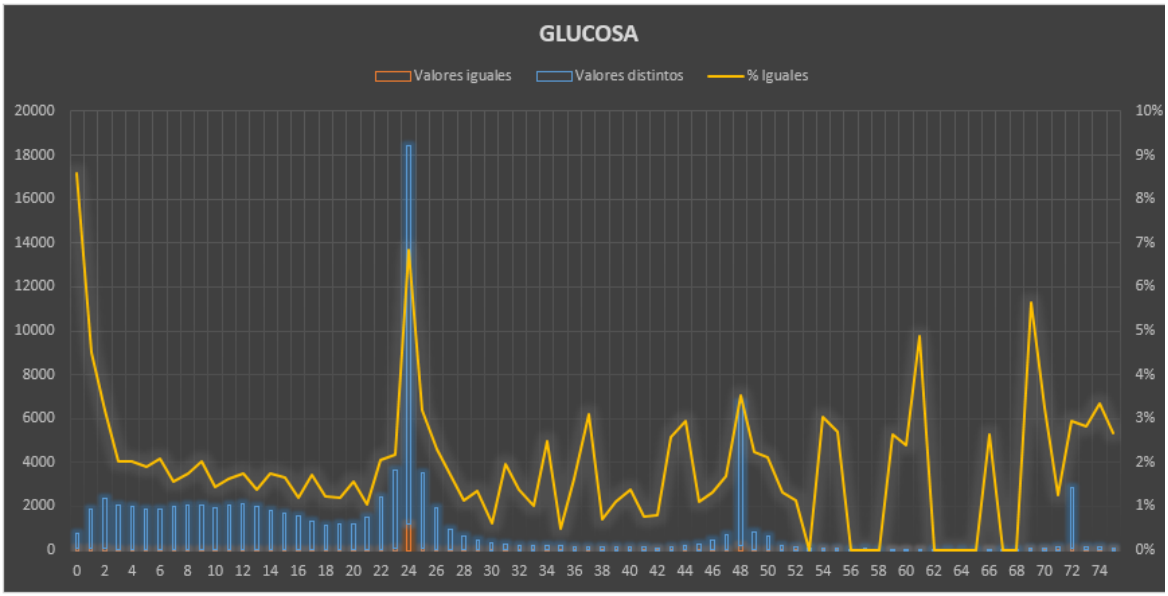


Figura 2 – Cantidad de internaciones por frecuencia de extracción

HEMATIES RECUENTO - En la Figura 3 se visualiza una silueta similar a la anterior y que se repetirá en el resto de las variables. Pero en este caso, la primera extracción, durante la primera hora los valores son iguales a la primera extracción en un 22%, que indicaría que la variable no varía tanto en la primera hora como la anterior y luego se mantiene hasta las 75 hs en un promedio del 2%. observando que a las 24 hs, 48 hs y 72 hs se producen los picos de extracciones y el porcentaje de valores iguales es 5%, 2% y 2% respectivamente, como en la variable anterior y también sus magnitudes. En las primeras 24 hs hay un promedio del 10% de extracciones cada una hora en referencia al total del pico de las 24 hs, igual que la variable anterior, en esas observaciones se pueden trabajar con los especialistas de laboratorio a fin de determinar si son necesarias o se puede reducir las extracciones o ampliar el rango de extracción.

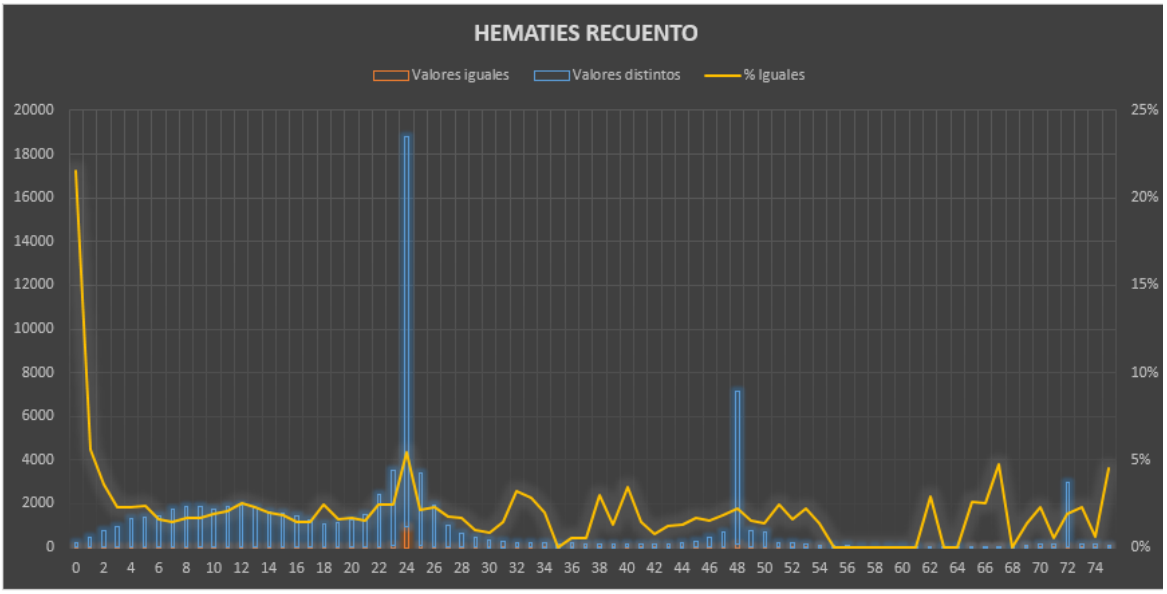


Figura 3 – Cantidad de internaciones por frecuencia de extracción

HEMATOCRITO - En la Figura 4 y visualizando el mismo comportamiento, nos focalizaremos en el porcentaje de muestras sin variación en la primera hora que es del 21%, y a diferencia de las variables anteriores, la caída del porcentaje es más escalonada que las anteriores, siendo del 19% para la segunda hora, 8% para la tercera, 6% para la cuarta y 5% para la quinta. El promedio para esta variable es del 3%. Y en las primeras 24 hs hay un promedio del 12% de extracciones cada una hora en referencia al total del pico de las 24 hs, en las cuales se pueden trabajar con los especialistas de laboratorio a fin de mejorar el proceso y el cuidado del paciente.

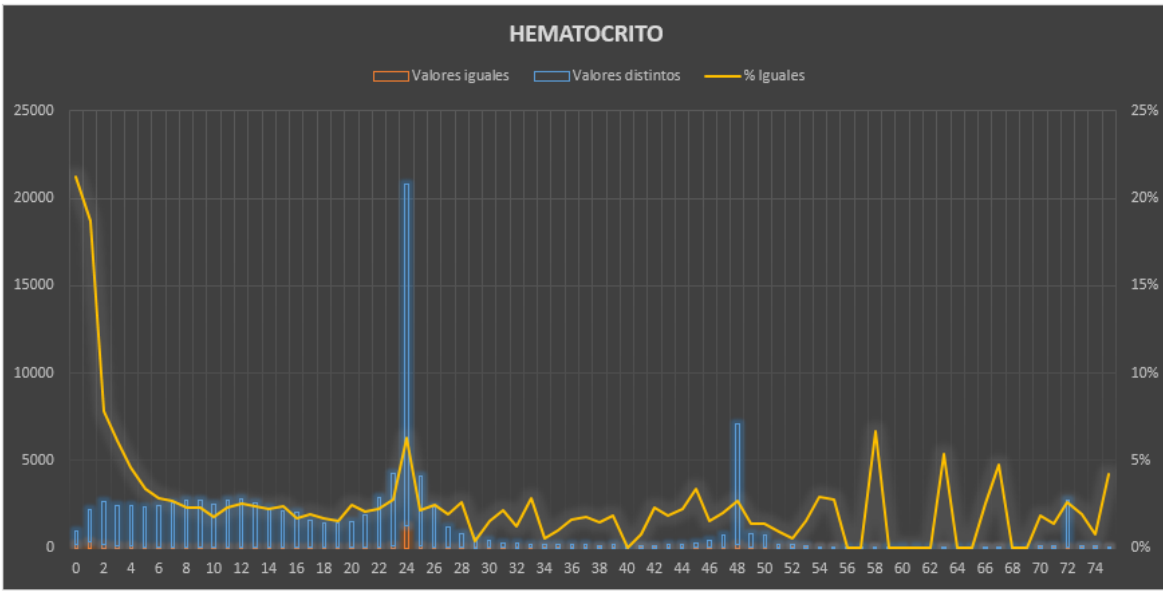


Figura 4 – Cantidad de internaciones por frecuencia de extracción

HEMOGLOBINA – Figura 5, como en las variables anteriores, en la primera hora tienen 29% de observaciones sin variación, hasta ahora el más alto, luego baja a 9% en la segunda hora, pero mantiene un promedio del 6%, mayor a las anteriores. El comportamiento dentro de las primeras 24 hs es similar a la variable "Hematies Recuento" en forma de campana, con un promedio del 9%.

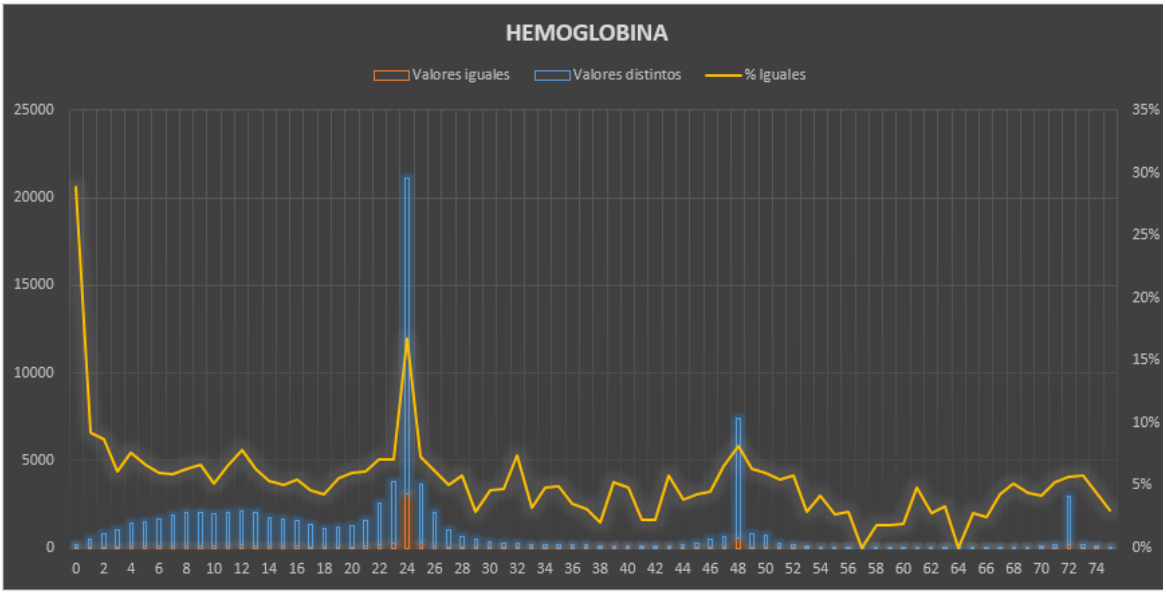


Figura 5 – Cantidad de internaciones por frecuencia de extracción

LEUCOCITOS RECUENTO – Figura 6, en la primera hora las observaciones sin variación son del 24% y luego se desploma a un promedio del 1%. En las primeras 24 hs su forma es acampanada con un promedio del 9% para analizar por los expertos (laboratorio).

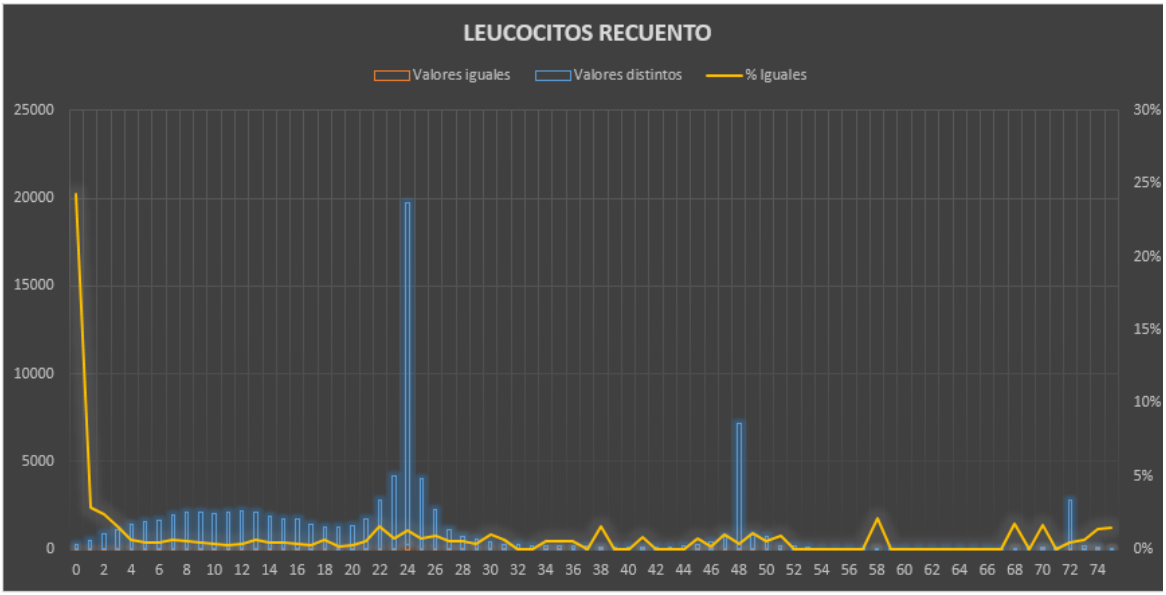


Figura 6 – Cantidad de internaciones por frecuencia de extracción

NEUTROFILOS MIELOCITOS – Figura 7, como se puede observar su comportamiento es totalmente distinto a las otras variables, ya que estas se miden en relación a la participación que tiene la variable, por lo tanto, las cantidades de observaciones que cambian con su medición anterior son muy pocas y por ello se mantienen cerca del 100%, para compararlas con las otras variables, podemos leer la línea de variación a la inversa.

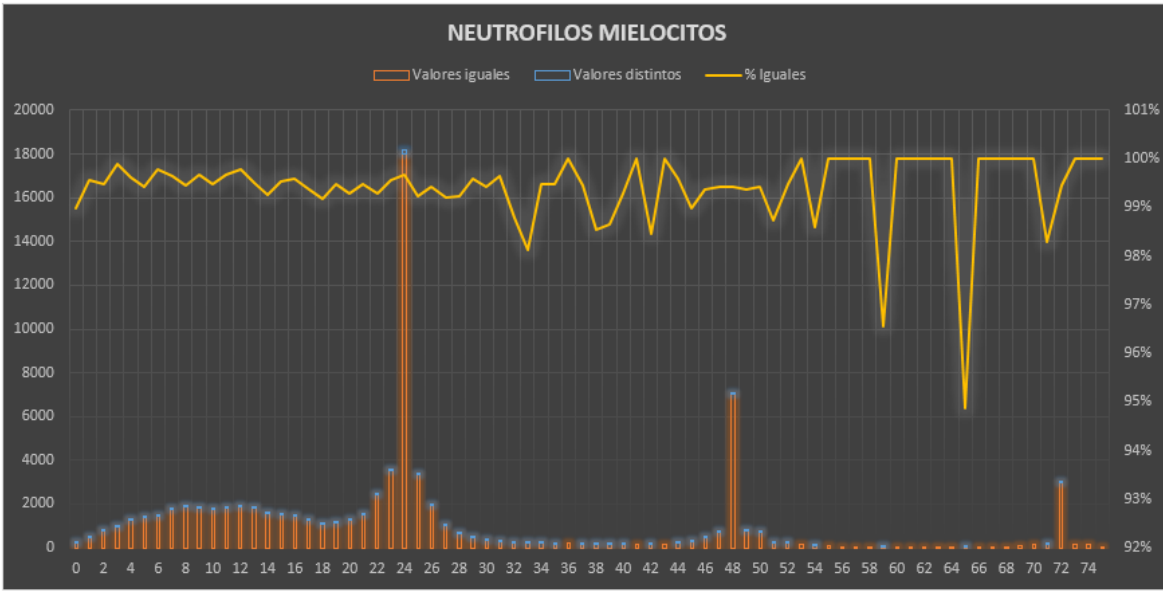


Figura 7 – Cantidad de internaciones por frecuencia de extracción

NEUTROFILOS SEGMENTADOS – Figura 8, en la primera hora las observaciones sin variación son del 19% y luego se desploma a un promedio menor al 1%. En las primeras 24 hs su forma es acampanada con un promedio del 8% para analizar.

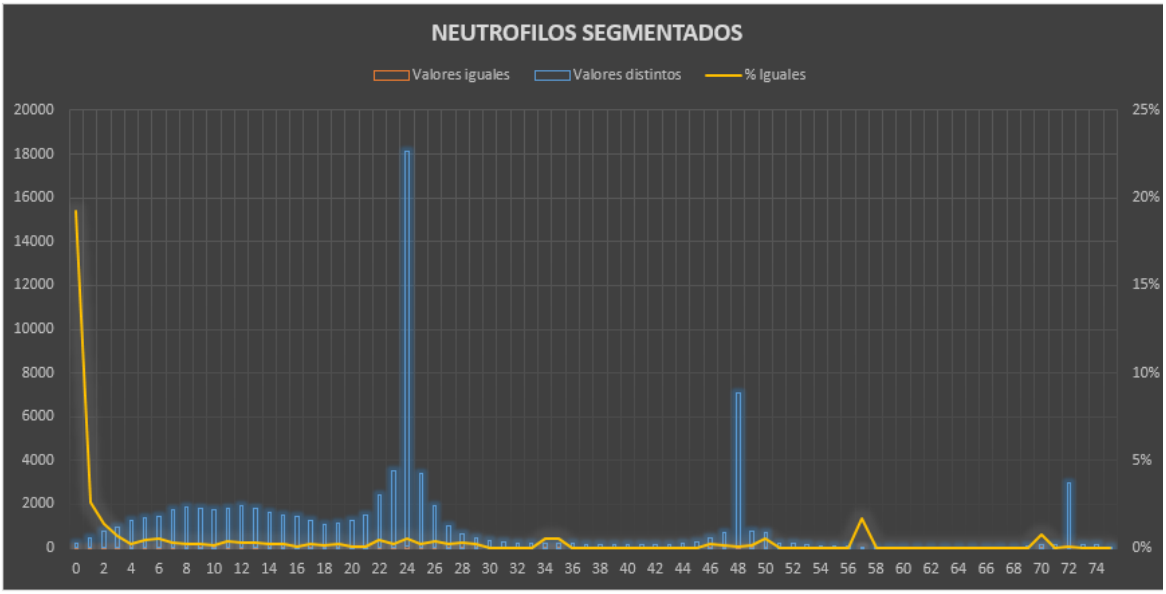


Figura 8 – Cantidad de internaciones por frecuencia de extracción

VCM - VOLUMEN CORPUSCULAR MEDIO – Figura 9, en la primera hora las observaciones sin variación son del 22% y luego mantiene un promedio de 4%. En las primeras 24 hs su forma es acampanada con un promedio del 8% para analizar.

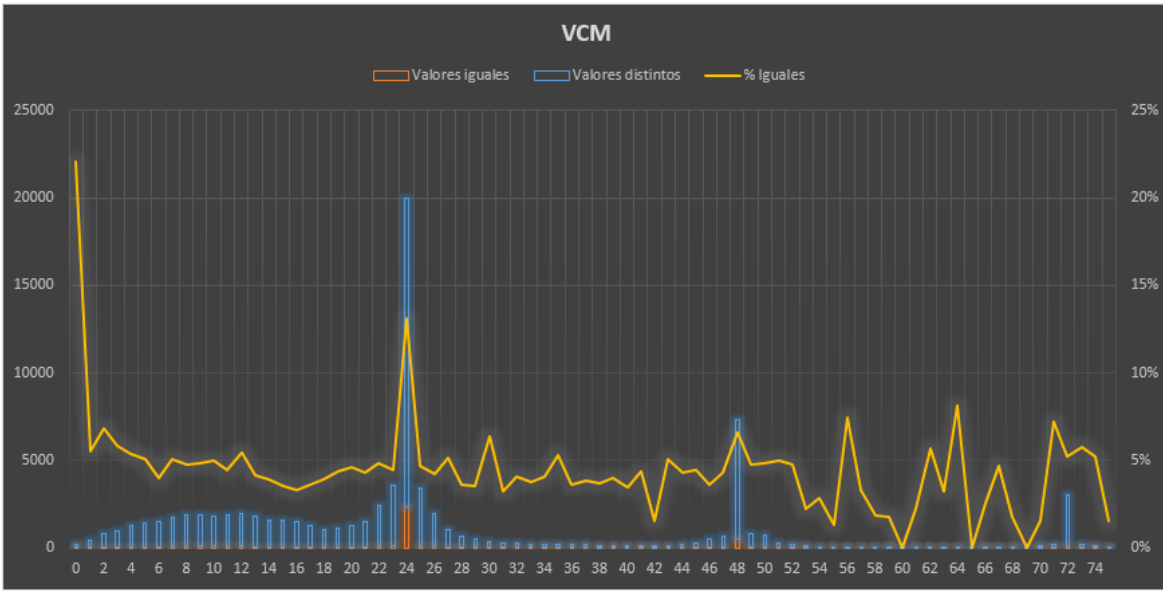


Figura 9 – Cantidad de internaciones por frecuencia de extracción

HCM - HEMOGLOBINA CORPUSCULAR MEDIA – Figura 10, en la primera hora las observaciones sin variación son del 29% y luego mantiene un promedio alto de 9%. En las primeras 24 hs su forma es acampanada con un promedio del 8% para analizar.

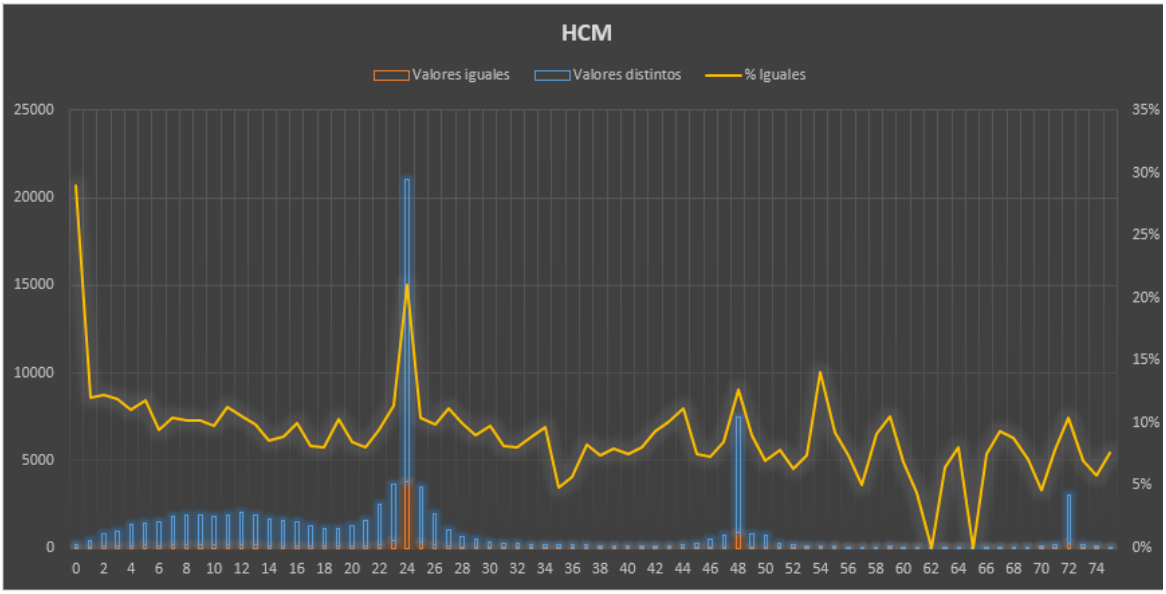


Figura 10 – Cantidad de internaciones por frecuencia de extracción

CHCM - CONCENTRACIÓN DE HCM – Figura 11, en la primera hora las observaciones sin variación son del 26% y luego mantiene un promedio de 7%. En las primeras 24 hs su forma es acampanada con un promedio del 8% para analizar.

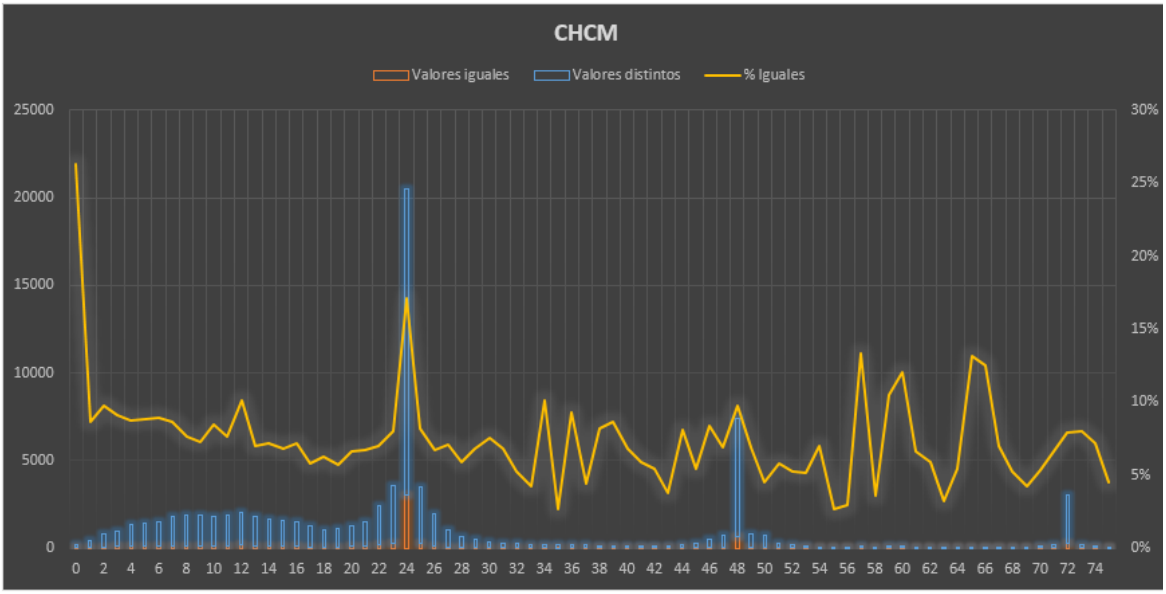


Figura 11 – Cantidad de internaciones por frecuencia de extracción

RDW – Figura 12, en la primera hora las observaciones sin variación son del 29% y luego mantiene un promedio de 8%. En las primeras 24 hs su forma es acampanada con un promedio del 12% para analizar.

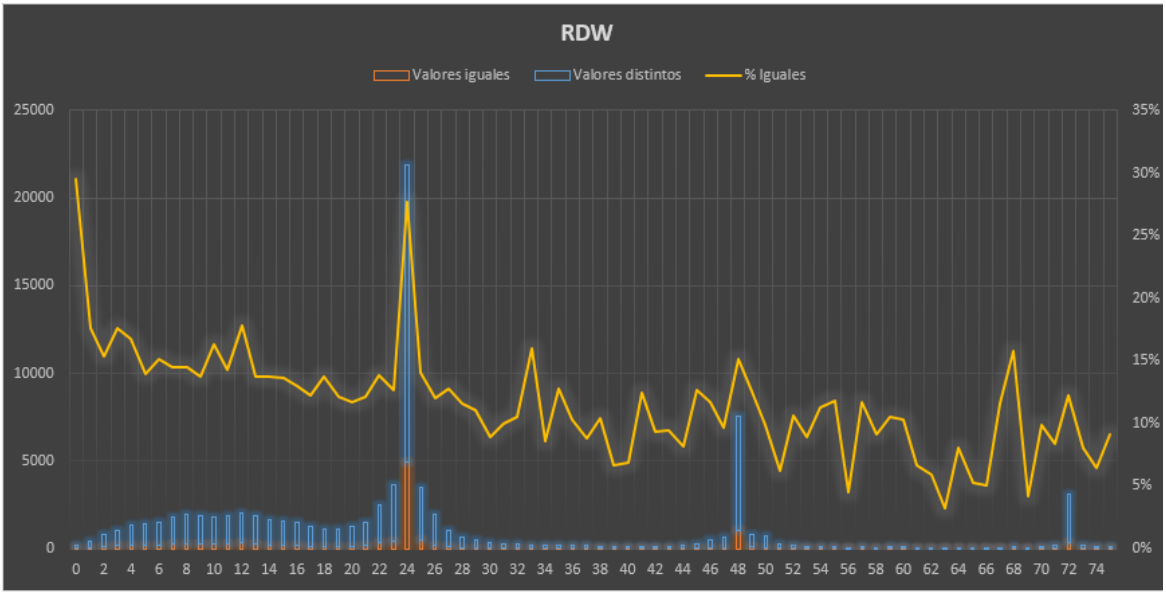


Figura 12 – Cantidad de internaciones por frecuencia de extracción

NEUTROFILOS METAMIELOCITOS, Figura 13, su comportamiento es como los "Neutrófilos Mielocitos", las cantidades de observaciones que cambian con su medición anterior son muy pocas y por ello se mantienen cerca del 100%.

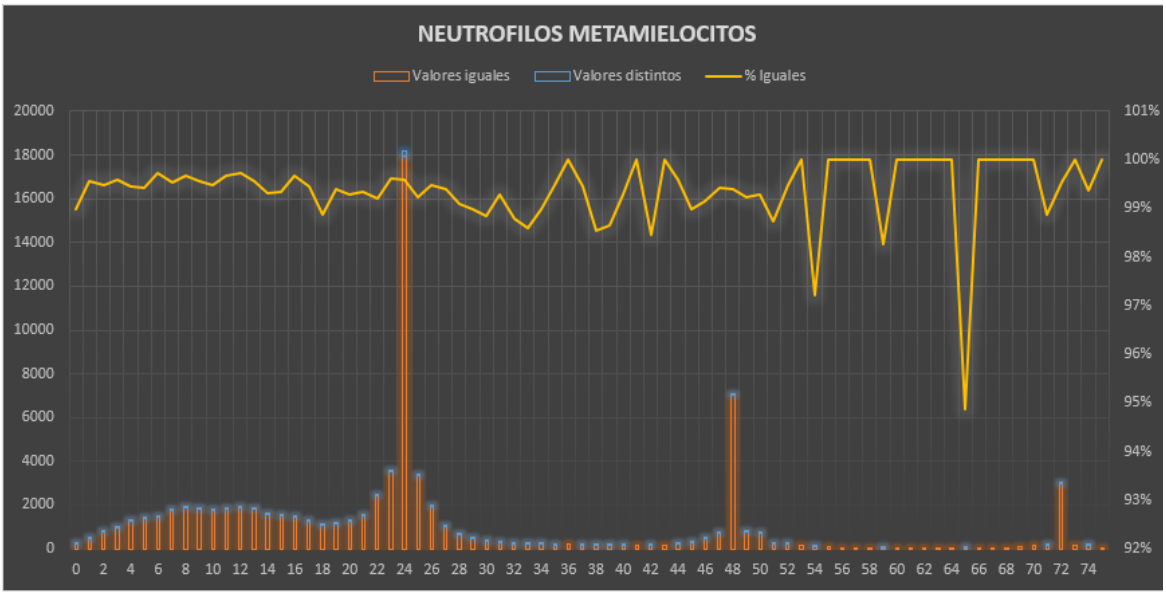


Figura 13 – Cantidad de internaciones por frecuencia de extracción

NEUTROFILOS EN CAYADO – Figura 14, su comportamiento es como los "Neutrófilos Mielocitos", las cantidades de observaciones que cambian con su medición anterior son muy pocas y por ello se mantienen cerca del 100%.

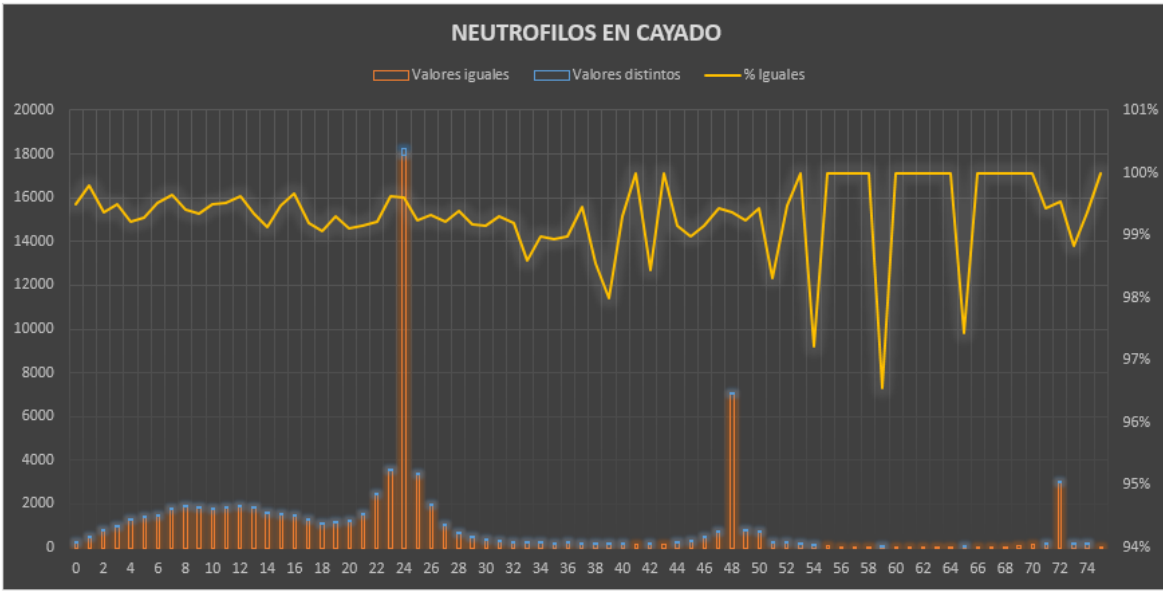


Figura 14 – Cantidad de internaciones por frecuencia de extracción

BASOFILOS – Figura 15, en la primera hora las observaciones sin variación son del 23% y luego mantiene un promedio de 5%. En las primeras 24 hs su forma es acampanada con un promedio del 8% para analizar.

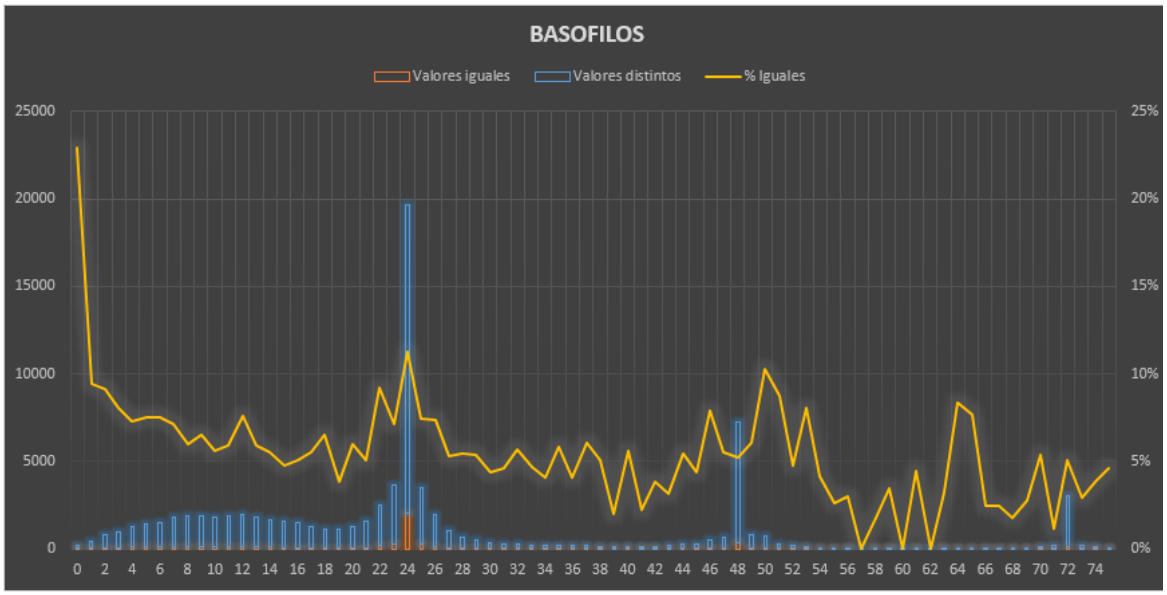


Figura 15 – Cantidad de internaciones por frecuencia de extracción

EOSINOFILOS – Figura 16, en la primera hora las observaciones sin variación son del 24%, luego decrece con valores altos entre el 12% al 9% y luego mantiene un promedio de 5%. En las primeras 24 hs su forma es acampanada con un promedio del 8% para analizar.

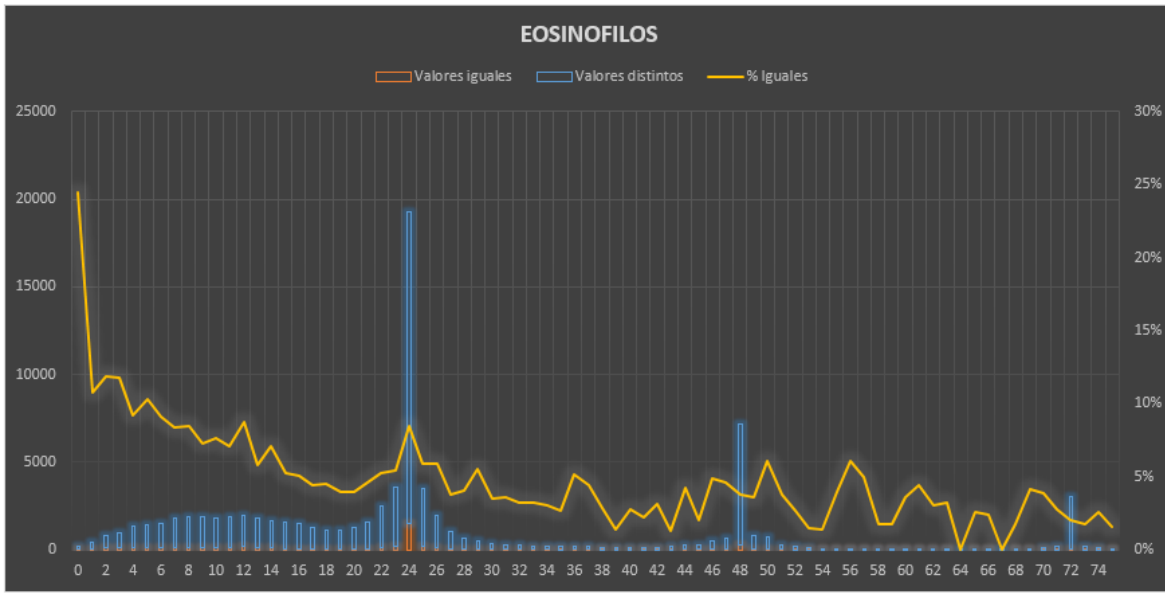


Figura 16 – Cantidad de internaciones por frecuencia de extracción

LINFOCITOS – Figura 17, en la primera hora las observaciones sin variación son del 20% y luego se plancha a un promedio de 1%. En las primeras 24 hs su forma es acampanada con un promedio del 8% para analizar.

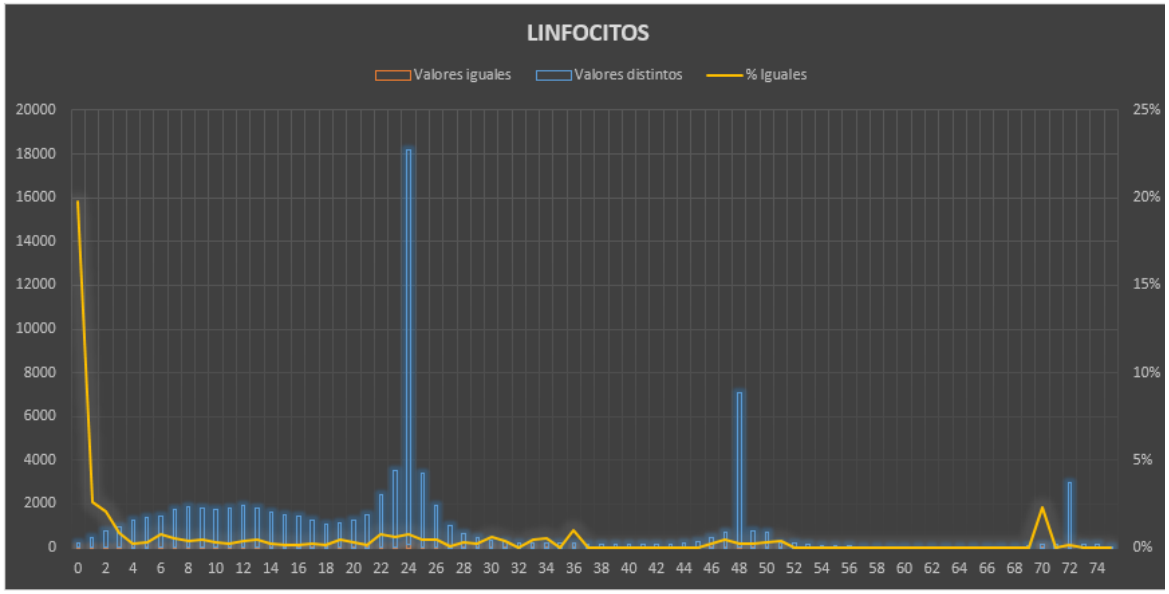


Figura 17 – Cantidad de internaciones por frecuencia de extracción

MONOCITOS – Figura 18, en la primera hora las observaciones sin variación son del 20% y luego se plancha a un promedio de 1%. En las primeras 24 hs su forma es acampanada con un promedio del 8% para analizar.

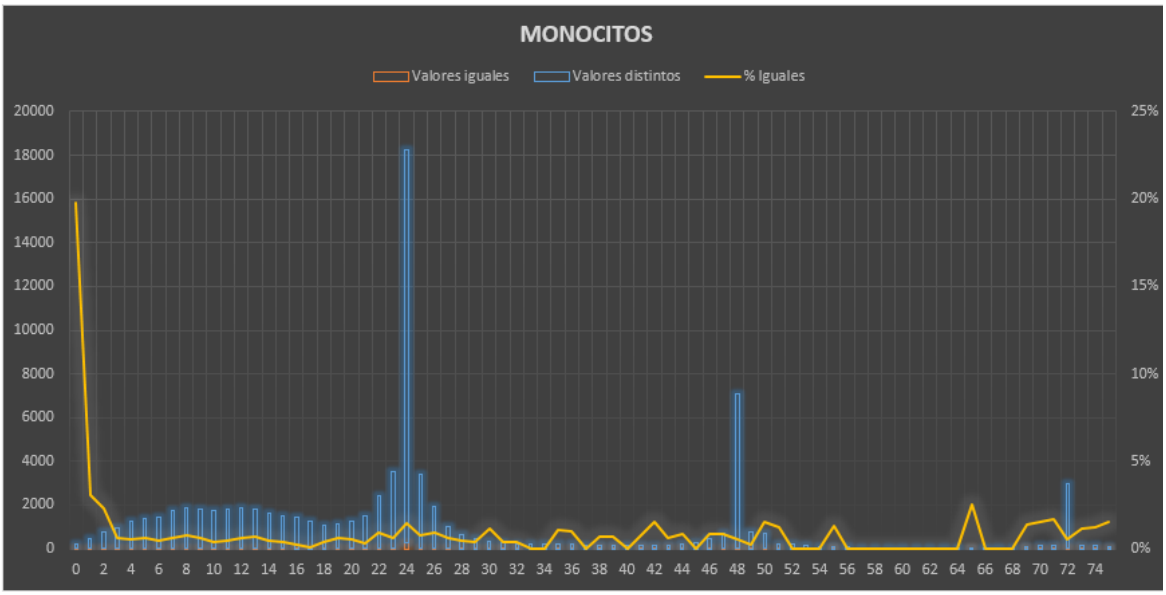


Figura 18 – Cantidad de internaciones por frecuencia de extracción

CELULAS DE DOWNEY – Figura 19, su comportamiento es como los “Neutrofilos Mielocitos”, las cantidades de observaciones que cambian con su medición anterior son muy pocas y por ello se mantienen cerca del 100%.

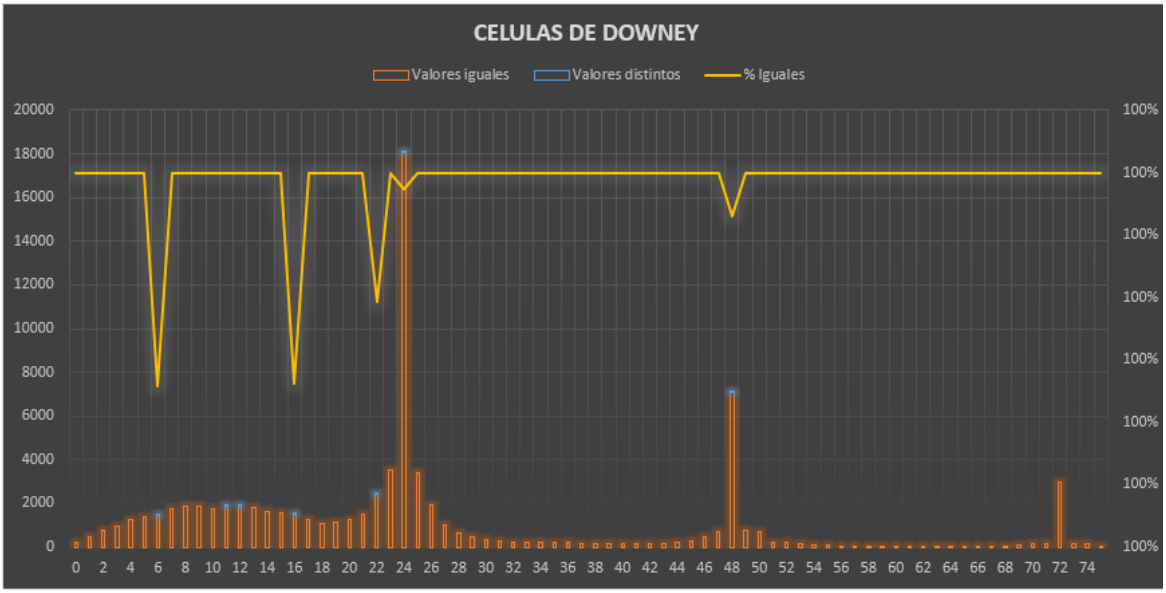


Figura 19 – Cantidad de internaciones por frecuencia de extracción

8.6. Pre procesamiento para métodos de minería de datos

Los métodos de minería de datos, necesitan de un formato diferente a los que se acostumbran a utilizar en las bases de datos transaccionales, donde para nuestro caso, por cada variable se genera un nuevo registro con su resultado de laboratorio.

Para ello se tomó para cada internación y fecha de la muestra, los resultados de las variables y se las traslado de forma columnar, como se las almacena tradicionalmente, a un formato de fila.

A resultado de esta transformación y nuevo conjunto de datos, se los denomina **dataset**. En la tabla 2 podemos observar el dataset a utilizar y su diccionario de datos.

Atributo	Descripción
id	Número de registro
fecha_realizacion	Fecha de la solicitud
fecha_hora_realizacion	Fecha y hora de la solicitud
hora_realizacion	Hora de la solicitud
minuto_realizacion	Minutos de la solicitud
id_paciente	Código del paciente dentro de la institución
edad	Edad del paciente al momento de la internación
grupo_etario_5	Rango de edad correspondiente en agrupaciones de 5 años
grupo_etario_10	Rango de edad correspondiente en agrupaciones de 10 años
sexo	Sexo de nacimiento
id_tipo_episodio	Código de tipo de episodio
tipo_episodio	Tipo de episodio, identifica el tipo de ingreso del paciente (Ambulatorio, General, Guardia, etc.)
nro_historia	Número de historia clinica que se le abre al paciente
id_internacion	Código de internación
razon_internacion	Causa de la internación
id_sector	Código de sector de internación
sector	Sector de internación (Terapia intensiva, Pediatria, Internación general, Oncología, Maternidad, etc.)
vn_348	Variable de analisis de laboratorio numero 1
vn_368	Variable de analisis de laboratorio numero 2
vn_369	Variable de analisis de laboratorio numero 3
vn_379	Variable de analisis de laboratorio numero 4
vn_463	Variable de analisis de laboratorio numero 5
vn_6082	Variable de analisis de laboratorio numero 6
vn_6083	Variable de analisis de laboratorio numero 7
vn_6098	Variable de analisis de laboratorio numero 8
vn_6099	Variable de analisis de laboratorio numero 9
vn_6100	Variable de analisis de laboratorio numero 10
vn_6101	Variable de analisis de laboratorio numero 11
vn_6520	Variable de analisis de laboratorio numero 12
vn_6521	Variable de analisis de laboratorio numero 13
vn_7911	Variable de analisis de laboratorio numero 14
vn_7912	Variable de analisis de laboratorio numero 15
vn_7913	Variable de analisis de laboratorio numero 16
vn_7914	Variable de analisis de laboratorio numero 17
vn_7915	Variable de analisis de laboratorio numero 18

Tabla 2 - Dataset

8.6.1. Análisis de estadística descriptiva

Antes de comenzar a realizar un estudio multivariante de nuestros datos es necesario conocer el comportamiento que tiene cada una de las variables individualmente mediante la estadística descriptiva.

Primero analizaremos los estadísticos descriptivos básicos (media, mediana, minino, máximo, primer y tercer cuartil) de las 18 variables. Tabla 3.

Mujeres	GLUCOSA	HEMATIES RECUENTO	HEMATO CRITO	HEMOGLO BINA	LEUCOCITOS RECUENTO	NEUTRO FILOS MIELOCITOS
Min.	0,0	1,500	6,00	4,00	80	0,00000
1st Qu.	90,0	3,370	29,80	9,90	6280	0,00000
Median	104,0	3,890	34,40	11,50	8300	0,00000
Mean	115,8	3,846	33,94	11,34	9351	0,01800
3rd Qu.	127,0	4,330	38,20	12,80	11000	0,00000
Max.	900,0	8,320	60,40	23,30	458800	30,00000
Hombres						
Min.	20,0	1,500	11,40	4,00	80	0,00000
1st Qu.	93,0	3,360	29,90	10,00	6599	0,00000
Median	108,0	3,980	35,60	11,90	8620	0,00000
Mean	122,5	3,968	35,34	11,85	9776	0,02146
3rd Qu.	135,0	4,570	40,70	13,70	11528	0,00000
Max.	996,0	7,630	65,00	22,90	215000	27,00000

Mujeres	NEUTRO FILOS SEGMENTADOS	VCM	HCM	CHCM	RDW	NEUTRO FILOS METAMIELOCITOS
Min.	0,00	51,9	18,00	24,30	11,00	0,00000
1st Qu.	62,50	85,5	28,40	32,80	13,60	0,00000
Median	72,59	89,0	29,90	33,50	14,60	0,00000
Mean	71,33	88,6	29,62	33,40	15,27	0,01837
3rd Qu.	82,05	92,3	31,10	34,00	16,10	0,00000
Max.	99,36	142,8	45,00	45,40	39,80	25,00000
Hombres						
Min.	0,00	53,2	18,00	22,40	11,00	0,00000
1st Qu.	63,89	86,2	28,80	33,00	13,60	0,00000
Median	73,68	89,7	30,20	33,60	14,60	0,00000
Mean	72,30	89,4	30,01	33,54	15,24	0,02105
3rd Qu.	82,88	93,1	31,50	34,20	16,20	0,00000
Max.	99,46	149,3	45,00	44,60	40,00	31,00000

Mujeres	NEUTRO FILOS EN CAYADO	BASOFILOS	EOSINO FILOS	LINFOCITOS	MONOCITOS	CELULAS DE DOWNEY
Min.	0,00000	0,0000	0,000	0,00	0,000	0,00000
1st Qu.	0,00000	0,2000	0,210	9,61	5,660	0,00000
Median	0,00000	0,4000	1,030	16,33	7,680	0,00000
Mean	0,01638	0,5033	1,750	18,29	7,964	0,00168
3rd Qu.	0,00000	0,6800	2,420	24,79	9,800	0,00000
Max.	24,00000	19,8000	81,500	99,05	60,000	40,00000
Hombres						
Min.	0,00000	0,0000	0,000	0,00	0,000	0,00000
1st Qu.	0,00000	0,2000	0,220	8,60	5,940	0,00000
Median	0,00000	0,3700	1,100	14,69	8,090	0,00000
Mean	0,02051	0,4604	1,847	16,83	8,378	0,00330
3rd Qu.	0,00000	0,6100	2,600	22,60	10,270	0,00000
Max.	31,00000	19,0600	79,900	100,00	60,000	40,00000

Tabla 3 – Estadística descriptiva

En un primer análisis, por la diferencia entre la media y la media nos determina que en la mayoría de las variables no son normales, en el siguiente puto lo veremos en detalle.

Por otro lado, analizando los mínimos, mediana y máximos, detectamos la existencia de valores mínimos y máximos que comparados con la mediana darían la presencia de outliers, pero en las variables de laboratorio se debe tener en cuenta que se miden por un rango aceptable para cada variable para definir si el valor observado es aceptable o normal, lo cual observamos en los resultados entregados a los pacientes, pero se debe saber, que internamente se maneja un segundo rango de mínimos y máximos que solo es observado por los profesionales de la salud. Por ello pueden figurarnos en los gráficos como outliers, pero no lo son.

8.6.2. Test de normalidad

Para los algoritmos que se utilizan en este análisis necesitamos que las variables cumplan con el test de normalidad.

La prueba de normalidad utilizada, es comparar la Media con la Mediana, dado que una de las propiedades del modelo normal es que se cumpla que la Media, Moda y Mediana coinciden (μ). Por ello que si la división de una sobre otra está en el rango de 0.95 y 1.05 la consideraremos de distribución normal.

	DATOS REALES					
	Mujer			Hombre		
Variable	Media	Mediana	Variación	Media	Mediana	Variación
GLUCOSA	115,821	104,000	1,114	122,515	108,000	1,134
HEMATIES RECUENTO	3,846	3,890	0,989	3,968	3,980	0,997
HEMATOCRITO	33,943	34,400	0,987	35,336	35,600	0,993
HEMOGLOBINA	11,335	11,500	0,986	11,851	11,900	0,996
LEUCOCITOS RECUENTO	9351,186	8300,000	1,127	9776,305	8620,000	1,134
NEUTROFILOS MIELOCITOS	0,018	0,000		0,021	0,000	
NEUTROFILOS SEGMENTADOS	71,333	72,590	0,983	72,303	73,680	0,981
VCM	88,606	89,000	0,996	89,403	89,700	0,997
HCM	29,620	29,900	0,991	30,006	30,200	0,994
CHCM	33,404	33,500	0,997	33,540	33,600	0,998
RDW	15,273	14,600	1,046	15,236	14,600	1,044
NEUTROFILOS METAMIELOCITOS	0,018	0,000		0,021	0,000	
NEUTROFILOS EN CAYADO	0,016	0,000		0,021	0,000	
BASOFILOS	0,503	0,400	1,258	0,460	0,370	1,244
EOSINOFILOS	1,750	1,030	1,699	1,847	1,100	1,679
LINFOCITOS	18,294	16,330	1,120	16,834	14,690	1,146
MONOCITOS	7,964	7,680	1,037	8,378	8,090	1,036
CELULAS DE DOWNEY	0,002	0,000		0,003	0,000	

Tabla 4 – Test de Normalidad

Para lograr la normalidad de las variables se utilizó la transformación de BOX – COX. Las transformaciones de Box y Cox son una familia de transformaciones potenciales usadas en estadística para corregir sesgos en la distribución de errores, para corregir varianzas desiguales (para diferentes valores de la variable predictora) y principalmente para corregir

la no linealidad en la relación (mejorar correlación entre las variables). Esta transformación recibe el nombre de los estadísticos George E. P. Box y David Cox.[4]

	BOX – COX					
	Mujer			Hombre		
Variable	Media	Mediana	Variación	Media	Mediana	Variación
GLUCOSA	0,113	0,115	0,983	0,004	0,004	0,972
HEMATIES RECUENTO	10,526	10,597	0,993	6,347	6,359	0,998
HEMATOCRITO	279,329	281,839	0,991	85,909	86,293	0,996
HEMOGLOBINA	97,354	98,231	0,991	25,896	25,938	0,998
LEUCOCITOS RECUENTO	5,180	5,168	1,002	7,545	7,507	1,005
NEUTROFILOS MIELOCITOS	0,997	1,000	0,997	0,996	1,000	0,996
NEUTROFILOS SEGMENTADOS	4564,286	4552,087	1,003	5574,612	5577,102	1,000
VCM	20066,410	20135,610	0,997	20664,130	20707,210	0,998
HCM	2370,448	2391,210	0,991	2408,012	2420,640	0,995
CHCM	3447,095	3457,440	0,997	3243,123	3249,000	0,998
RDW	0,001	0,001	0,972	0,001	0,001	0,974
NEUTROFILOS METAMIELOCITOS	0,997	1,000	0,997	0,996	1,000	0,996
NEUTROFILOS EN CAYADO	0,997	1,000	0,997	0,996	1,000	0,996
BASOFILOS	0,652	0,652	1,000	0,649	0,645	1,007
EOSINOFILOS	0,706	0,699	1,009	0,716	0,708	1,011
LINFOCITOS	2,360	2,374	0,994	1,834	1,844	0,995
MONOCITOS	2,479	2,501	0,991	2,308	2,333	0,989
CELULAS DE DOWNEY	1,000	1,000	1,000	1,000	1,000	1,000

Tabla 5 – Test de Normalidad

8.6.3. Correlación de Variables

Se considera que **dos variables** cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos de la otra: si tenemos **dos variables** (A y B) existe **correlación** entre ellas si al disminuir los valores de A lo hacen también los de B y viceversa.[5]

El índice de correlación varía en el intervalo [-1,1], estableciendo el signo el sentido de la relación, y la interpretación de cada resultado es el siguiente:

- **Si r = 1:** Correlación positiva perfecta. El índice refleja la dependencia total entre ambas dos variables, la que se denomina relación directa: cuando una de las variables aumenta, la otra variable aumenta en proporción constante.
- **Si 0 < r < 1:** Refleja que se da una correlación positiva.
- **Si r = 0:** En este caso no hay una relación lineal. Aunque no significa que las variables sean independientes, ya que puede haber relaciones no lineales entre ambas variables.
- **Si -1 < r < 0:** Indica que existe una correlación negativa.
- **Si r = -1:** Indica una **correlación negativa perfecta** y una dependencia total entre ambas variables lo que se conoce como **"relación inversa"**, que es cuando una de las variables aumenta, la otra variable en cambio disminuye en proporción constante.

La correlación refleja la medida de **asociación entre variables.** Si se aplica en probabilidad y estadística, la correlación permite conocer la fuerza y dirección de la relación lineal que se dé entre dos variables aleatorias.[6]

Se separaron los índices de correlación en dos tablas, una para el sexo Femenino (Tabla 6) y otra para el sexo Masculino (Tabla 7).

Analizando y observando las dos tablas en conjunto podemos determinar que independientemente del sexo la correlación entre las variables hay una similitud entre ellas, tanto en la dirección (positiva o negativa) como en la fortaleza, cuanto más se acerca a 1 o a -1 su correlación es más fuerte.

MUJERES																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1,00	0,04	0,04	0,05	-0,18	0,00	-0,34	0,01	0,01	0,01	0,02	0,00	0,01	-0,22	-0,30	0,31	0,22	0,00
2	0,04	1,00	0,90	0,87	0,11	0,06	-0,08	-0,32	-0,33	-0,18	0,25	0,06	0,05	-0,12	-0,02	0,14	-0,01	0,00
3	0,04	0,90	1,00	0,99	0,10	0,06	-0,09	0,10	0,07	-0,04	0,38	0,07	0,06	-0,12	-0,01	0,15	-0,01	0,00
4	0,05	0,87	0,99	1,00	0,07	0,06	-0,10	0,15	0,16	0,12	0,44	0,07	0,06	-0,12	-0,01	0,16	-0,01	0,00
5	-0,18	0,11	0,10	0,07	1,00	-0,10	0,47	-0,05	-0,09	-0,15	0,02	-0,11	-0,12	0,22	0,28	-0,47	-0,30	0,00
6	0,00	0,06	0,06	0,06	-0,10	1,00	0,06	0,00	0,00	0,01	0,08	0,89	0,72	0,00	-0,03	0,03	-0,03	0,01
7	-0,34	-0,08	-0,09	-0,10	0,47	0,06	1,00	-0,02	-0,03	-0,04	-0,09	0,06	0,03	0,47	0,59	-0,93	-0,56	0,02
8	0,01	-0,32	0,10	0,15	-0,05	0,00	-0,02	1,00	0,95	0,32	0,24	0,00	0,00	0,02	0,01	0,01	0,02	0,00
9	0,01	-0,33	0,07	0,16	-0,09	0,00	-0,03	0,95	1,00	0,60	0,31	0,00	0,00	0,03	0,02	0,03	0,02	0,00
10	0,01	-0,18	-0,04	0,12	-0,15	0,01	-0,04	0,32	0,60	1,00	0,35	0,01	0,01	0,04	0,03	0,05	0,02	0,00
11	0,02	0,25	0,38	0,44	0,02	0,08	-0,09	0,24	0,31	0,35	1,00	0,08	0,08	-0,06	-0,05	0,14	-0,01	0,00
12	0,00	0,06	0,07	0,07	-0,11	0,89	0,06	0,00	0,00	0,01	0,08	1,00	0,80	-0,01	-0,04	0,04	-0,02	0,01
13	0,01	0,05	0,06	0,06	-0,12	0,72	0,03	0,00	0,00	0,01	0,08	0,80	1,00	-0,02	-0,04	0,05	0,00	0,02
14	-0,22	-0,12	-0,12	-0,12	0,22	0,00	0,47	0,02	0,03	0,04	-0,06	-0,01	-0,02	1,00	0,45	-0,43	-0,27	0,00
15	-0,30	-0,02	-0,01	-0,01	0,28	-0,03	0,59	0,01	0,02	0,03	-0,05	-0,04	-0,04	0,45	1,00	-0,48	-0,31	0,00
16	0,31	0,14	0,15	0,16	-0,47	0,03	-0,93	0,01	0,03	0,05	0,14	0,04	0,05	-0,43	-0,48	1,00	0,34	-0,01
17	0,22	-0,01	-0,01	-0,01	-0,30	-0,03	-0,56	0,02	0,02	0,02	-0,01	-0,02	0,00	-0,27	-0,31	0,34	1,00	0,00
18	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,00	0,00	0,00	0,00	0,01	0,02	0,00	0,00	-0,01	0,00	1,00

Tabla 6 – Correlación de Variables – Sexo Femenino

HOMBRES																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1,00	0,05	0,05	0,05	-0,16	-0,01	-0,29	0,01	0,02	0,02	0,01	-0,01	0,00	-0,18	-0,26	0,26	0,18	-0,01
2	0,05	1,00	0,93	0,91	0,12	0,06	-0,11	-0,28	-0,27	-0,10	0,34	0,06	0,06	-0,13	-0,07	0,16	0,03	0,00
3	0,05	0,93	1,00	0,99	0,11	0,06	-0,11	0,08	0,07	0,01	0,44	0,06	0,06	-0,12	-0,06	0,17	0,03	0,00
4	0,05	0,91	0,99	1,00	0,09	0,06	-0,12	0,11	0,14	0,15	0,49	0,06	0,06	-0,11	-0,06	0,18	0,04	0,00
5	-0,16	0,12	0,11	0,09	1,00	-0,10	0,44	-0,02	-0,08	-0,17	0,04	-0,11	-0,12	0,21	0,28	-0,45	-0,28	0,00
6	-0,01	0,06	0,06	0,06	-0,10	1,00	0,08	-0,01	0,00	0,01	0,08	0,88	0,76	-0,02	-0,05	0,02	-0,04	0,00
7	-0,29	-0,11	-0,11	-0,12	0,44	0,08	1,00	0,00	-0,03	-0,07	-0,09	0,07	0,06	0,46	0,58	-0,92	-0,55	0,03
8	0,01	-0,28	0,08	0,11	-0,02	-0,01	0,00	1,00	0,94	0,28	0,21	-0,01	-0,01	0,04	0,01	0,00	0,00	0,00
9	0,02	-0,27	0,07	0,14	-0,08	0,00	-0,03	0,94	1,00	0,58	0,29	0,00	0,00	0,04	0,02	0,02	0,02	0,00
10	0,02	-0,10	0,01	0,15	-0,17	0,01	-0,07	0,28	0,58	1,00	0,36	0,01	0,01	0,04	0,02	0,08	0,04	-0,01
11	0,01	0,34	0,44	0,49	0,04	0,08	-0,09	0,21	0,29	0,36	1,00	0,07	0,07	-0,05	-0,08	0,13	0,03	-0,01
12	-0,01	0,06	0,06	0,06	-0,11	0,88	0,07	-0,01	0,00	0,01	0,07	1,00	0,86	-0,02	-0,05	0,04	-0,03	0,00
13	0,00	0,06	0,06	0,06	-0,12	0,76	0,06	-0,01	0,00	0,01	0,07	0,86	1,00	-0,03	-0,05	0,04	-0,02	0,00
14	-0,18	-0,13	-0,12	-0,11	0,21	-0,02	0,46	0,04	0,04	0,04	-0,05	-0,02	-0,03	1,00	0,47	-0,42	-0,27	0,00
15	-0,26	-0,07	-0,06	-0,06	0,28	-0,05	0,58	0,01	0,02	0,02	-0,08	-0,05	-0,05	0,47	1,00	-0,48	-0,31	0,00
16	0,26	0,16	0,17	0,18	-0,45	0,02	-0,92	0,00	0,02	0,08	0,13	0,04	0,04	-0,42	-0,48	1,00	0,32	-0,02
17	0,18	0,03	0,03	0,04	-0,28	-0,04	-0,55	0,00	0,02	0,04	0,03	-0,03	-0,02	-0,27	-0,31	0,32	1,00	0,00
18	-0,01	0,00	0,00	0,00	0,00	0,00	0,03	0,00	0,00	-0,01	-0,01	0,00	0,00	0,00	0,00	-0,02	0,00	1,00

Tabla 7 – Correlación de Variables – Sexo Masculino

Rangos de definición:

- **Si r = 1:** Correlación positiva perfecta.
- **Si 0.75 < r < 1:** Correlación positiva muy fuerte.
- **Si 0.5 < r < 0.75:** Correlación positiva fuerte.
- **Si 0.25 < r < 0.5:** Correlación positiva débil.
- **Si 0 < r < 0.25:** Correlación positiva muy débil.

- **Si r = 0:** No hay una relación lineal.
- **Si -0.25 < r < 0:** Correlación negativa muy débil.
- **Si -0.5 < r < 0.25:** Correlación negativa débil.
- **Si -0.75 < r < 0.5:** Correlación negativa fuerte.
- **Si -1 < r < 0.75:** Correlación negativa muy fuerte.
- **Si r = -1:** Correlación negativa perfecta.

Según el rango de definición, las correlaciones entre las variables se componen como se observa en la Tabla 8.

	Total		Positivo		Negativo	
Hasta 0,25	239	78,1%	141	76,6%	98	80,3%
Hasta 0,50	45	14,7%	25	13,6%	20	16,4%
Hasta 0,75	8	2,6%	6	3,3%	2	1,6%
Hasta 0,99	14	4,6%	12	6,5%	2	1,6%

Tabla 8 – Distribución de las correlaciones según sus rangos.

- Un 92,8% del total se ubica dentro de una correlación débil y solo el 7,2% del total en una correlación fuerte.
- Un 60,1% corresponde a una correlación positiva y un 39,9 a una correlación negativa.
- Dentro del 7,2% con una correlación fuerte, el 81,8% es positiva y solo el 18,2% negativa.
- En las correlaciones fuertes, en las positivas, las muy fuertes es el doble de las fuertes, mientras que en las negativas son iguales.

En la siguiente tabla (Tabla 9), podemos observar aquellas variables que tienen una fuerte correlación, tanto negativa como positiva. Es importante aclarar, que la dirección de las flechas, en la tabla, son meramente indicativas a fin de señalar que el comportamiento de la variable es igual a la de su correlación u opuesta a ella, en el caso de los negativos. Esto quiere decir, que si tienen el mismo sentido, cuando una variable sube, su correlativa también subirá o si su valor baja su correlativa bajará también, mientras que, donde el sentido es opuesto si el valor de una variable baja su correlativa subirá o viceversa.

Variable		M	H		Variable
3 HEMATOCRITO	↑	0,986	0,990	↑	4 HEMOGLOBINA
8 VCM	↑	0,947	0,943	↑	9 HCM
2 HEMATIES RECUENTO	↑	0,902	0,930	↑	3 HEMATOCRITO
6 NEUTROFILOS MIELOCITOS	↑	0,887	0,883	↑	12 NEUTROFILOS METAMIELOCITOS
2 HEMATIES RECUENTO	↑	0,866	0,906	↑	4 HEMOGLOBINA
12 NEUTROFILOS METAMIELOCITOS	↑	0,801	0,856	↑	13 NEUTROFILOS EN CAYADO
6 NEUTROFILOS MIELOCITOS	↑	0,718	0,756	↑	13 NEUTROFILOS EN CAYADO
9 HCM	↑	0,604	0,575	↑	10 CHCM
7 NEUTROFILOS SEGMENTADOS	↑	0,586	0,578	↑	15 EOSINOFILOS
7 NEUTROFILOS SEGMENTADOS	↑	-0,557	-0,549	↓	17 MONOCITOS
7 NEUTROFILOS SEGMENTADOS	↑	-0,932	-0,923	↓	16 LINFOCITOS

Tabla 9 – Variables con fuerte correlación

Correlación muy fuerte positiva:

- El hematocrito mide la cantidad de sangre compuesta por glóbulos rojos. Los glóbulos rojos contienen una proteína llamada hemoglobina que transporta oxígeno de los pulmones al resto del cuerpo. Por lo que es correcto que estas dos variables estén fuertemente correlacionadas y que cuando una de ellas presenta niveles elevados o bajos, la otra la acompaña.
- Las variables VCM (tamaño promedio de los glóbulos rojos) y HCM (cantidad de hemoglobina por glóbulo rojo) corresponden a los índices de glóbulos rojos, lo que nos indica esta correlación es que si el tamaño medio de los hematíes aumenta la cantidad media de hemoglobina que contiene cada hematíes o glóbulos rojos o lo mismo sucede si disminuye.
- Los hematíes recuento es el conteo de glóbulos rojos en sangre y el hematocrito que mide la cantidad de sangre, están fuertemente correlacionadas, ya que si crece o disminuye el valor del recuento la cantidad de sangre hará lo mismo.
- Los mielocitos se derivan de los promielocitos y dan lugar a metamielocitos. Los promielocitos son los que darán origen a los diferentes mielocitos y, éstos, a su vez, madurarán en metamielocitos. Por lo mencionado estas dos variables tienen una fuerte correlación.
- Como mencionamos anteriormente, los hematíes recuento es el conteo de glóbulos rojos en sangre, mientras que la hemoglobina es una proteína que transporta oxigeno de los pulmones al resto del cuerpo. Estas variables están fuertemente correlacionadas, dado que, si el conteo de glóbulos rojo crece o decrece en sangre, habrá más o menos hemoglobina. Es interesante observar que en el sexo masculino la correlación es mayor al sexo femenino, debido a la contextura muscular que posee y que requiere el transporte de mayor cantidad de oxígeno.

- Los neutrófilos se forman en la médula ósea a partir de un precursor común y se van diferenciando en una serie de células (mieloblasto, promielocito, mielocito, metamielocito, cayado) para finalmente llegar a segmentado. Estas variables también están fuertemente correlacionadas.

Correlación fuerte positiva:

- En una correlación menor entre los neutrófilos, se encuentran los mielocitos con los en cayado o blandos.
- En este grupo de variables también encontramos a HCM (cantidad de hemoglobina por glóbulo rojo) y CHCM que es la cantidad de hemoglobina relativa al tamaño de la célula (concentración de hemoglobina) por glóbulo rojo, al formar parte de los índices de glóbulos rojos esta última variable esta correlacionada con HCM en menor medida.
- Por último, dentro de este grupo encontramos a los neutrófilos segmentados que tienen una correlación fuerte con los eosinófilos, estos corresponden a tipos de glóbulos blancos en sangre, también llamados leucocitos.

Correlación negativa:

Entre ellas encontramos una correlación fuerte entre los neutrófilos segmentados y los monocitos y otra muy fuerte entre los primeros y los linfocitos, estos dos corresponden también a tipos de glóbulos blancos en sangre, pero a diferencia de los antes mencionados en el grupo de correlación fuerte positiva, el comportamiento de estas es a la inversa de los neutrófilos segmentados. Lo que quiere decir es que si los glóbulos blancos de tipos segmentados aumentan las otras mencionadas disminuyen o viceversa.

8.7. Clusterización

El análisis de clúster hace referencia a la familia de algoritmos que permiten agrupar registros similares de un conjunto de datos en grupos. A cada uno de estos grupos es a lo que se denomina un clúster. El objetivo final del análisis es asignar a cada clúster los registros que son similares entre sí. Al mismo tiempo que los registros del resto de clústeres son diferentes. Estas son técnicas de aprendizaje no supervisado, con las que es posible descubrir patrones ocultos en los conjuntos de datos.

El algoritmo de k-means [7] es uno de los más utilizados en análisis de clúster. Esto es debido a su simplicidad y ser fácilmente interpretable. En este método, para llevar a cabo el análisis de clúster, solamente es necesario indicar el número de clústeres. Inicialmente se generan aleatoriamente tantos puntos, a los que se les denomina centroides, en el espacio de propiedades como clústeres. Cada registro se asigna al clúster del centroide más cercano. En

cada uno de los clústeres se actualiza la posición de los centroides con el valor promedio de sus registros. Volviéndose a realizar la asignación de los registros a un clúster con los nuevos centroides. Este proceso se repite hasta que la posición de los centroides no cambie por encima de un umbral de una iteración a la siguiente.[8]

La diferencia más notable entre el aprendizaje supervisado y el no supervisado radica en los resultados. El aprendizaje no supervisado crea una nueva variable, la etiqueta, mientras que el aprendizaje supervisado predice un resultado. La máquina ayuda al profesional en la búsqueda de etiquetar los datos en función de la estrecha relación.

Los seres humanos más allá de su sexo, se diferencian por el grupo etario al que pertenecen a través de su vida. Por ello hemos armado 4 grupos de estudio para este análisis. En el primero ubicaremos a los que van de 0 (cero) años a 9 (nueve) años y lo denominaremos INFANCIA, el segundo de 10 (diez) años a 24 (veinticuatro) años llamado ADOLESCENCIA Y JUVENTUD, tercero de 25 (veinticinco) a 64 (sesenta y cuatro) años llamado ADULTOS y por cuarto y último de 65 (sesenta y cinco) años en adelante y lo denominaremos ADULTOS MAYORES.

La distribución de nuestras muestras, según su sexo y grupo etario la podemos observar en la Figura 20.

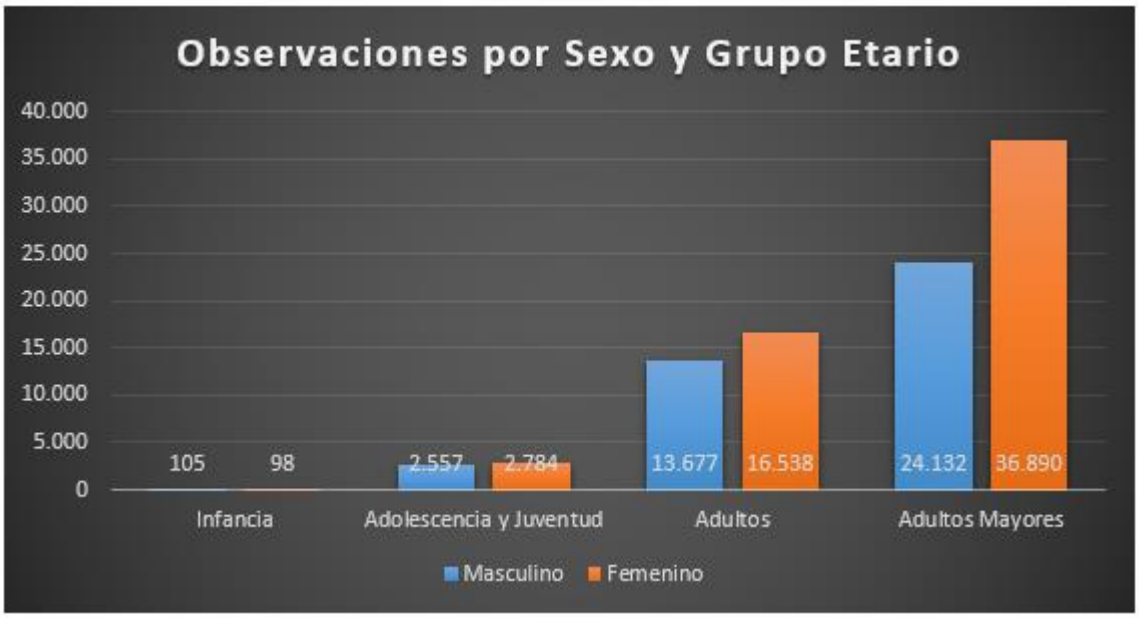


Figura 20 – Cantidad de observaciones por Sexo y Grupo Etario

Podemos observar el crecimiento de observaciones a través de los grupos etarios y salvo en el primer grupo, en los siguientes grupos es mayor el del sexo Femenino.

8.7.1 Validación del clustering

La validación de clúster es el proceso por el cual se evalúa la veracidad de los grupos obtenidos. A modo general, este proceso consta de tres partes: estudio de la tendencia de clustering, elección del número óptimo de clústeres y estudio de la calidad/significancia de los clústeres generados.

Antes de aplicar un método de clustering a los datos es conveniente evaluar si hay indicios de que realmente existe algún tipo de agrupación en ellos.

Hopkins statistics

El estadístico Hopkins permite evaluar la tendencia de clustering de un conjunto de datos mediante el cálculo de la probabilidad de que dichos datos procedan de una distribución uniforme, es decir, estudia la distribución espacial aleatoria de las observaciones. La forma de calcular este estadístico es la siguiente:

- Extraer una muestra uniforme de n observaciones ($p1..., pn$) del set de datos estudiado.
- Para cada observación pi seleccionada, encontrar la observación vecina más cercana pj y calcular la distancia entre ambas, $xi=dist\ (pi, pj)$.
- Simular un conjunto de datos de tamaño n ($q1..., qn$) extraídos de una distribución uniforme con la misma variación que los datos originales.
- Para cada observación simulada qi, encontrar la observación vecina más cercana qj y calcular la distancia entre ambas, $yi=dist\ (qi, qj)$.
- Calcular el estadístico Hopkins (H) como la media de las distancias de vecinos más cercanos en el set de datos simulados, dividida por la suma de las medias de las distancias vecinas más cercanas del set de datos original y el simulado.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

Valores de H en torno a 0.5 indican que $\Sigma xini=1$ y $\Sigma yini=1$ son muy cercanos el uno al otro, es decir, que los datos estudiados se distribuyen uniformemente y que por lo tanto no tiene sentido aplicar clustering. Cuanto más se aproxime a 0 el estadístico H, más evidencias se tienen a favor de que existen agrupaciones en los datos y de que, si se aplica clustering correctamente, los grupos resultantes serán reales.

Femenino			
Infancia	Adolescencia y Juventud	Adultos	Adultos Mayores
0,3477618	0,0716197	0,0363290	0,0211636

Masculino			
Infancia	Adolescencia y Juventud	Adultos	Adultos Mayores
0,2977835	0,0783116	0,0482599	0,0469932

Tabla 10 - Resultados H para cada grupo etario por sexo.

Los resultados, en la Tabla 10, muestran evidencias de que las observaciones del set de datos, para los grupos de Adolescencia y Juventud, Adultos y Adultos Mayores, no siguen una distribución espacial uniforme, su estructura contiene algún tipo de agrupación. Por contra, el valor del estadístico H obtenido para el set de datos de Infancia es muy próximo a 0.5, aunque no tanto por lo que puede indicar que los datos están uniformemente distribuidos y desaconseja la utilización de métodos de clustering.

8.7.2 Número óptimo de clústeres

La selección de la cantidad de clústeres es una dificultad encontrada con k-mean. Puede establecer un valor alto de, es decir, una gran cantidad de grupos, para mejorar la estabilidad, pero puede terminar con un ajuste excesivo de datos (overfit). Sobreajuste (overfitting) significa que el rendimiento del modelo disminuye sustancialmente durante nuevos datos próximos. Una técnica para elegir la mejor k se llama método del codo. Este método utiliza homogeneidad intragrupal o heterogeneidad intragrupal para evaluar la variabilidad. En otras palabras, le interesa el porcentaje de la varianza explicada por cada grupo. Puede esperar que la variabilidad aumente con el número de clústeres; alternativamente, la heterogeneidad disminuye. Nuestro desafío es encontrar la k que esté más allá de los rendimientos decrecientes. Agregar un nuevo grupo no mejora la variabilidad de los datos porque queda muy poca información por explicar.

La muestra tomada para realizar la clusterización fue tomada del set de datos originales, expresados en el documento. Como el mismo (dataset) muestra varias tomas para un mismo paciente y en la misma internación se decidió tomar solo la primera muestra tomada al momento de la internación. Ya que la misma no se ve influenciada o modificada por los fármacos o evolución dentro de la internación.

No existe una forma única de averiguar el número adecuado de clústeres. Es un proceso bastante subjetivo que depende en gran medida del tipo de clustering empleado y de si se dispone de información previa sobre los datos con los que se está trabajando, por

ejemplo, estudios anteriores pueden sugerir o acotar las posibilidades. A pesar de ello, se han desarrollado varias estrategias que ayudan en el proceso.

Método Elbow

El método Elbow sigue una estrategia comúnmente empleada para encontrar el valor óptimo de un parámetro. La idea general es probar un rango de valores del parámetro en cuestión, representar gráficamente los resultados obtenidos con cada uno e identificar aquel punto de la curva a partir del cual la mejora deja de ser sustancial (principio de verosimilitud).

Método promedio Silhouette

El método de average silhouette es muy similar al de Elbow, con la diferencia de que, en lugar minimizar el total inter-clúster sum of squares (wss), se maximiza la media de los silhouette coeficient (si). Este coeficiente cuantifica cómo de buena es la asignación que se ha hecho de una observación comparando su similitud con el resto de observaciones de su clúster frente a las de los otros clústeres. Su valor puede estar entre -1 y 1, siendo valores altos un indicativo de que la observación se ha asignado al clúster correcto.

Método estadístico Gap

El estadístico gap fue publicado por R.Tibshirani, G.Walther y T. Hastie, autores también del magnífico libro Introduction to Statistical Learning. Este estadístico compara, para diferentes valores de k, la varianza total intra-cluster observada frente al valor esperado acorde a una distribución uniforme de referencia. La estimación del número óptimo de clústeres es el valor k con el que se consigue maximizar el estadístico gap, es decir, encuentra el valor de k con el que se consigue una estructura de clústeres lo más alejada posible de una distribución uniforme aleatoria. Este método puede aplicarse a cualquier tipo de clustering.

Se aplicaron los tres métodos a los 4 (cuatro) grupos etarios definidos oportunamente y por la variable de sexo, femenino y masculino.

El primer análisis es para el grupo etario Infancia, en sus dos sexos. En la Figura 21 podemos ver los tres métodos para cada uno de los sexos. Para el sexo femenino podemos observar que los tres métodos coinciden con que el numero óptimo de clústeres es de 2 (dos), mientras que, para el sexo masculino, los dos primeros métodos coinciden que el valor óptimo de k (cantidad de clústeres) es de 3 (tres), mientras que el estadístico Gap es de 1 (uno).

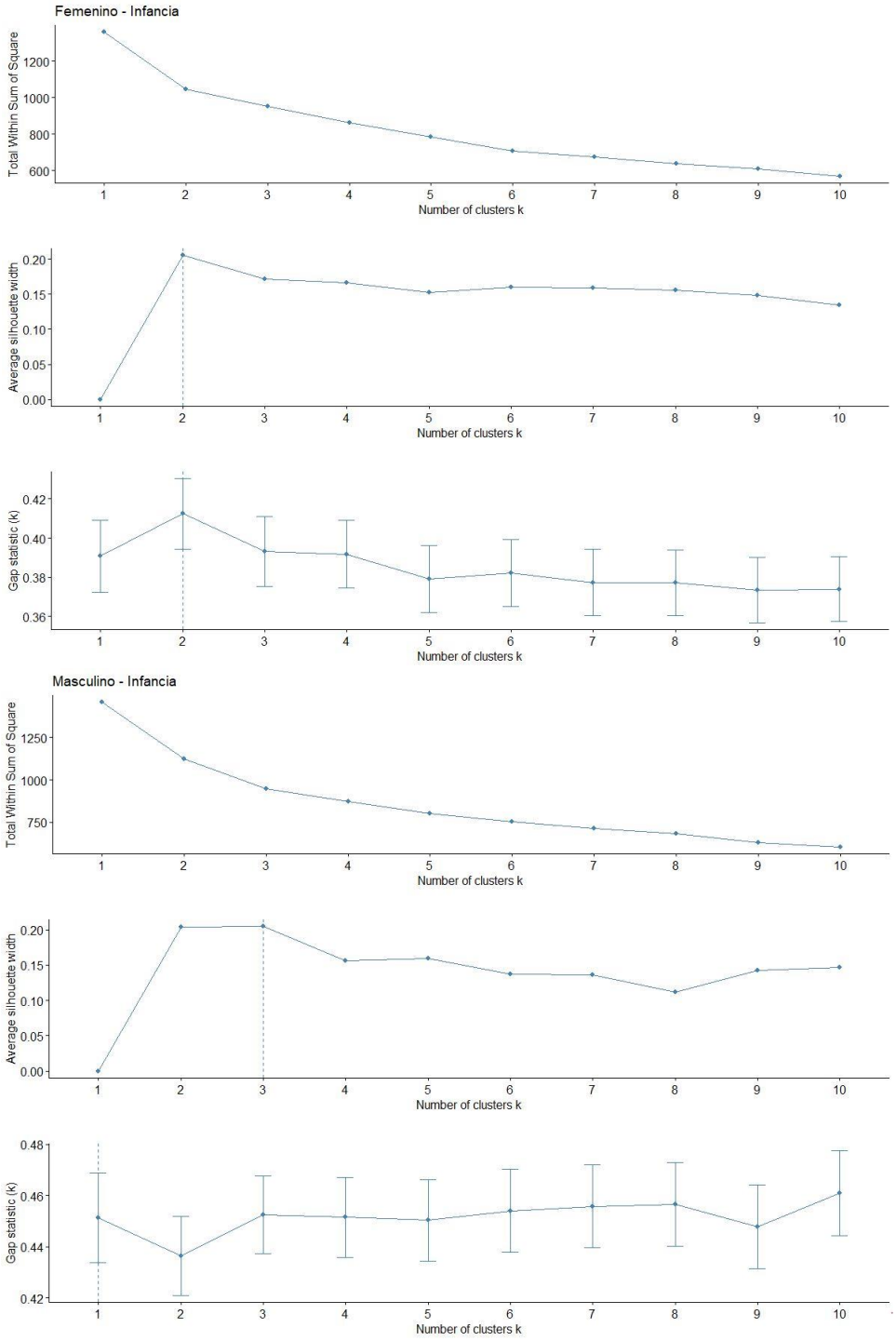


Figura 21 – Grupo etario Infancia

El segundo grupo etario es el de Adolescencia y Juventud (Figura 22), al igual que el grupo anterior, los dos primeros métodos coinciden en 4 (cuatro) clústeres como cantidad optima mientras que el tercer método opta por 1 (un) solo clúster.

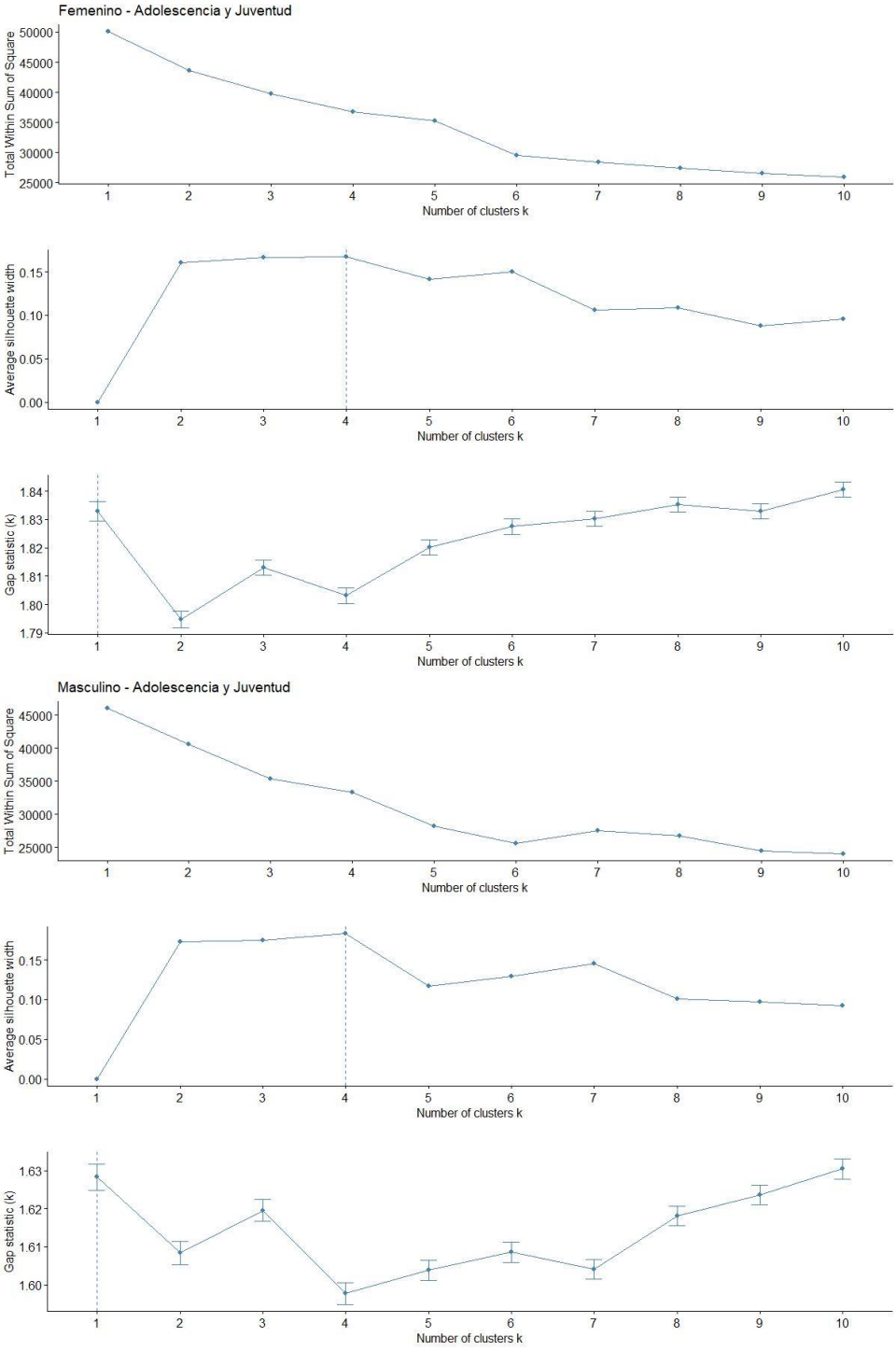


Figura 22 – Grupo etario Adolescencia y Juventud

En el caso del grupo etario de Adultos (Figura 23), los dos primeros muy similares, donde para el sexo femenino es de 3 (tres) y para el masculino de 2 (dos) clústeres, mientras que el tercero sigue determinando como optimo 1 (un) clúster.

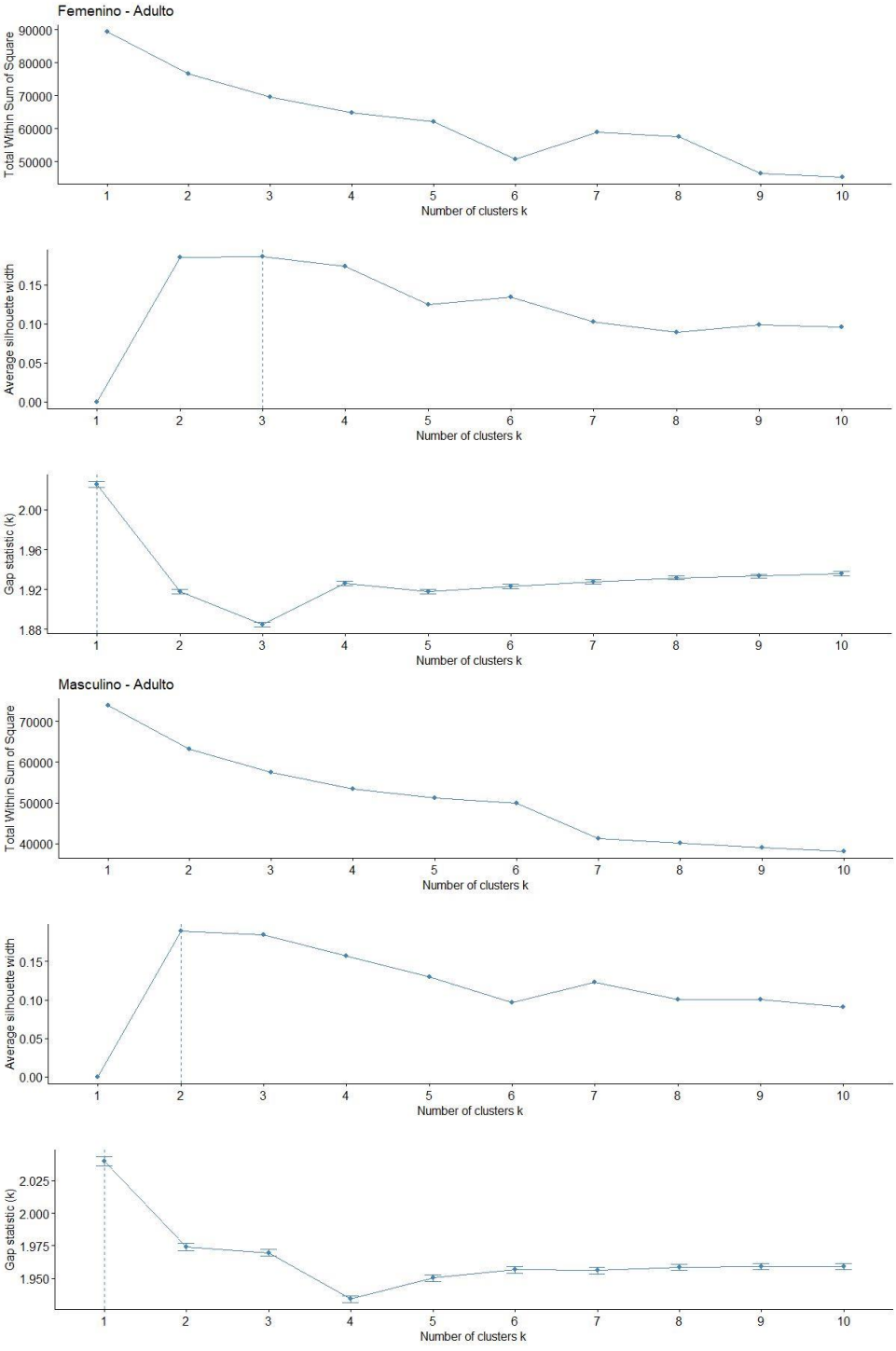


Figura 23 – Grupo etario Adulto

Por último, el grupo etario de Adultos Mayores (Figura 24), la cantidad optima de clústeres es 2 (dos) para ambos sexos. Para el sexo femenino lo tres métodos dan lo mismo,

mientras que, para el masculino, el primer y segundo método propone 2 (dos) clústeres y el tercero 1 (uno).

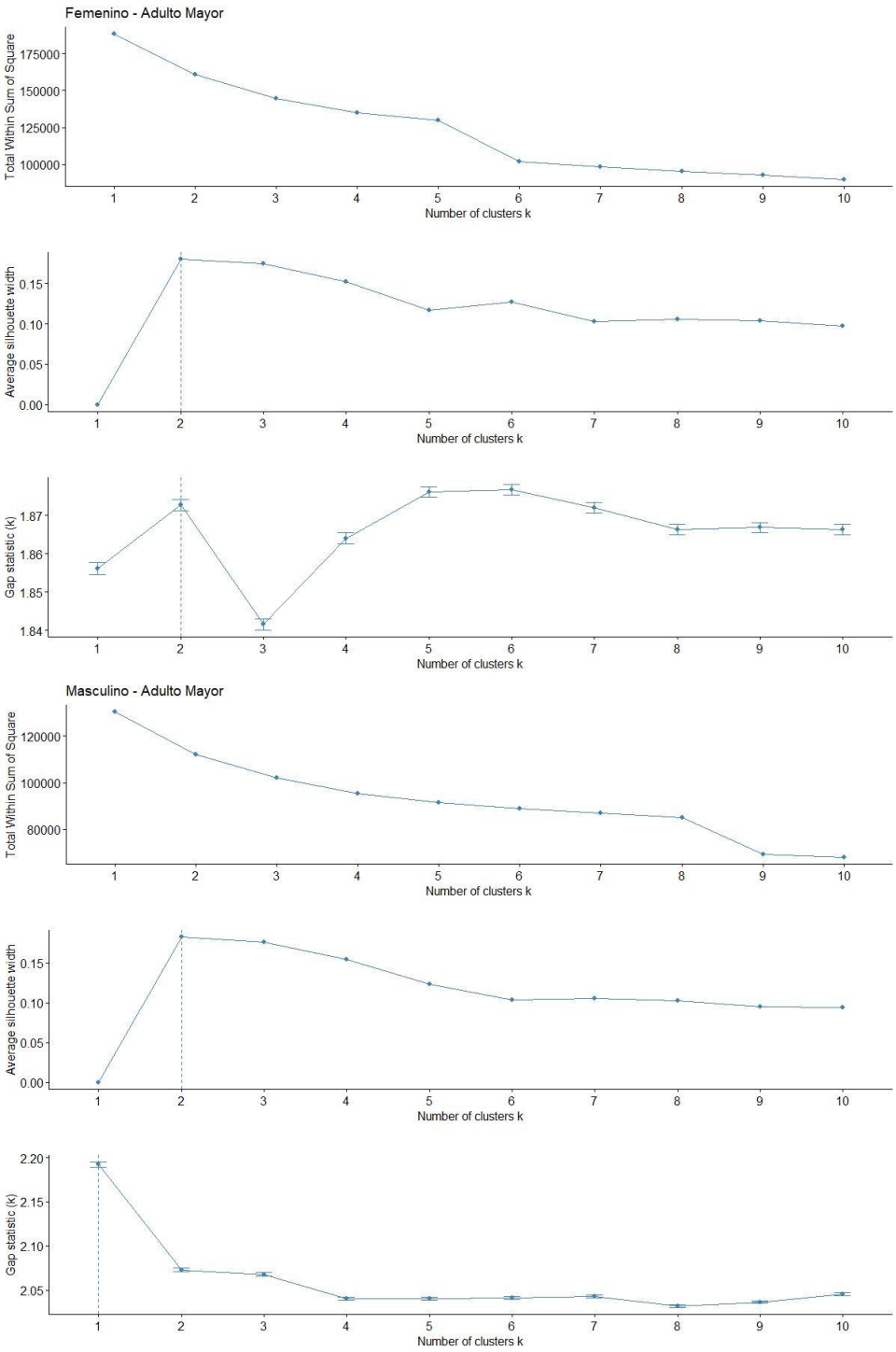


Figura 24 – Grupo etario Adulto Mayor

Como resumen, de los diferentes grupos etarios por sexo, la cantidad optima de clústeres propuestos por los tres métodos los podemos observar en la Tabla 11.

	Infancia	Adolescencia y Juventud	Adultos	Adultos Mayores
Femenino	2	4	3	2
Masculino	3	4	2	2

Tabla 11 – Resumen cantidad de clústeres propuestos.

8.7.3. Calidad de los Clústeres

Una vez seleccionado el número adecuado de clústeres y aplicado el algoritmo de clustering pertinente se tiene que evaluar la calidad de los de los mismos, de lo contrario, podrían derivarse conclusiones de agrupación que no se corresponden con la realidad.

La idea principal detrás del clustering es agrupar las observaciones de forma que sean similares a aquellas que están dentro de un mismo clúster y distintas a las de otros clústeres.

Para verificar la calidad del clúster se utilizó el método **"Silhouette width"** que se basa en cuantificar cómo de buena es la asignación que se ha hecho de una observación comparando su similitud con el resto de observaciones del mismo clúster frente a las de los otros clústeres.

Su valor puede estar entre -1 y 1, siendo valores altos un indicativo de que la observación se ha asignado al clúster correcto. Cuando su valor es próximo a cero significa que la observación se encuentra en un punto intermedio entre dos clústeres. Valores negativos apuntan a una posible asignación incorrecta de la observación.

Comencemos con el grupo etario **"Infancia"**, para ambos sexos. Según el análisis realizado en el punto anterior, nos proponía 2 (dos clústeres para el sexo femenino y 3 (tres) para el masculino.

Para el primero observamos en la Figura 7 una distribución bastante achatada, con un promedio entre los dos clústeres de 0.21, cuanto más bajo este promedio indica que habrá más observaciones que estén en el límite entre los clústeres definidos. Incluso vemos en la figura que el clúster 1 posee valores negativos, lo que indicaría que esa o esas observaciones podrían tener una incorrecta clasificación.

En la Tabla 12, observamos primero la clasificación del clúster con la cantidad de observaciones que hay en cada uno y por último el promedio de las similitudes. En este caso, para el clúster 1 es de 0.18 lo que nos indica que sus observaciones están más cerca del 0 y por ende están más cerca del límite con el otro clúster. También observamos que lo que figuraba en la imagen (Figura 25) como negativo al clúster 1, se corresponde a una observación, justamente del clúster 1 que se sitúa en la frontera del clúster 2.

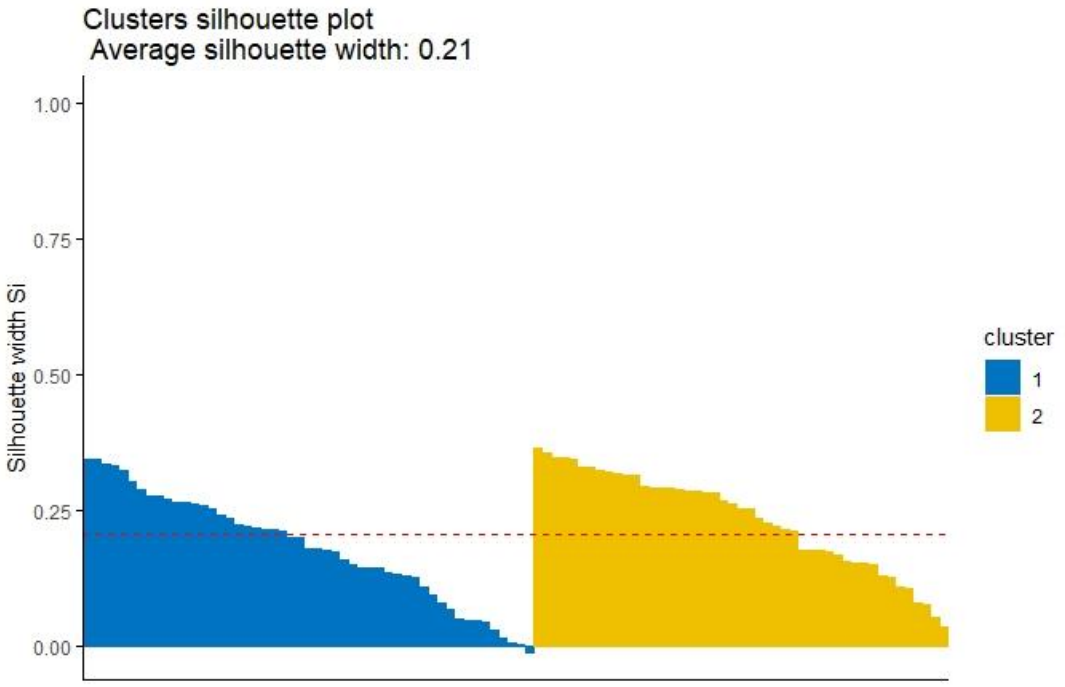


Figura 25 – Femenino - Infancia

id	cluster	size	ave.sil.width
1	1	51	0,18
2	2	47	0,23

id	cluster	neighbor	sil_width
63	1	2	-0,01057628

Tabla 12 – Clasificación de observaciones (Femenino – Infancia)

Por último, podemos observar en la Figura 26 como quedan distribuidos en un plano las observaciones clasificadas en los dos clústeres y las observaciones en conflicto quedaran marcadas en rojo. Como el set de datos analizado está formado por más de dos variables, cada uno de sus ejes corresponden a las 2 primeras variables del algoritmo de Componentes Principales (PCA) que veremos más adelante en este documento.

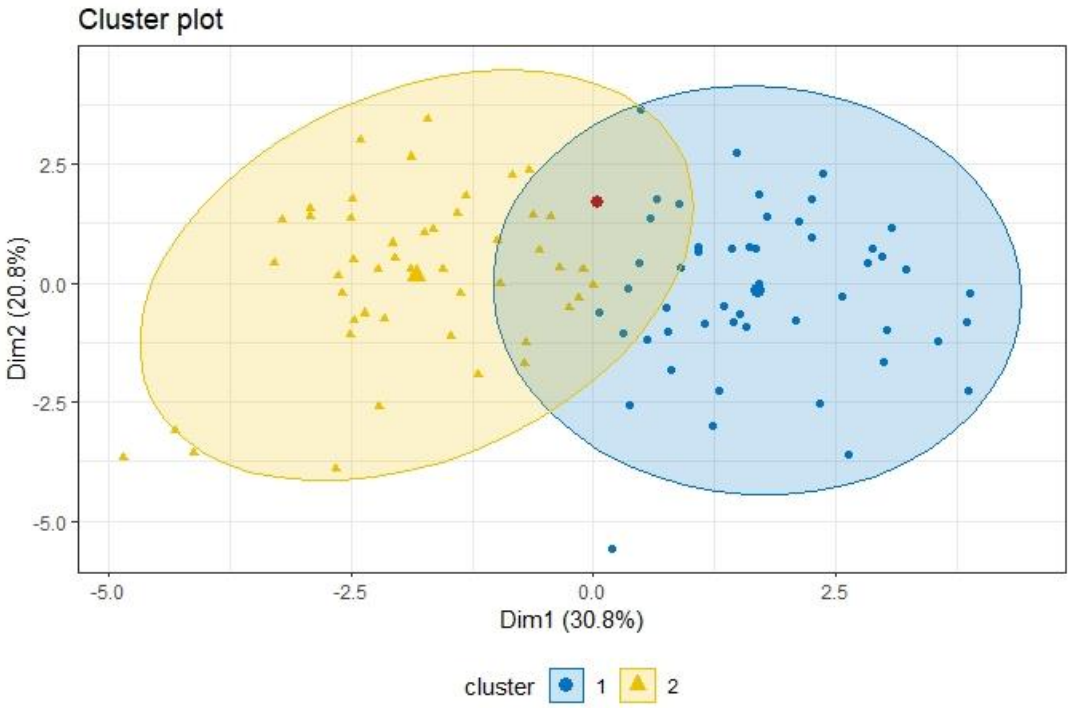


Figura 26 – Plot Femenino - Infancia

En el segundo, masculino, nos recomendaban 3 (tres) clústeres, pero cuando aplicamos el algoritmo para verificar la calidad de los mismo nos encontramos que hay varios valores por debajo del 0 (cero), ver Figura 27. El promedio 0.20 es similar al femenino, pero la distribución de las observaciones varia, siendo para el clúster 1 no tan aplanada lo que indica una agrupación de observaciones más definida, mientras que el clúster 2 es más grande y distribuido y por último el 3 (tres) que aparentemente es el mejor definido, pero tanto el 1 (uno) como el 3 (tres) poseen valores negativos.

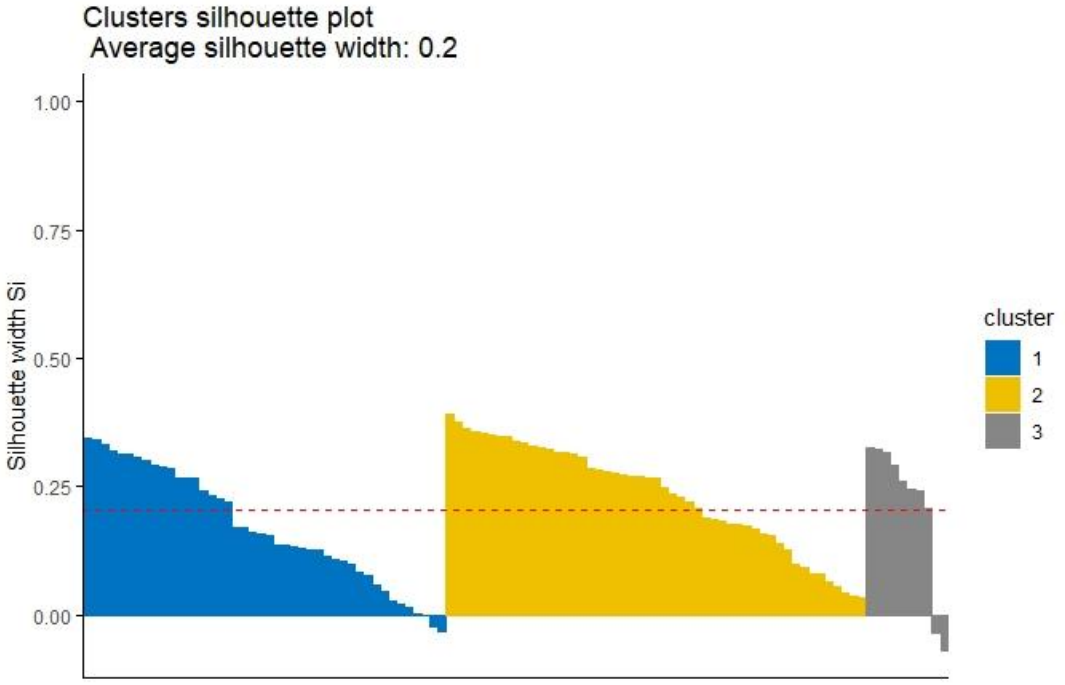


Figura 27 – 3 Clúster Masculino - Infancia

Se decidió aplicar el algoritmo con 2 (dos) clúster para ver su resultado.

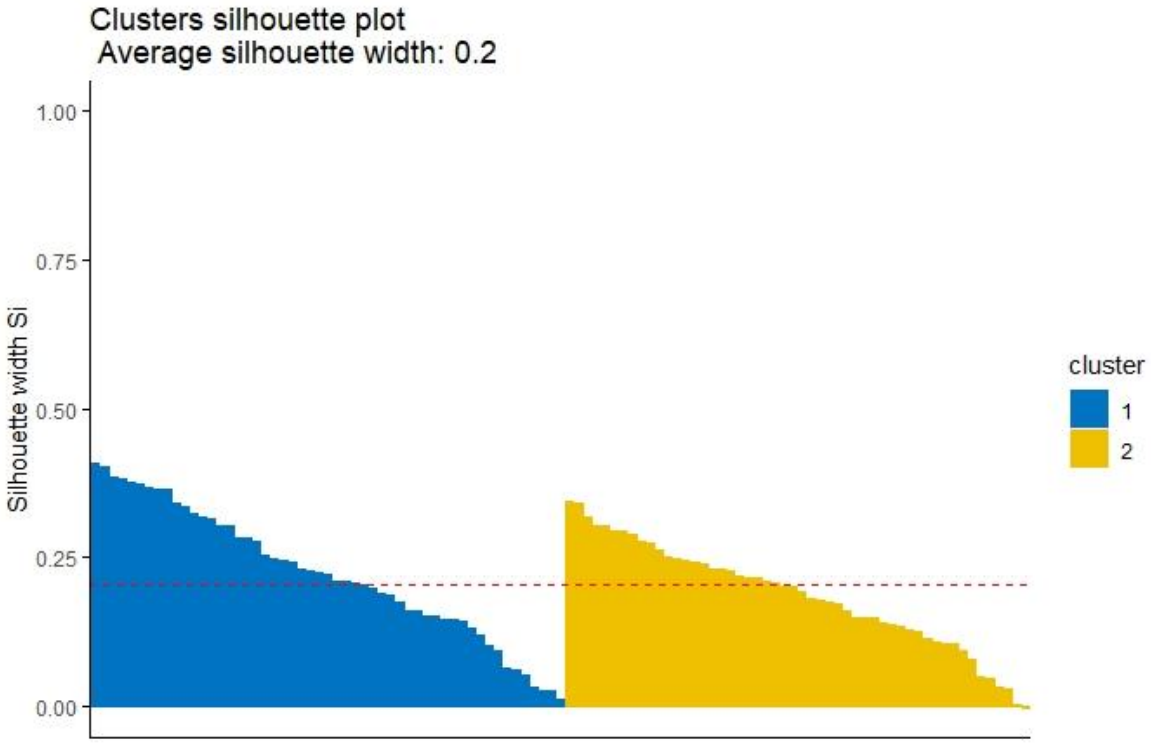


Figura 28 – 2 Clúster Masculino - Infancia

En la Figura 28 podemos observar como su clasificación mejora dejando solo una pequeña parte por debajo de 0 (cero) en el clúster 2 (dos).

Si analizamos los datos de las dos alternativas, con la formación de 3 (tres) y 2 (dos) clúster, podemos observar en la Tabla 13, que en la primera hay 3 (tres) valores por debajo de cero en el clúster 1 (uno) y dos en el 3 (tres), pero el segundo de este último tiene una distancia del cero del doble que el tercer valor del clúster uno, por ello en la Figura 8 se ve más profundo en el clúster 3 (tres) que en el uno. Pero si analizamos la corrida con 2 (dos) clústeres vemos que prácticamente el clúster dos quedo con los mismos valores, solo se incrementó en uno, el promedio del clúster 1 subió de 0.17 a 0.22 y del 2 disminuyo de 0.23 a 0.19, quedando un solo valor por debajo de cero correspondiente el clúster 2 que en la primera corrida con 3 clúster no tenía valores negativos.

id	cluster	size	ave.sil.width
1	1	44	0,17
2	2	51	0,23
3	3	10	0,21

id	cluster	neighbor	sil_width
99	1	2	-0,000615847
51	1	2	-0,021849694
89	1	2	-0,031885766
40	3	2	-0,035402701
54	3	2	-0,068443511

id	cluster	size	ave.sil.width
1	1	53	0,22
2	2	52	0,19

id	cluster	neighbor	sil_width
35	2	1	-0,001859423

Tabla 13 – Clasificación de observaciones (Masculino – Infancia)

Por último, observamos cómo quedan distribuidos en un plano las observaciones clasificadas en los tres clústeres y en los dos clústeres y las observaciones en conflicto quedaran marcadas en rojo, en las Figuras 29 y 30 respectivamente. Se utiliza el mismo método de gráfica, como en el femenino.

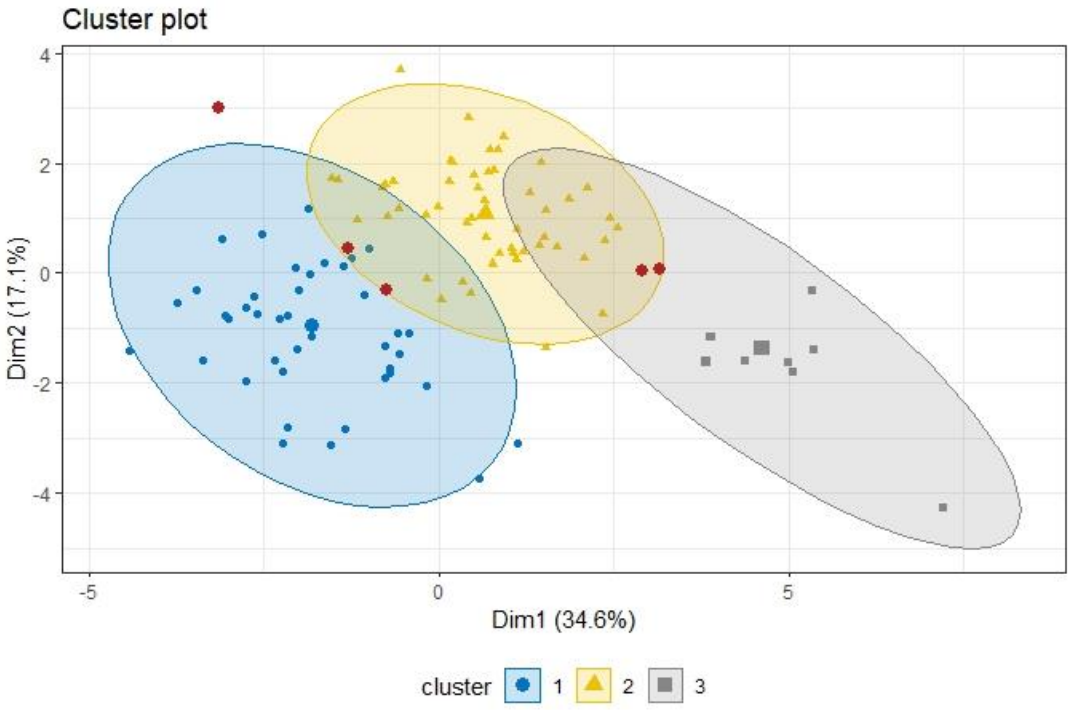


Figura 29 – Plot 3 Clúster Masculino - Infancia

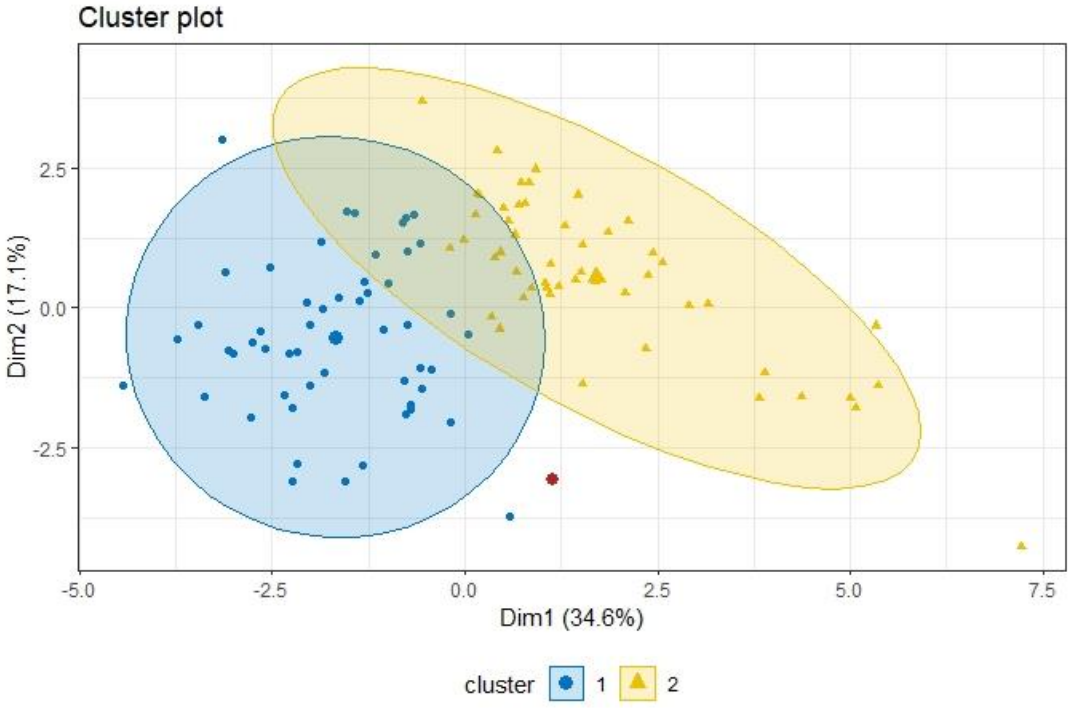


Figura 30 – Plot 2 Clúster Masculino - Infancia

Como este primer grupo etario tiene pocas observaciones y fácil ver cada observación, al clúster que pertenece y cuales están en conflicto. A medida que avancemos

en los otros grupos etarios esto será más difícil, donde los puntos se convertirán en zonas de color y ver las observaciones en conflicto no se listaran por su gran cantidad, se realizó en este grupo por su cantidad, de todas formas, quedaron documentadas para otros análisis.

En el segundo grupo etario **"Adolescencia y Juventud"**, para ambos sexos, la propuesta era de 4 (cuatro) clústeres para cada uno.

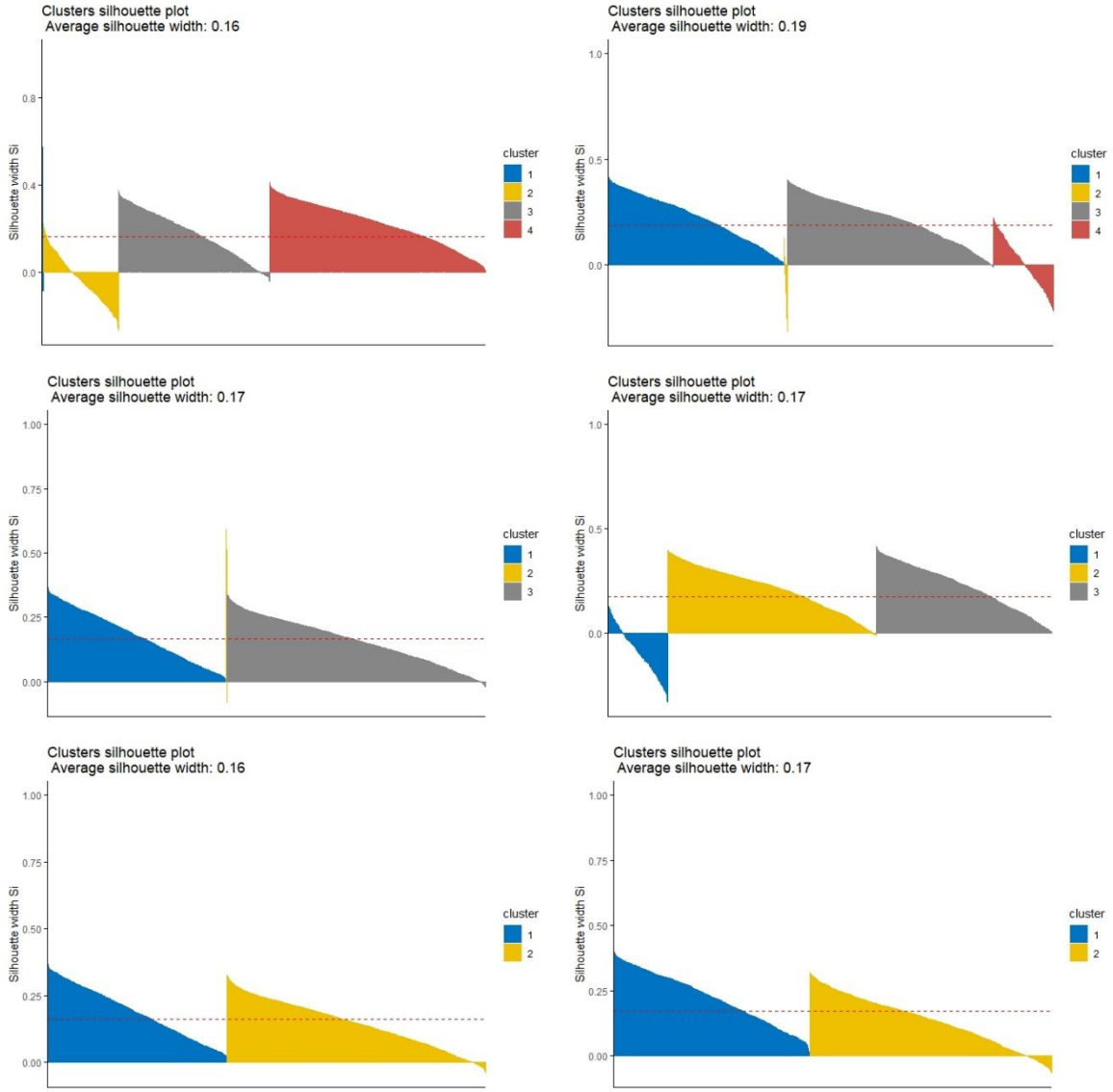


Figura 31 – Para cada sexo, Femenino y Masculino, variantes de cantidad de clusters

En la Figura 31, se muestran los resultados de ejecución para cada sexo, en la columna de la izquierda el Femenino y en el de la derecha el Masculino. En la primera fila con la cantidad de clúster propuestos, se visualiza una gran existencia de observaciones que entran en conflicto con otros clústeres ya que están por debajo de cero. Ya para la segunda fila, con dos clústeres podemos observar cómo se redistribuyen las observaciones dentro de los

clústeres, pero sigue habiendo una cantidad considerable de conflictos. En la tercera fila, con dos clústeres, vemos que el algoritmo logra acomodar las observaciones, pero siguen quedando conflictos entre los clústeres.

Analicémoslo desde los números (Tabla 14), si observamos primero el sexo femenino, es interesante ver la formación de un clúster con pocas observaciones (8) con un coeficiente de 0.36 el cual indica que hay valores de Silhouette más altos y estos están bien diferenciados de otros clústeres y se ven reflejados en la corrida de 4 y 3 clúster.

En el masculino hay un clúster con un promedio de valores negativo (-0.09), tanto para la corrida con 4 clústeres como para 3 clúster, lo que representa que la mayoría de sus valores están en conflicto. Otra observación en la generación de 4 clúster es el número 4 que tiene un promedio de 0.01 para sus observaciones, lo que nos indica que prácticamente la mitad de sus valores están en negativo y la otra mitad en positivo.

Para la última corrida de 2 clústeres las cantidades se distribuyen prácticamente uniformemente y los valores promedios son similares en los dos sexos.

id	cluster	size	ave.sil.width
1	1	8	0,36
2	2	473	-0,03
3	3	947	0,18
4	4	1356	0,22

id	cluster	size	ave.sil.width
1	1	1011	0,22
2	2	18	-0,09
3	3	1184	0,21
4	4	344	0,01

id	cluster	size	ave.sil.width
1	1	1134	0,18
2	2	8	0,36
3	3	1642	0,16

id	cluster	size	ave.sil.width
1	1	344	-0,09
2	2	1201	0,21
3	3	1012	0,22

id	cluster	size	ave.sil.width
1	1	1142	0,19
2	2	1642	0,14

id	cluster	size	ave.sil.width
1	1	1146	0,22
2	2	1411	0,13

Tabla 14 – Valores correspondiente a cada corrida para diferente cantidad de clústeres

Observemos, ahora como se representan los valores en el plano, bajo el mismo método de PCA (Componentes Principales). En la Figura 32, vemos primero la representación para el sexo Femenino y luego para el sexo Masculino y en los dos la gráfica de los valores en conflicto entre los dos clústeres.

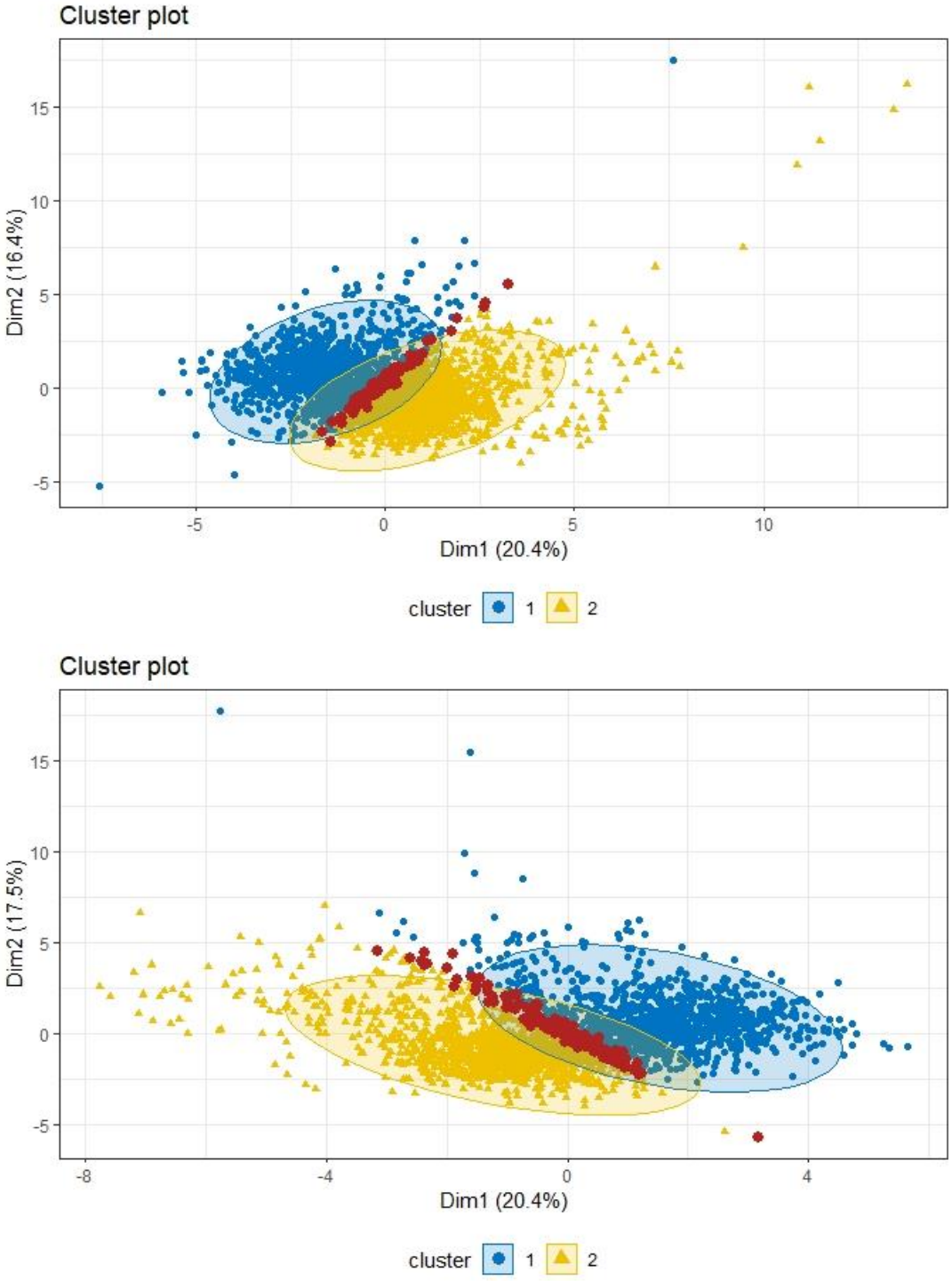


Figura 32 – Plot Femenino y Masculino de 2 clúster para el grupo etario 2

El tercer grupo etario **"Adultos"**, para ambos sexos, la propuesta era de 3 (tres) clústeres para el sexo femenino y de 2 (dos) para el masculino.

Comenzamos analizando la calidad de los clústeres para el sexo femenino, donde en la Figura 33 vemos que, para tres clústeres, el primero debe tener un valor promedio muy próximo a cero, lo validaremos luego con la tabla de valores para cada figura. Ya en la segunda figura con dos clústeres mejoran la cantidad de los valores negativos.

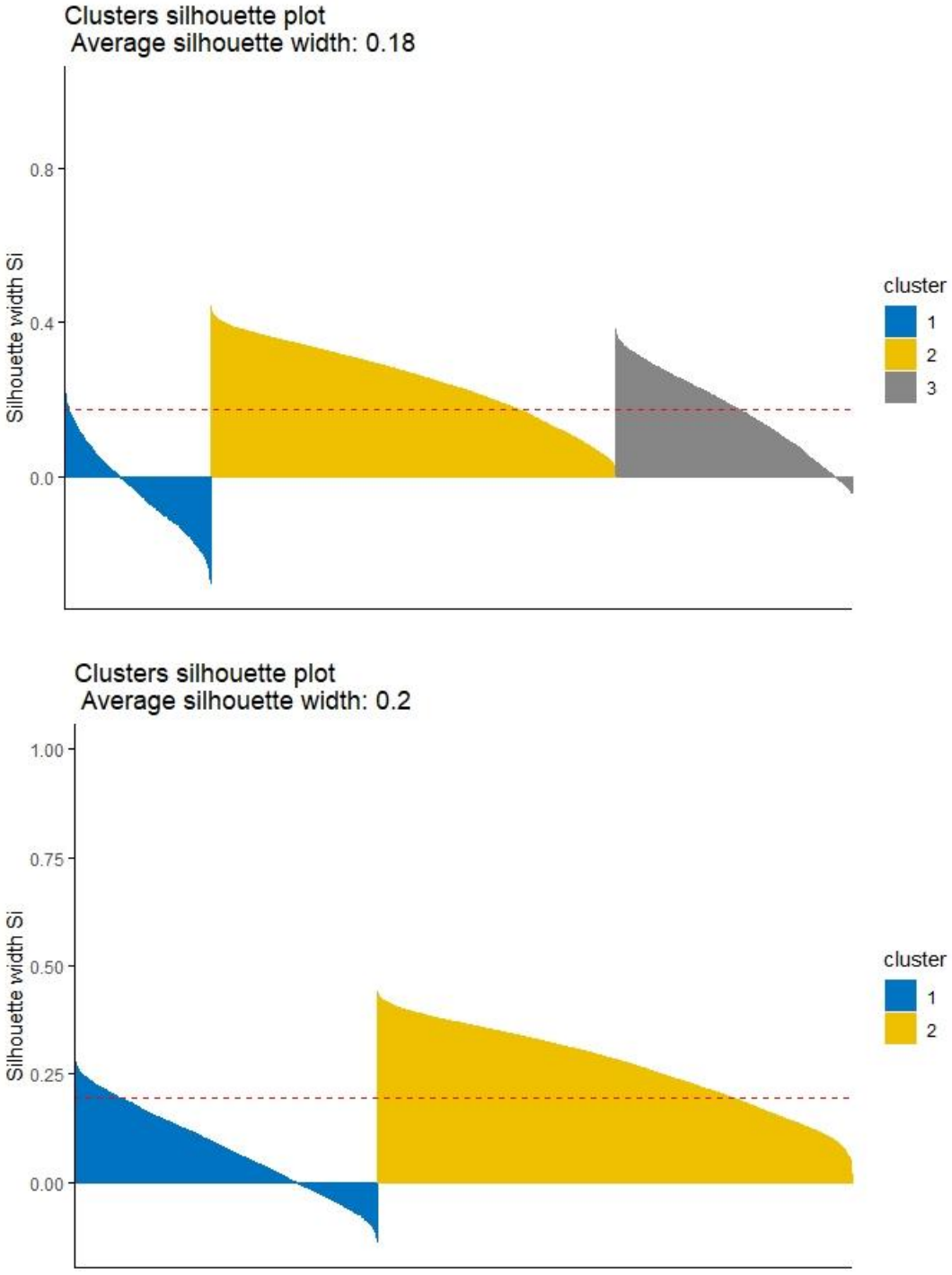


Figura 33 – Análisis Silhouette para 3 y 2 clústeres Femenino

Para el sexo masculino, Figura 34, la distribución es muy similar a la femenina.

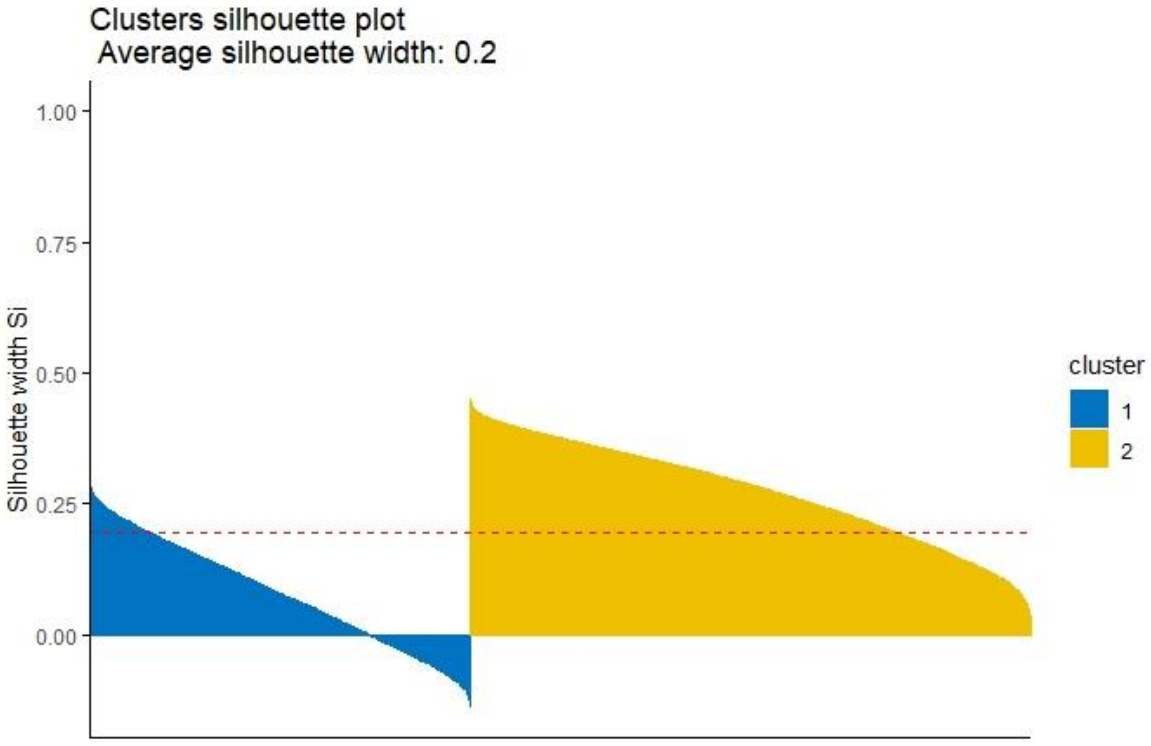


Figura 34 – Análisis Silhouette para 3 y 2 clústeres Masculino

Si vemos los valores obtenidos, para los dos sexos en la Tabla 15, confirmamos primero lo dicho para el sexo femenino cuando se generaron tres clústeres, donde el clúster 1 (uno) tiene un promedio de valor de Silhouette de -0.03. Y en la segunda corrida, para 2 clúster, donde las imágenes de los dos sexos son muy similares, lo confirmamos con los números, donde lo único que varía es la cantidad de observaciones para cada uno, pero sus promedios son iguales para cada clúster.

id	cluster	size	ave.sil.width
1	1	3059	-0,03
2	2	8497	0,25
3	3	4982	0,17

id	cluster	size	ave.sil.width
1	1	6431	0,08
2	2	10107	0,27

id	cluster	size	ave.sil.width
1	1	5531	0,08
2	2	8146	0,27

Tabla 16 – Valores correspondiente a cada corrida para diferente cantidad de clústeres

Veamos en la Figura 15 su distribución en un plano de los valores para cada clúster y los valores en conflicto.

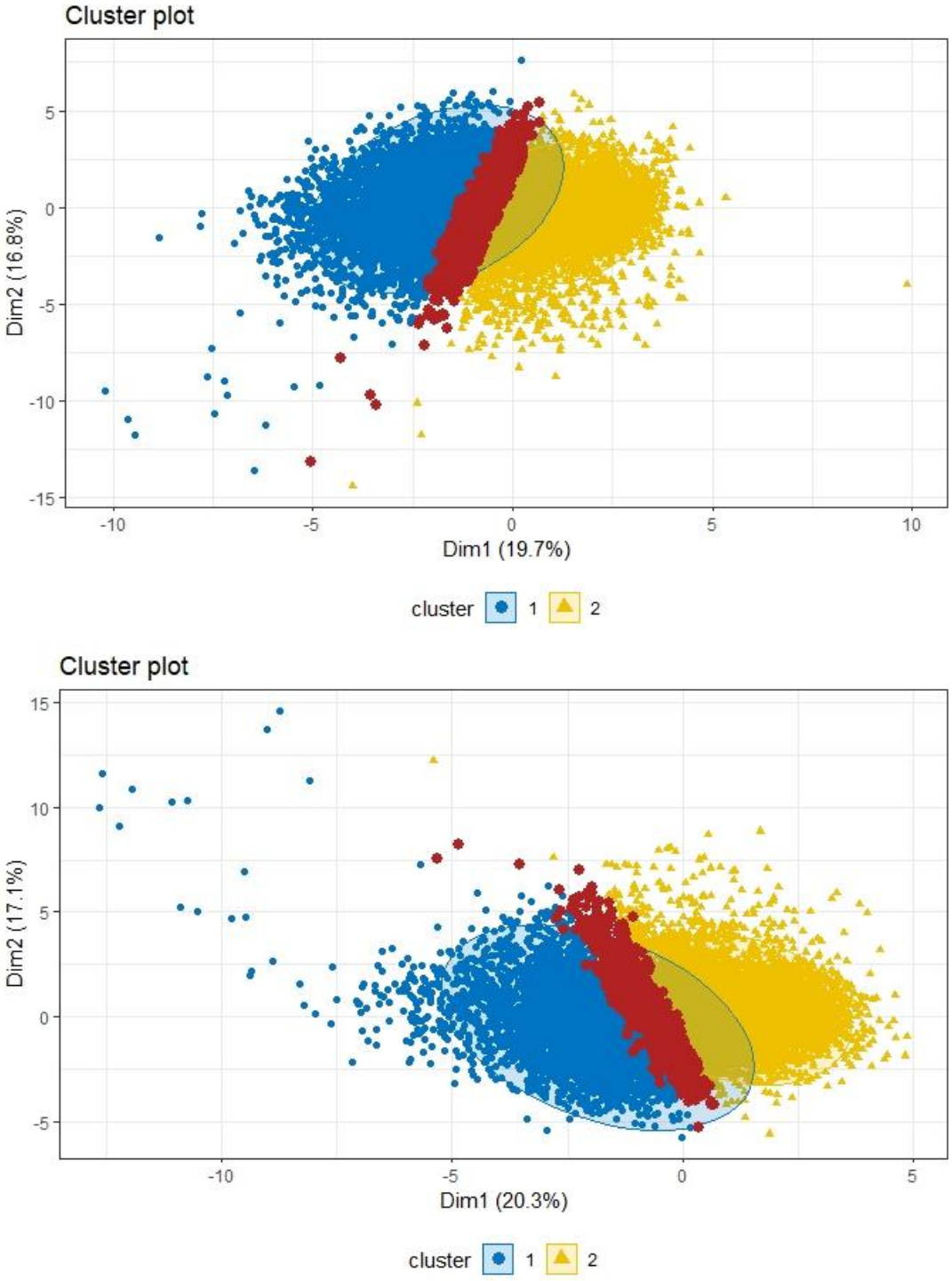


Figura 35 – Plot Femenino y Masculino de 2 clúster para el grupo etario 3

Por ultimo nos queda el cuarto grupo etario, denominado **"Adulto Mayor"**, donde la propuesta de cantidad de clústeres para cada sexo era de 2 (dos).

Comencemos primero analizando, en la Figura 35, la formación de los clústeres para los dos sexos. Podemos observar que la distribución de los valores en los clústeres es similar en ambos sexos a diferencia que se invierten los clústeres, en el femenino es similar el clúster 1 al masculino clúster 2 y el dos femenino al 1 masculino.

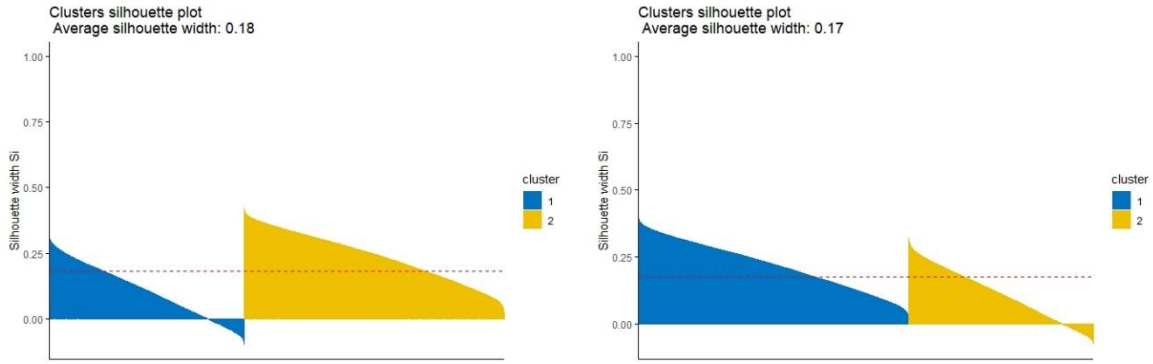


Figura 18 – Valores Silhouette para sexo femenino y masculino

En valores, Tabla 16, confirmamos lo observado en la Figura 16, donde los valores son similares en clúster invertidos.

id	cluster	size	ave.sil.width
1	1	15781	0,11
2	2	21109	0,24

id	cluster	size	ave.sil.width
1	1	14347	0,22
2	2	9785	0,11

Tabla 16 – Valores para cada sexo del grupo etario Adulto Mayor

Veamos en la Figura 36 su distribución en un plano de los valores para cada clúster y los valores en conflicto.

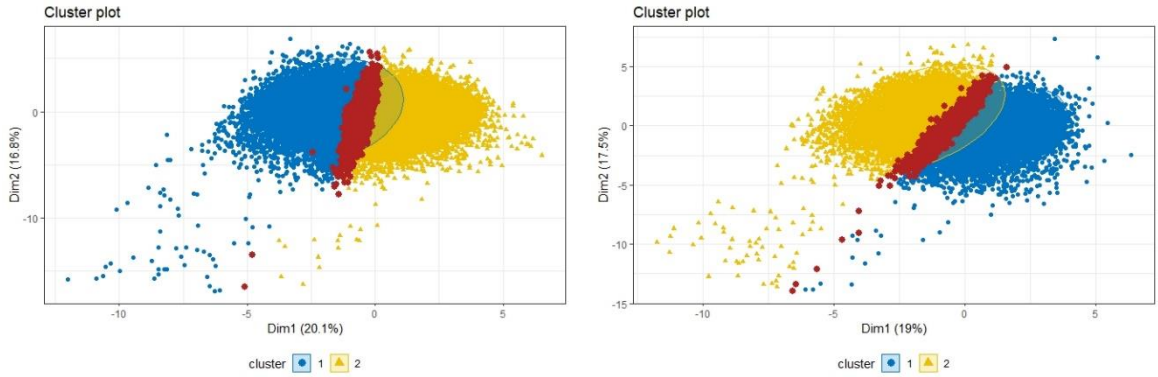


Figura 36 – Plot para grupo etario Adulto Mayor

8.7.4. Nube de palabras

Para identificar la relación de las variables y sus clústeres correspondientes se realizó un análisis de los diagnósticos de los pacientes mediante un análisis de palabras y procesamiento de lenguaje natural.[9]

En primer lugar, se analizaron todos los diagnósticos y se generó un diagnostico abreviado, el cual agrupa varios de los diagnósticos originales. De ellos se realizó por sexo, por cada grupo etario y para cada clúster una nube de palabras, a fin de poder analizar los grupos generados y las patologías entre ambos y determinar si sus agrupaciones tienen algún sentido sobre los diagnósticos y clúster.

Comenzamos por el primer grupo etario “Infancia”.

Figura 37 – Nube de palabras

Para el sexo femenino, Figura 37, podemos observar que los diagnósticos prevalentes son, para el primer clúster son vómitos, agudo, abdomen y fiebre, mientras que para el segundo clúster dolor, convulsión e infección.

Figura 38 – Nube de palabras

Mientras que, para el sexo masculino, Figura 38, podemos observar que los diagnósticos prevalentes son, para el primer clúster son fiebre, vómitos, broncoespasmos y neumonía, mientras que para el segundo clúster control, abdominal y hepático.

Segundo grupo etario “Adolescencia y Juventud”

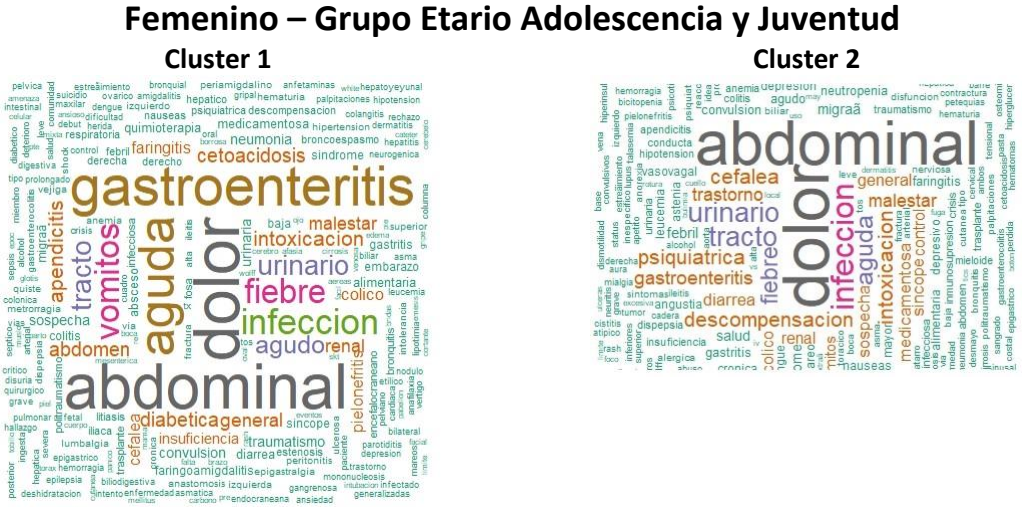


Figura 39 – Nube de palabras

En este grupo, en el sexo femenino, Figura 39, podemos observar que los diagnósticos prevalentes son, para el primer clúster son dolor, abdominal, gastroenteritis y aguda, mientras que para el segundo clúster dolor, abdominal e infección.

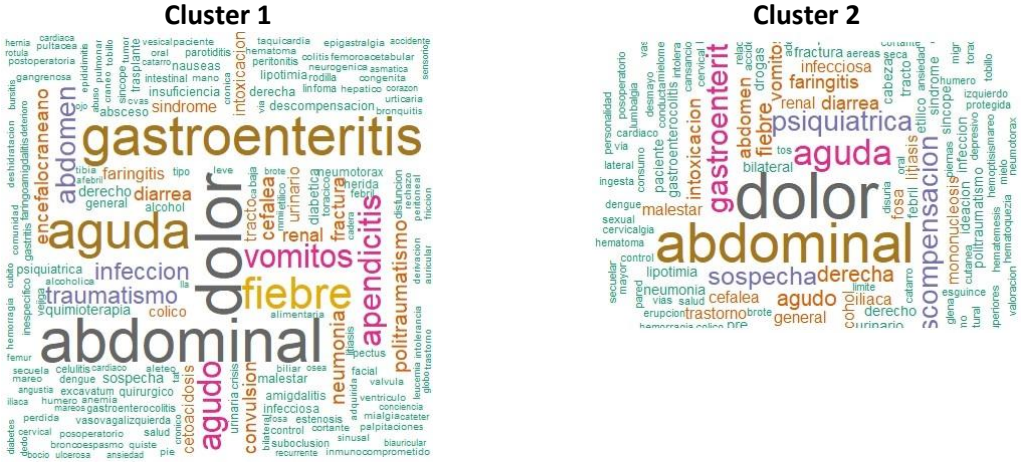


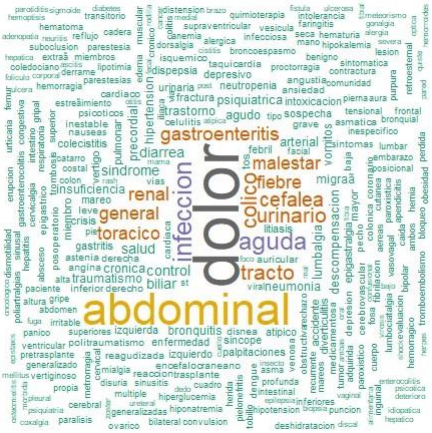
Figura 40 – Nube de palabras

En el sexo masculino, Figura 40, podemos observar que los diagnósticos prevalentes son, para el primer clúster son dolor, abdominal, gastroenteritis y aguda, mientras que para el segundo clúster dolor y abdominal.

Tercer grupo etario “Adulto”

Femenino – Grupo Etario Adulto

Cluster 1



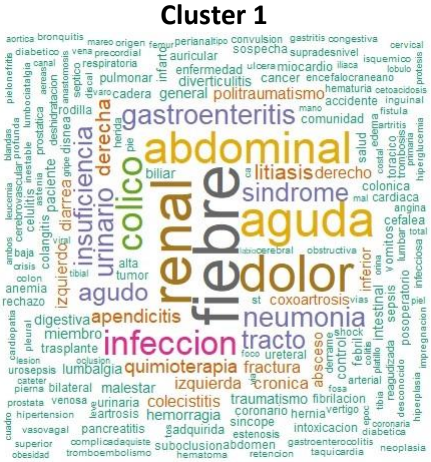
Cluster 2

Figura 41 – Nube de palabras

El sexo femenino, Figura 41, podemos observar que los diagnósticos prevalentes son, para el primer clúster son dolor, abdominal, infección y aguda, mientras que para el segundo clúster dolor, abdominal, infección, fiebre, aguda y urinario.

Masculino – Grupo Etario Adulto

Cluster 1



Cluster 2

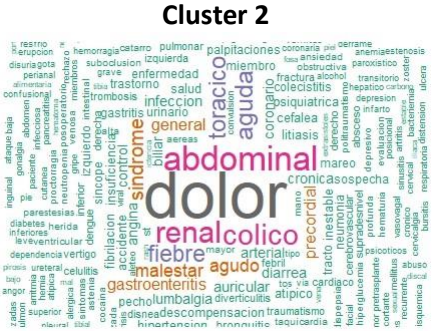


Figura 42 – Nube de palabras

En tanto por el sexo masculino, Figura 42, podemos observar que los diagnósticos prevalentes son, para el primer clúster son fiebre, renal, abdominal, dolor y aguda, mientras que para el segundo clúster dolor, renal, cólico y abdominal.

Cuarto y último grupo etario “Adulto Mayor”

Figura 43 – Nube de palabras

En este grupo de personas mayores y donde las patologías son mayores que en los grupos anteriores, en el sexo femenino, Figura 43, podemos observar que los diagnósticos prevalentes son, para el primer clúster son dolor, insuficiencia y cardiaca, mientras que para el segundo clúster fractura, insuficiencia, neumonía, cadera y cardiaca.

Figura 44 – Nube de palabras

En el sexo masculino, Figura 44, podemos observar que los diagnósticos prevalentes son, para el primer clúster son insuficiencia, cardiaca y dolor, mientras que para el segundo clúster neumonía, insuficiencia, fiebre e infección.

8.7.5. Boxplots

Se utiliza otro método para identificar la relación de las variables y sus clústeres correspondientes, se realizó un análisis descriptivo mediante la generación de Boxplots o también conocidos como Diagrama de caja o Diagrama de caja y bigote. Un diagrama de caja es una forma estandarizada de mostrar la distribución de datos basada en un resumen de cinco números ("mínimo", primer cuartil (Q1), mediana, tercer cuartil (Q3) y "máximo"). Puede informarle sobre sus valores atípicos y cuáles son sus valores. También puede indicarle si sus datos son simétricos, qué tan estrechamente están agrupados sus datos y si sus datos están sesgados y de qué manera. [10].

De las 18 (dieciocho) variables analizadas para cada sexo y grupo etario, documentaremos solo aquellas por la que consideramos que varían entre cada variable categórica y dan origen a la clusterización.

La primera variable corresponde a los **"Neutrófilos Segmentados",** qué son, valores normales y cuándo se alteran. Los **neutrófilos** son un tipo de leucocito o de célula blanca responsables por la defensa del organismo, los cuales aumentan en la sangre cuando hay alguna infección o algún proceso inflamatorio. Su medición, para las muestras observadas del hospital, es de 0 a 100 ya que la misma mide en porcentaje y sus valores normales van entre el 40 y el 60. En el sexo femenino, Figura 45, se observa que las observaciones entre primer cuartil (Q1), mediana y tercer cuartil (Q3) se encuentran entre los valores normales para el **clúster 2**, mientras que para el **clúster 1** están por encima de la máxima.

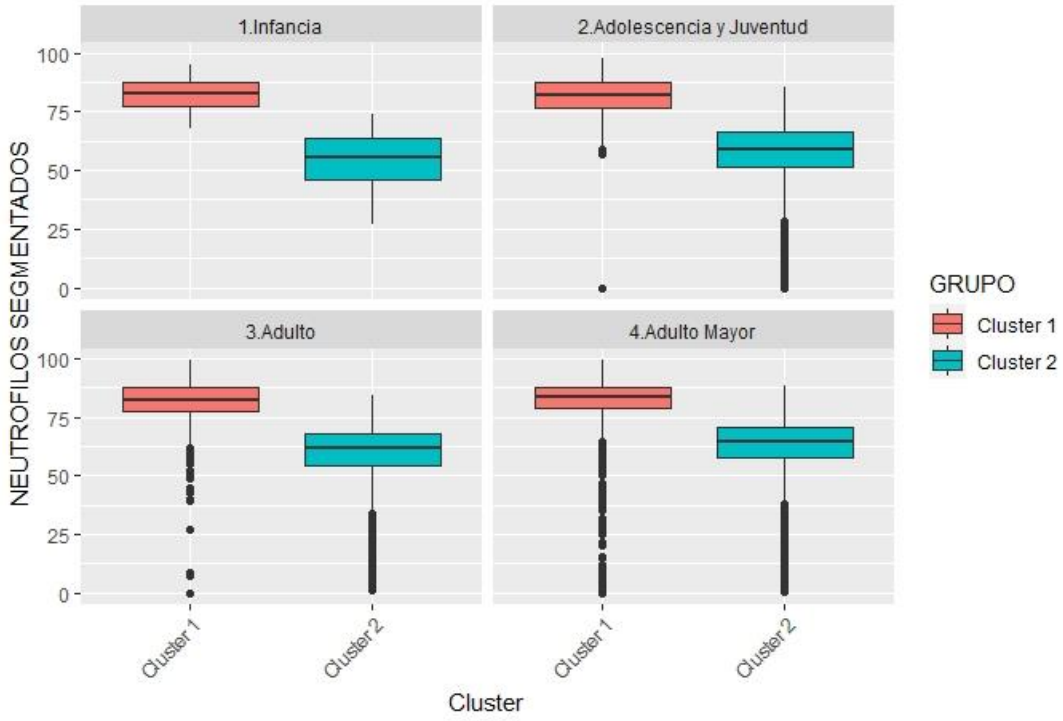


Figura 45 – Boxplots sexo Femenino

En el sexo masculino (Figura 46) es igual para sus tres primeros grupos etarios, pero se invierte en el cuarto.

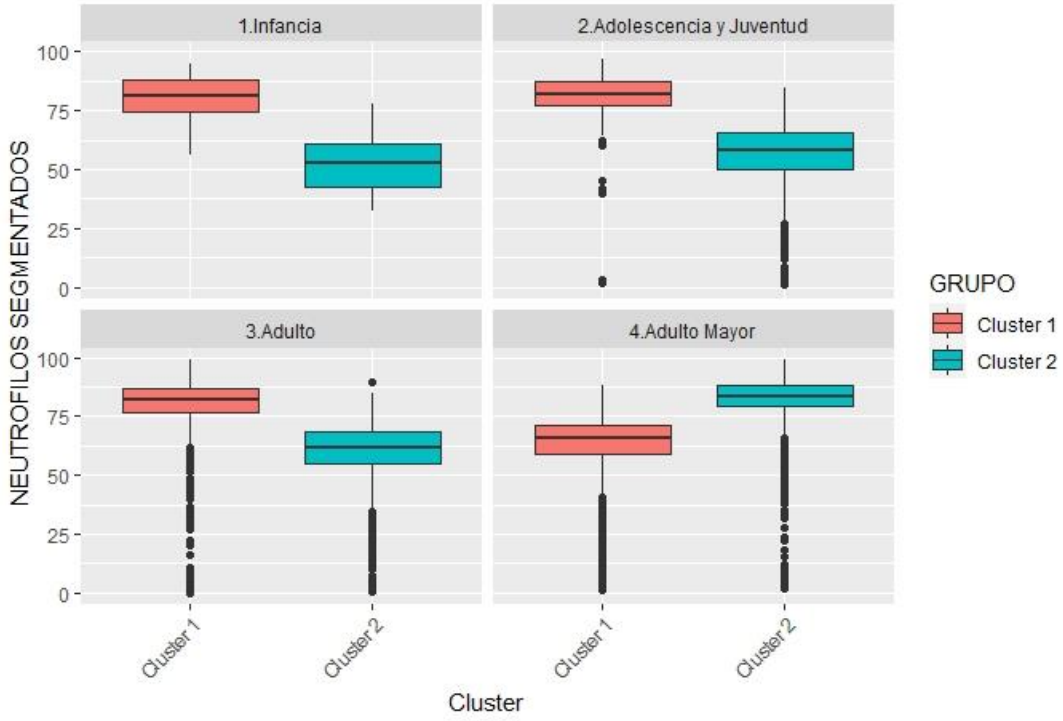


Figura 46 – Boxplots sexo Masculino

La segunda variable, **"Linfocitos"**, una clase de glóbulos blancos que se eleva sobre todo en infecciones víricas y que producen anticuerpos. A igual que la variable anterior, su medición, para las muestras observadas del hospital, es de 0 a 100 ya que la misma mide en porcentaje y sus valores normales van entre el 20 y 40. Para el sexo femenino, Figura 47, el **clúster 2** se mantiene dentro de los valores normales y el **clúster 1** por debajo de ellas. Mientras que, en el sexo masculino, Figura 48, para los grupos etarios de Infancia, Adolescencia y Juventud y Adultos también se mantiene el **clúster 2** dentro de los parámetros normales y el **clúster 1** por debajo de ellos y en el grupo etario Adultos Mayores es el **clúster 1** el que se encuentra dentro de los parámetros normales y el **clúster 2** por debajo de ellos.

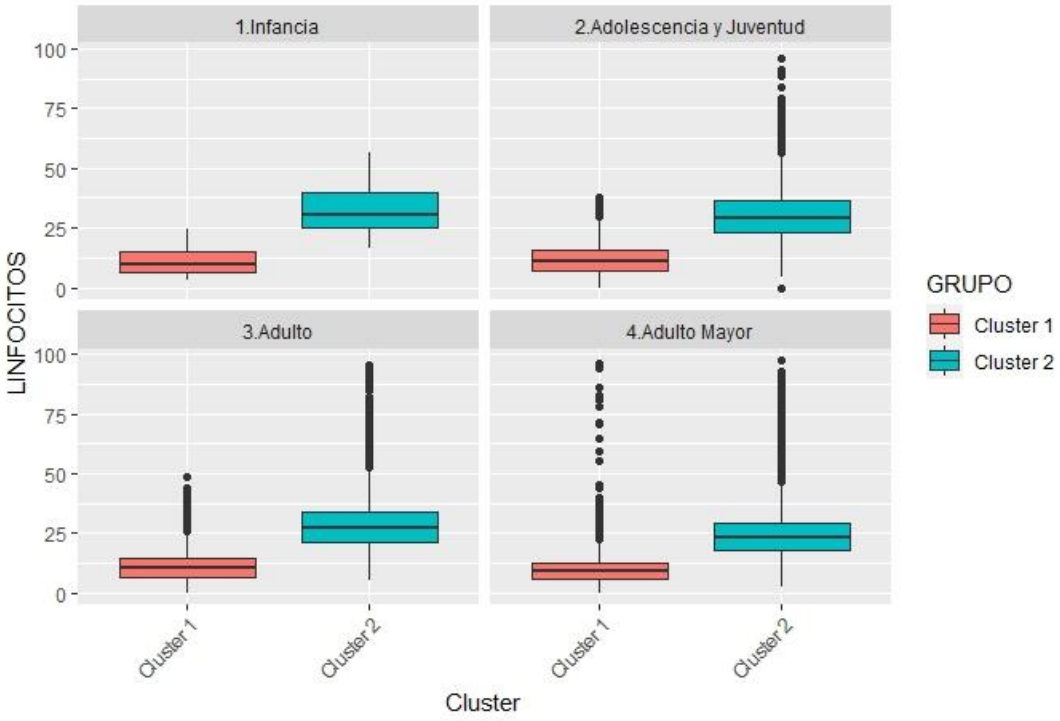


Figura 47 – Boxplots sexo Femenino

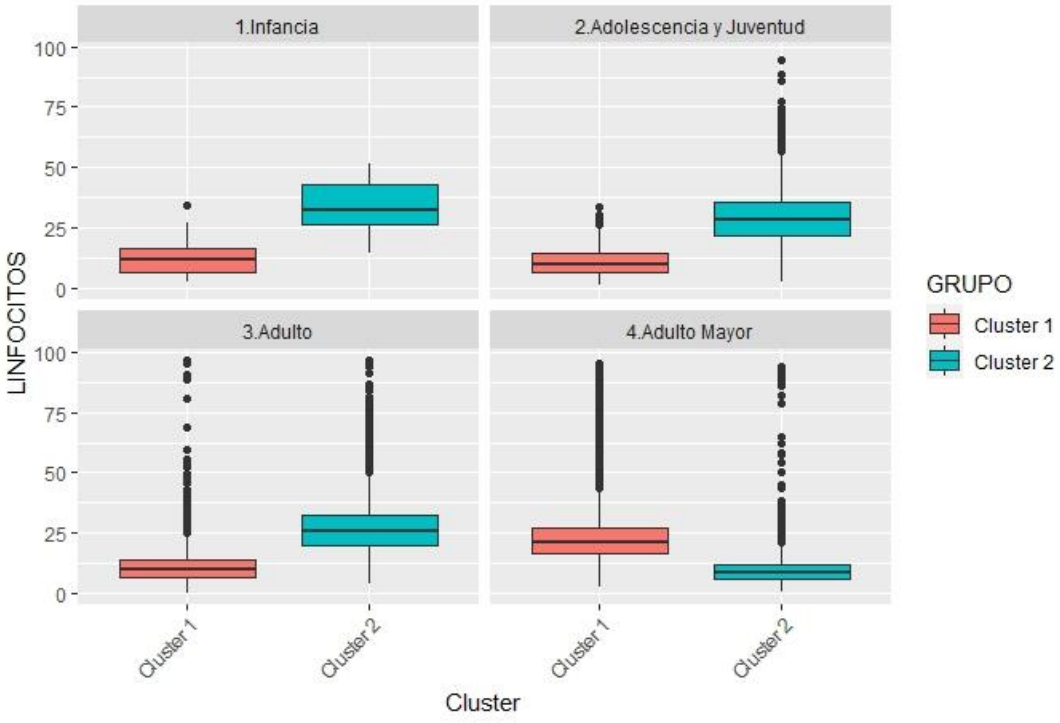


Figura 48 – Boxplots sexo Masculino

La tercera variable, **“Monocitos”**, una clase de glóbulos blancos que se producen en la médula ósea y luego entran en el torrente sanguíneo y luchan contra determinadas infecciones y ayudan a otros leucocitos a eliminar tejidos muertos o dañados, destruir células cancerosas y regular la inmunidad contra sustancias extrañas. A igual que la variable anterior, su medición, para las muestras observadas del hospital, es de 0 a 100 ya que la misma mide en porcentaje y sus valores normales van entre el 2 y 8.

Para el sexo femenino, Figura 49, el **clúster 1** se mantiene dentro de los valores normales y el **clúster 2** por encima del primer cuartil (Q1) y haciendo base en el límite superior de los valores normales para la variable.

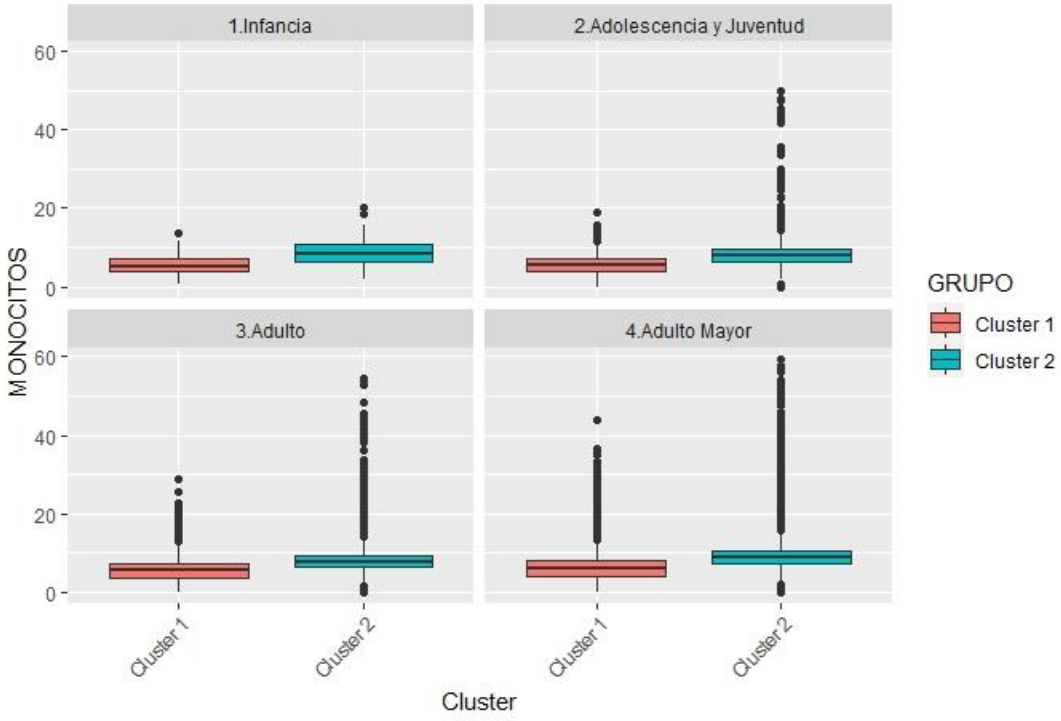


Figura 49 – Boxplots sexo Femenino

Mientras que, en el sexo masculino, Figura 50, para los grupos etarios de Infancia, Adolescencia y Juventud y Adultos también se mantiene el **clúster 1** dentro de los parámetros normales y el **clúster 2** por encima de ellos haciendo base, al igual que el sexo femenino, en el límite superior y en el grupo etario Adultos Mayores es el **clúster 2** el que se encuentra dentro de los parámetros normales y el **clúster 1** por encima de ellos.

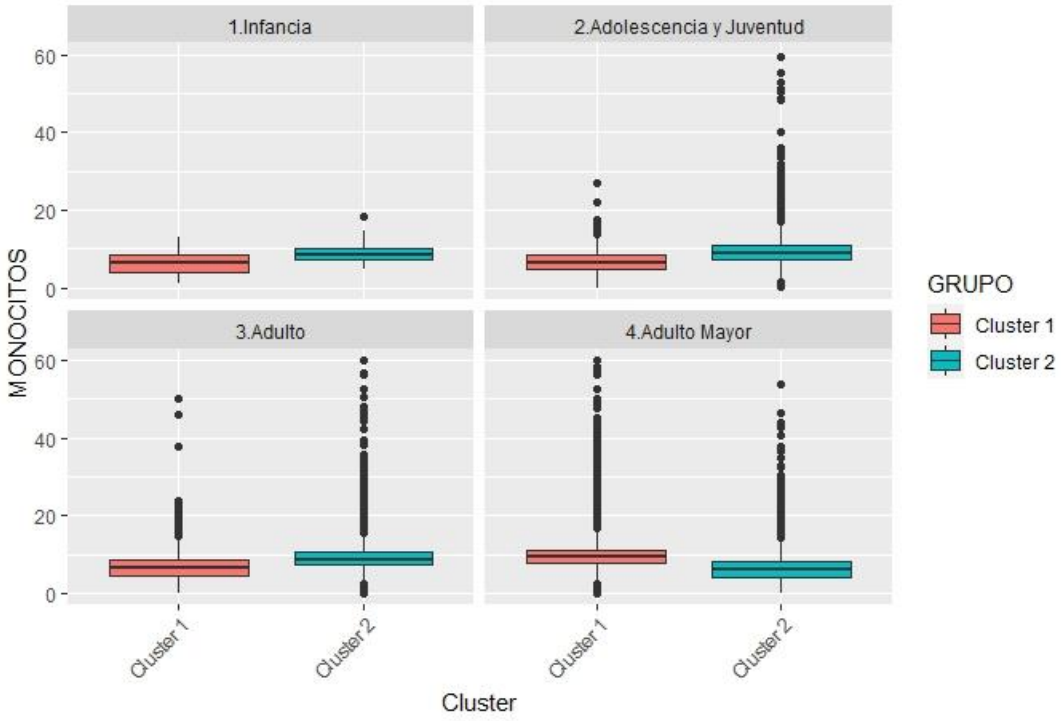


Figura 50 – Boxplots sexo Masculino

Todas las variables inciden en la formación del clúster, unas en mayor medida que otras.

8.8. Componentes Principales

Principal Component Analysis (PCA) [11] es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información. Supóngase que existe una muestra con n individuos cada uno con p variables (X1, X2, ..., Xp), es decir, el espacio muestral tiene p dimensiones. PCA permite encontrar un número de factores subyacentes ($z < p$) que explican aproximadamente lo mismo que las p variables originales. Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores. Cada una de estas z nuevas variables recibe el nombre de componente principal.

El método de PCA permite por lo tanto "condensar" la información aportada por múltiples variables en solo unas pocas componentes. Esto lo convierte en un método muy útil de aplicar previa utilización de otras técnicas estadísticas tales como regresión, clustering... Aun así, no hay que olvidar que sigue siendo necesario disponer del valor de las variables originales para calcular las componentes.

Los componentes principales son las combinaciones lineales de las variables originales que explican la varianza en los datos. El número máximo de componentes extraídos siempre es igual al número de variables. Los vectores propios, compuestos por los coeficientes que corresponden a cada variable, se utilizan para calcular las puntuaciones de

los componentes principales. Los coeficientes indican la ponderación relativa de cada variable en el componente.

S - Covarianza	MUJERES								
Importance of components:	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1814,14390	873,70240	437,70664	128,90310	11,51108	4,25800	1,22700	0,33980	0,20040
Proportion of Variance	0,77210	0,17910	0,04494	0,00390	0,00003	0,00000	0,00000	0,00000	0,00000
Cumulative Proportion	0,77210	0,95110	0,99607	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
Importance of components:	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
Standard deviation	0,18630	0,15000	0,12310	0,05826	0,02244	0,01718	0,01463	0,01181	0,00016
Proportion of Variance	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
Cumulative Proportion	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000

R - Correlación	MUJERES								
Importance of components:	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1,89110	1,73980	1,61270	1,58800	1,00061	0,92774	0,91490	0,89432	0,82324
Proportion of Variance	0,19870	0,16820	0,14450	0,14010	0,05562	0,04782	0,04650	0,04443	0,03765
Cumulative Proportion	0,19870	0,36690	0,51130	0,65140	0,70706	0,75488	0,80140	0,84581	0,88346
Importance of components:	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
Standard deviation	0,78821	0,70985	0,69371	0,54897	0,38070	0,14837	0,13673	0,05277	0,03856
Proportion of Variance	0,03452	0,02799	0,02673	0,01674	0,00805	0,00122	0,00104	0,00015	0,00008
Cumulative Proportion	0,91798	0,94597	0,97271	0,98945	0,99750	0,99872	0,99976	0,99992	1,00000

Tabla 17 – PCA por Covarianza y por Correlación

Primero analizaremos el resultado del PCA utilizando la matriz S (covarianzas).

Como podemos observar el primer componente tiene valores muy altos con respecto al resto. Esto lo podemos observar en la “Desviación Estándar” de cada una de las variables principales.

Lo que produce que la variable principal con mayor desviación concentre la mayor proporción.

Como podemos observar en la Tabla 17, el PC1, con un valor muy elevado al resto concentra el 0,77210 de proporción, prácticamente todo el peso y la PC2 el 0,17910 lo cual contra las restantes su desvío estándar es muy elevado, descartando del análisis el resto de las variables. Esto puede producir un análisis erróneo.

A diferencia de la utilización de la matriz de covarianzas, la cual como vimos mantiene la diferencia que existe entre las unidades de medida de las variables, se puede utilizar la matriz de correlación (R), la cual genera un rango uniforme entre todas las variables, de -1 a 1.

Rápidamente podemos observar como la desviación estándar de las variables principales son más homogéneas, máx. 1,90, min. 0,03.

Siendo estas SD más parejas la proporción de variación es más distribuida, recién alcanza el 0,95 en la PC11.

Para el análisis de cuantas variables principales debemos tomar, utilizaremos un gráfico lineal representando la proporción de variación de cada una. Ver Figura 51 para la de Covarianza y 52 para la de Correlación.

Podemos observar en la de Covarianza que el codo se produce ya en la PC3 y luego prácticamente se mantiene en una curva recta.

Proporción de las varianzas explicadas

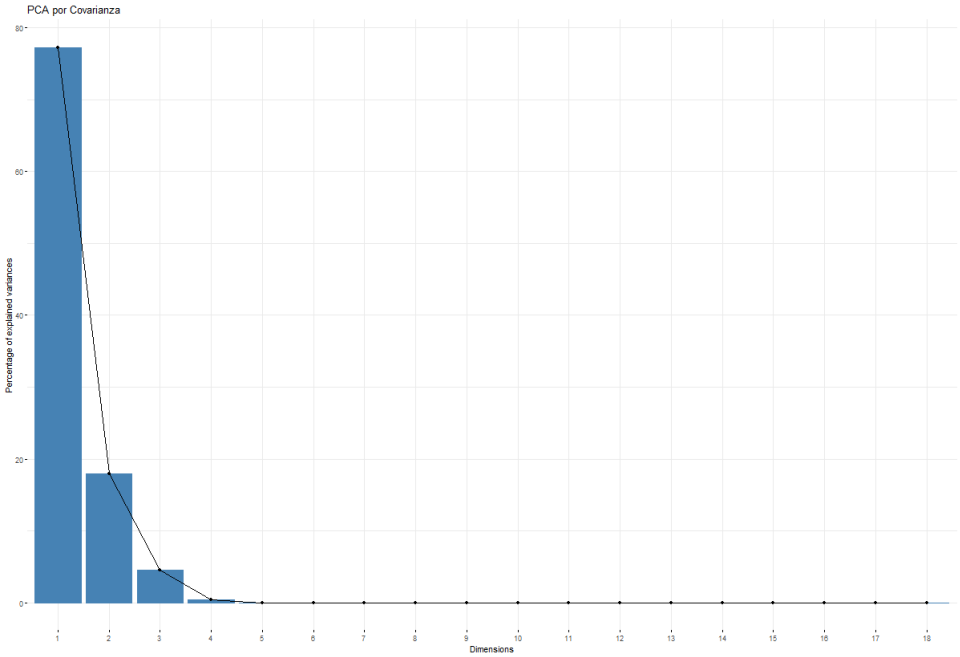


Figura 51 – Matriz de Covarianza

Mientras que en la de Correlación recién alcanza el codo en el PC5 y luego un segundo codo en el 14.

Proporción de las varianzas explicadas

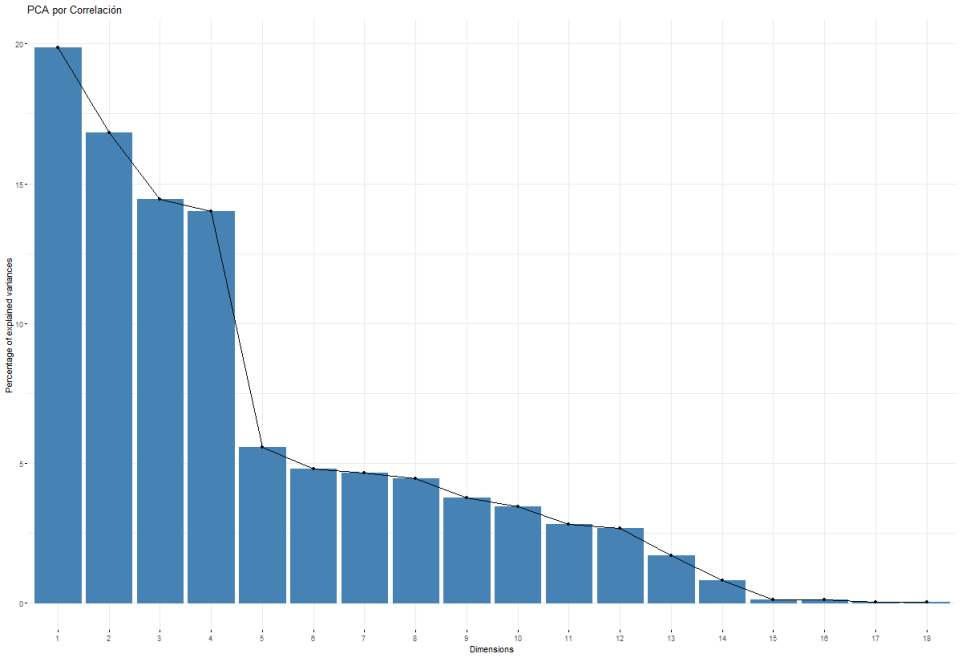


Figura 52 – Matriz de Correlación

Para el sexo Masculino sucede lo mismo que para el Femenino, ya que las variables a evaluar son las mismas.

S - Covarianza	HOMBRES								
Importance of components:	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1903,08530	1274,68260	626,04359	133,49029	12,40089	4,52200	1,61600	0,33060	0,20820
Proportion of Variance	0,64030	0,28730	0,06929	0,00315	0,00003	0,00000	0,00000	0,00000	0,00000
Cumulative Proportion	0,64030	0,92750	0,99682	0,99997	1,00000	1,00000	1,00000	1,00000	1,00000
Importance of components:	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
Standard deviation	0,17520	0,13870	0,10290	0,06583	0,02512	0,02098	0,01396	0,00031	0,00016
Proportion of Variance	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
Cumulative Proportion	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000

R - Correlación	HOMBRES								
Importance of components:	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1,88110	1,76660	1,60930	1,56110	1,00243	0,95491	0,93824	0,88104	0,83018
Proportion of Variance	0,19660	0,17340	0,14390	0,13540	0,05583	0,05066	0,04891	0,04312	0,03829
Cumulative Proportion	0,19660	0,37000	0,51380	0,64920	0,70505	0,75571	0,80462	0,84774	0,88603
Importance of components:	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
Standard deviation	0,78975	0,71654	0,69468	0,54510	0,29731	0,15176	0,14002	0,04637	0,03859
Proportion of Variance	0,03465	0,02852	0,02681	0,01650	0,00491	0,00128	0,00109	0,00012	0,00008
Cumulative Proportion	0,92068	0,94920	0,97601	0,99250	0,99743	0,99871	0,99980	0,99992	1,00000

Tabla 18 – PCA por Covarianza y por Correlación

En la Tabla 18, para la matriz de covarianza S el PC1, concentra el 0,64030 de proporción, y la PC2 el 0,28730, prácticamente alcanzando el 0.95 solo entre estos dos componentes, lo cual contra las restantes su desvío estándar es muy elevado, descartando del análisis el resto de las variables. Esto puede producir un análisis erróneo.

Mientras que para el de correlación R el PC1 es de 0.19660 y el PC2 del 0.17340, recién alcanzando el 0,95 en la PC11, siendo estas SD más parejas la proporción de variación es más distribuida.

Para el análisis de cuantas variables principales debemos tomar, utilizaremos un gráfico lineal representando la proporción de variación de cada una. Ver Figura 53 para la de Covarianza y 54 para la de Correlación.

En la de Covarianza el codo se produce ya en la PC3 y luego prácticamente se mantiene en una curva recta.

Proporción de las varianzas explicadas

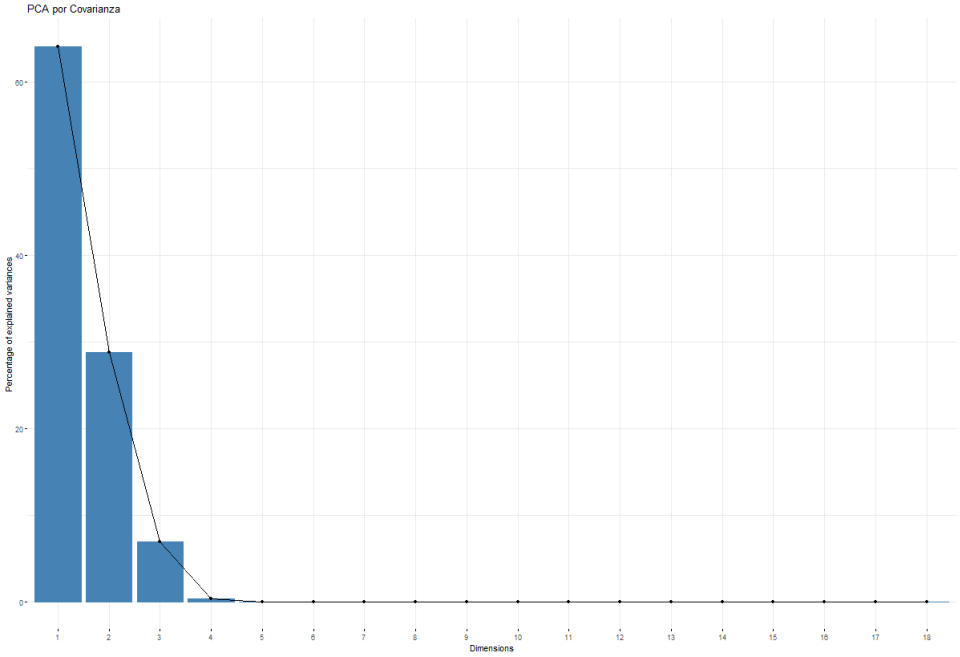


Figura 53 – Matriz de Covarianza

Y en la de Correlación realiza un primer codo en la PC5 y luego un segundo codo en el 14, igual que para el sexo femenino.

Proporción de las varianzas explicadas

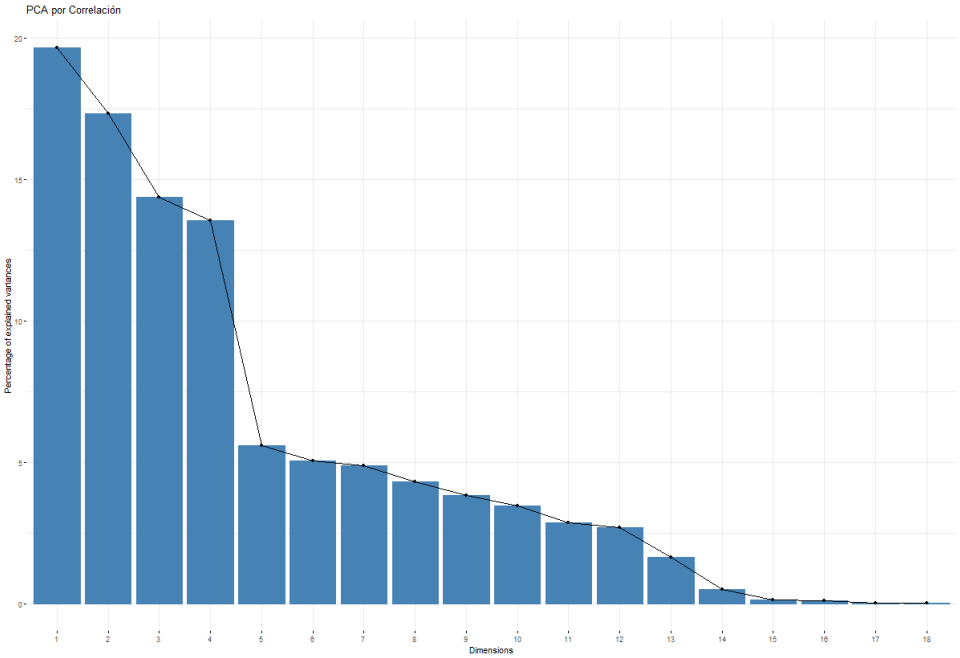


Figura 54 – Matriz de Correlación

Luego del análisis realizado en el ítem anterior, concluimos que para este set de datos es mejor la utilización de la “Matriz de Correlación”, dado que la unidad de medida del set de datos no es homogénea.

Analizaremos el grafico con la resultante de dicha matriz para los dos sexos, femenino y masculino.

Primero analizaremos el grafico resultante de la matriz femenina (figura 55). Sobre el eje X ubicamos el primer componente principal PC1 y sobre el eje Y el segundo componente principal PC2, cuyo peso es 19.9% y 16.8% respectivamente.

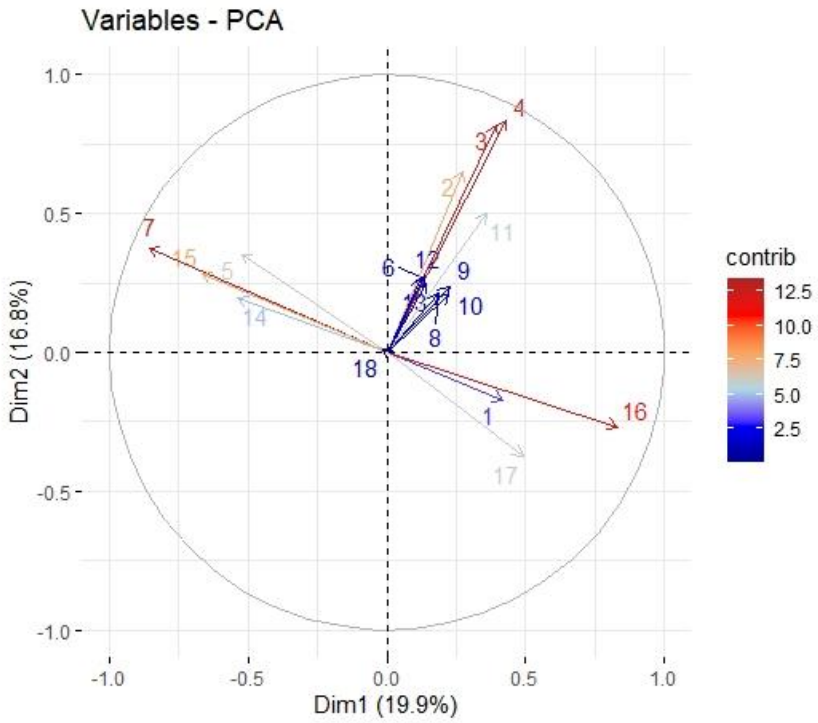


Figura 55 - Fuerza de las variables sobre las dos primeras componentes principales

El análisis que hacemos sobre esta figura es que las dos primeras componentes principales representan un 0.3669 de proporción del total del desvío estándar.

Segundo las flechas representan a las 18 variables de análisis, donde el color representa el peso o fuerza que tiene dicha variable entre las dos componentes principales graficadas, a cuanta mayor fuerza su longitud es mayor en el plano, esta se obtiene con la suma de los pesos asignados a una variable, en valores absolutos, para cada componente principal.

Tabla 19 – Variables por cuadrante y fuerza

En la Tabla 19, se puede observar mejor, con los nombres de las variables cuales tienen mayor peso en cada cuadrante. Esto significa que los puntos que estén ubicados en ese cuadrante, equivalente a cada internación, muestran un claro patrón entre las patologías de esos pacientes y las variables que influencian en ese sector.

Como mencionamos anteriormente, los pesos asignados a cada variable, deben ser tomados en valores absolutos, para identificar su patrón. En la Tabla 20, los valores figuran como fueron calculados por el algoritmo utilizado, donde fueron ordenados por la suma de las dos primeras componentes en forma descendente y sus colores nos permite ubicarlos rápidamente a que cuadrante corresponde, por ej. si son los dos verdes, arriba a la derecha, si el primero es verde y el segundo rojo, abajo a la derecha, si el primero es rojo y el segundo verde, arriba a la izquierda y por último si son los dos rojos abajo a la izquierda.

FEMENINO									
Var.	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
4	0,2254761	0,4810571	-0,0312342	-0,1784583	-0,0000786	0,0573090	0,0232017	-0,1333541	0,0289548
3	0,2084952	0,4697957	-0,1048641	-0,1951166	0,0037658	0,0586664	-0,0512354	-0,2205862	0,0680613
7	-0,4530309	0,2158580	0,0361457	0,0578939	0,0107245	-0,1297447	-0,0366180	-0,0251193	-0,0703238
16	0,4403575	-0,1546096	-0,0506277	-0,0225876	0,0024449	0,1058383	0,0208831	0,1346784	0,2823565
2	0,1450535	0,3745813	-0,3593184	-0,2255608	-0,0034302	0,0626974	0,0981857	-0,0522650	0,0109315
15	-0,3525512	0,1609421	0,0795618	0,0081757	-0,0232080	0,1613823	0,2680279	-0,1616078	-0,1103762
5	-0,2758092	0,2032384	-0,0404562	-0,1290271	0,0113314	-0,3296980	-0,3085981	0,0551101	-0,3961603
11	0,1891141	0,2881121	0,1656007	-0,0275487	-0,0035542	-0,1293434	0,1258296	0,5359655	-0,2320672
17	0,2591816	-0,2163408	0,0059859	-0,0152580	-0,0044201	0,2325790	0,2633701	-0,4313684	-0,7112860
14	-0,2840267	0,1132473	0,0897743	0,0311476	-0,0411779	0,0658156	0,6268733	-0,1715863	0,3478312
1	0,2219597	-0,1003685	-0,0191814	-0,0089022	-0,0068184	-0,8663531	0,2905431	-0,2562597	0,0482267
9	0,1189794	0,1366231	0,5651683	0,0912623	0,0061164	0,0051831	-0,1294031	-0,1605854	0,0495480
10	0,1170463	0,1201201	0,3956504	0,0638980	-0,0200211	0,0208838	0,3800458	0,4081768	-0,1878542
12	0,0702285	0,1581692	-0,1524301	0,5434393	-0,0207411	-0,0106206	-0,0156083	-0,0328822	-0,0332152
6	0,0651145	0,1548110	-0,1476525	0,5268088	-0,0233461	-0,0132632	-0,0143679	-0,0192718	-0,0161007
8	0,0979861	0,1209531	0,5193784	0,0824004	0,0146648	-0,0038721	-0,3021432	-0,3468123	0,1278617
13	0,0745282	0,1428302	-0,1430029	0,5141805	0,0252191	0,0063827	-0,0092381	-0,0180691	-0,0449346
18	-0,0086541	0,0061988	-0,0007743	0,0122354	0,9975707	0,0057502	0,0525289	0,0009577	0,0066031
Var.	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
4	-0,0988321	-0,1014259	0,0046948	0,0021076	0,0045491	0,4397472	0,0594578	-0,6208234	-0,2450318
3	-0,0339249	-0,0069954	0,0313599	0,0029187	0,0056872	0,3006264	0,0552096	0,6490660	0,3323372
7	-0,2460023	0,0608965	0,3005353	0,0057435	-0,0028945	0,1054981	-0,7442145	0,0083539	-0,0009691
16	0,2276058	0,0125863	-0,4737277	-0,0049175	-0,0040588	0,0929319	-0,6200522	0,0087962	-0,0007906
2	-0,0984992	-0,0715155	0,0117780	0,0010691	-0,0075101	-0,7770288	-0,1087978	-0,0483811	-0,0962413
15	-0,2144598	0,4736508	-0,6547897	-0,0222658	-0,0002777	-0,0089939	0,0837816	-0,0015517	-0,0006806
5	0,4695906	-0,3645089	-0,3827497	0,0086307	0,0071701	-0,0022120	-0,0166288	0,0020217	0,0024177
11	0,3390703	0,5637205	0,2142184	-0,0092189	-0,0031038	-0,0217170	0,0118884	-0,0014274	0,0025541
17	0,1364481	0,0041661	0,1349985	-0,0175046	0,0072346	0,0261887	-0,1798998	0,0012390	0,0001836
14	0,5424356	-0,1847887	0,1385665	0,0207249	0,0033422	0,0018550	0,0044880	0,0004888	-0,0001297
1	-0,1528241	0,0980265	-0,0641566	0,0063507	0,0028386	-0,0025267	0,0072063	0,0011050	0,0005342
9	-0,0117593	-0,0638077	-0,0257691	0,0022062	-0,0011799	-0,2534869	-0,0438796	-0,3039028	0,6564785
10	-0,3609267	-0,4916735	-0,1302219	-0,0024047	-0,0052276	-0,0052520	-0,0024884	0,2162017	-0,1682296
12	0,0344232	-0,0308898	-0,0305263	-0,1585702	-0,7873131	0,0035163	0,0206790	0,0008327	0,0003582
6	0,0295221	-0,0364171	-0,0232246	-0,5987643	0,5562265	-0,0084696	0,0138218	-0,0004735	0,0003745
8	0,1290205	0,1172744	0,0220537	0,0027188	0,0024073	-0,1644658	-0,0216675	0,2273569	-0,6008067
13	0,0147463	-0,0076480	-0,0421072	0,7832240	0,2652293	-0,0059205	0,0153417	-0,0002253	-0,0003877
18	0,0066695	-0,0031859	-0,0092840	-0,0370963	-0,0101057	-0,0013291	0,0114635	0,0000644	-0,0001708

Tabla 20 – Fuerza de las variables por componente principal

Por ultimo si analizamos particularmente, los pesos de la componente principal 1 (PC1) y los ordenamos en forma descendente, podemos observar (Tabla 21) que la primera componente recoge mayoritariamente la información correspondiente a las variables que componen la serie blanca y luego a la roja. Para otros estudios de investigación queda pendiente la relación o patrón entre esas variables, ya que no se pudo contar con expertos en la materia.

#	Variable	Tipo	PC1	PC1 (abs)
7	NEUTROFILOS SEGMENTADOS	Serie blanca	-0,4530309	0,4530309
16	LINFOCITOS	Serie blanca	0,4403575	0,4403575
15	EOSINOFILOS	Serie blanca	-0,3525512	0,3525512
14	BASOFILOS	Serie blanca	-0,2840267	0,2840267
5	LEUCOCITOS RECUENTO	Serie blanca	-0,2758092	0,2758092
17	MONOCITOS	Serie blanca	0,2591816	0,2591816
4	HEMOGLOBINA	Serie roja	0,2254761	0,2254761
1	GLUCOSA	Bioquímica básica	0,2219597	0,2219597
3	HEMATOCRITO	Serie roja	0,2084952	0,2084952
11	RDW	Serie roja	0,1891141	0,1891141
2	HEMATIES RECUENTO	Serie roja	0,1450535	0,1450535
9	HCM	Serie roja	0,1189794	0,1189794
10	CHCM	Serie roja	0,1170463	0,1170463
8	VCM	Serie roja	0,0979861	0,0979861
13	NEUTROFILOS EN CAYADO	Serie blanca	0,0745282	0,0745282
12	NEUTROFILOS METAMIELOCITOS	Serie blanca	0,0702285	0,0702285
6	NEUTROFILOS MIELOCITOS	Serie blanca	0,0651145	0,0651145
18	CELULAS DE DOWNEY	Serie blanca	-0,0086541	0,0086541

Tabla 21 – Pesos asignados de la PC1

Pero si tomamos como fueron asignados los pesos para la componente principal número 2, vemos (Tabla 22) que prioriza los de la serie roja.

#	Variable	Tipo	PC2	PC2 (abs)
4	HEMOGLOBINA	Serie roja	0,4810571	0,4810571
3	HEMATOCRITO	Serie roja	0,4697957	0,4697957
2	HEMATIES RECUENTO	Serie roja	0,3745813	0,3745813
11	RDW	Serie roja	0,2881121	0,2881121
17	MONOCITOS	Serie blanca	-0,2163408	0,2163408
7	NEUTROFILOS SEGMENTADOS	Serie blanca	0,2158580	0,2158580
5	LEUCOCITOS RECUENTO	Serie blanca	0,2032384	0,2032384
15	EOSINOFILOS	Serie blanca	0,1609421	0,1609421
12	NEUTROFILOS METAMIELOCITOS	Serie blanca	0,1581692	0,1581692
6	NEUTROFILOS MIELOCITOS	Serie blanca	0,1548110	0,1548110
16	LINFOCITOS	Serie blanca	-0,1546096	0,1546096
13	NEUTROFILOS EN CAYADO	Serie blanca	0,1428302	0,1428302
9	HCM	Serie roja	0,1366231	0,1366231
8	VCM	Serie roja	0,1209531	0,1209531
10	CHCM	Serie roja	0,1201201	0,1201201
14	BASOFILOS	Serie blanca	0,1132473	0,1132473
1	GLUCOSA	Bioquímica básica	-0,1003685	0,1003685
18	CELULAS DE DOWNEY	Serie blanca	0,0061988	0,0061988

Tabla 22 – Pesos asignados de la PC2

Segundo analizaremos el grafico resultante de la matriz masculina (figura 56). Sobre el eje X ubicamos el primer componente principal PC1 y sobre el eje Y el segundo componente principal PC2, cuyo peso es 19.7% y 17.3% respectivamente.

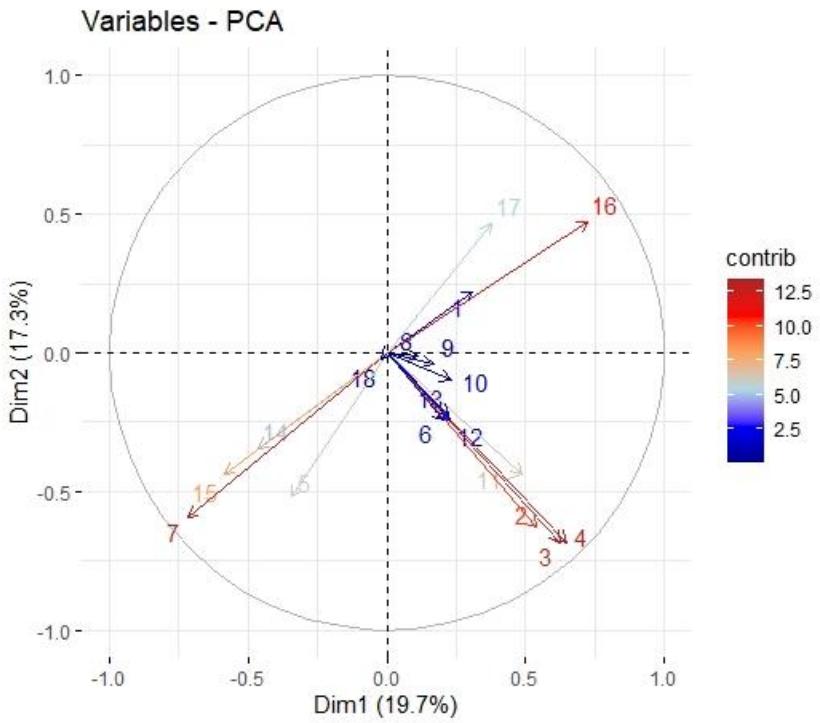


Figura 56 – Fuerza de las variables sobre las dos primeras componentes principales

El análisis que hacemos sobre esta figura es que las dos primeras componentes principales representan un 0.3700 de proporción del total del desvío estándar, muy similar al sexo Femenino.

Segundo, al igual que el femenino, las flechas representan a las 18 variables de análisis, los colores y longitudes son representados de igual manera. Lo que se observa es que las variables por cuadrante son la mismas, lo único que cambian de cuadrante, que en el femenino se encontraban en la parte superior (positivo de la PC2) ahora se encuentran en la parte inferior (negativo de la PC2).

Tabla 23 – Variables por cuadrante y fuerza

En la Tabla 23, se puede observar mejor, con los nombres de las variables cuales tienen mayor peso en cada cuadrante.

En la Tabla 24, observamos todos los valores calculados, diferenciados por color para identificar rápidamente su ubicación en el grafico o cuadrante al que se lo ubica, como vimos en el sexo femenino.

MASCULINO									
Var.	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
4	0,3449529	-0,3868227	0,1620864	0,0473899	0,0079858	0,0190857	-0,0147228	0,1189931	0,0505223
3	0,3331425	-0,3867880	0,1432445	0,1087021	0,0182757	-0,0351399	-0,0885610	0,1982104	0,0946083
7	-0,3815336	-0,3356387	-0,0389304	-0,0579901	0,0380431	-0,0961005	0,0129815	-0,0463975	-0,1589284
16	0,3856560	0,2679291	-0,0073379	0,0431665	-0,0381153	0,0749696	-0,0244693	-0,1034643	0,3863905
2	0,2864437	-0,3563592	0,0509501	0,3327544	0,0000220	0,0544294	0,0358264	0,0821633	0,0414692
15	-0,3121421	-0,2471257	0,0306960	-0,0598563	-0,0471100	0,1524076	0,1845510	0,2555953	0,0114845
11	0,2599224	-0,2469148	0,1263120	-0,1307211	-0,0264394	0,0874706	0,1345833	-0,3389043	-0,2271793
5	-0,1835377	-0,2939590	0,1115635	0,0874947	-0,0036997	-0,3319361	-0,1959067	-0,0766960	-0,3207975
17	0,2015411	0,2641165	0,0194675	0,0205484	0,0294433	0,2339146	0,1692030	0,6250384	-0,5927413
14	-0,2456917	-0,1944638	0,0230723	-0,0859348	-0,1031674	0,1843841	0,5007917	0,2855095	0,4736380
1	0,1645333	0,1253650	0,0070706	0,0230684	-0,0404561	-0,7690387	0,5814976	0,0317237	-0,0324254
12	0,1198595	-0,1379332	-0,5423924	-0,1655353	0,0003929	-0,0175608	-0,0122879	0,0316186	-0,0241465
6	0,1101253	-0,1370947	-0,5136781	-0,1584073	-0,0013320	-0,0259422	-0,0076252	0,0060164	-0,0061932
13	0,1166781	-0,1264984	-0,5059182	-0,1546541	0,0020381	-0,0135352	-0,0132033	0,0529401	-0,0362899
10	0,1238361	-0,0557867	0,1415324	-0,3730265	-0,0581696	0,3198117	0,4108057	-0,4189255	-0,2342918
9	0,0892663	-0,0216591	0,2137789	-0,5796273	0,0190688	-0,0822828	-0,1129286	0,1008475	0,0425282
8	0,0585475	-0,0076708	0,1974792	-0,5308406	0,0445326	-0,2201493	-0,2912716	0,2778977	0,1375751
18	-0,0101446	-0,0113831	0,0005081	0,0001543	0,9873634	0,0266018	0,1226373	-0,0307892	0,0572244
Var.	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
4	0,1411580	-0,0709078	-0,1073655	0,0068960	-0,0008626	0,0570757	-0,4328933	-0,3991062	-0,5468382
3	0,1192105	-0,0043421	-0,0570448	0,0071113	0,0004683	0,0276959	-0,3257384	0,3648988	0,6259931
7	0,2765677	0,2364954	-0,0761047	-0,0090033	-0,0063961	0,7422948	0,0958731	0,0104009	0,0018093
16	-0,2504198	-0,3262484	0,2224706	-0,0137227	-0,0097949	0,6205957	0,0756518	0,0098235	0,0027594
2	0,1254684	-0,0484732	-0,0798424	0,0051668	0,0007628	-0,1059458	0,7869070	0,0232060	-0,0971614
15	0,1576282	-0,4292910	0,7036604	0,0119599	-0,0057098	-0,0858968	-0,0130305	-0,0021189	-0,0005832
11	-0,3755597	0,5200967	0,4824869	-0,0226451	0,0053060	-0,0194575	0,0144602	-0,0057679	0,0003660
5	-0,6192637	-0,4412834	-0,1529395	-0,0038881	-0,0016637	0,0273103	0,0064133	0,0019608	0,0038120
17	-0,1386639	0,0430143	-0,0465458	0,0329322	-0,0066232	0,1939692	0,0243049	0,0026004	0,0004397
14	-0,4340274	0,1704546	-0,2738664	-0,0125504	0,0044771	-0,0005730	-0,0021894	0,0004425	0,0001137
1	0,1292930	-0,0161839	0,0797609	-0,0032060	0,0023468	-0,0037800	-0,0021402	0,0004208	0,0007855
12	-0,0378095	-0,0464842	-0,0057437	0,0700526	0,7964395	-0,0112023	-0,0014341	-0,0002237	-0,0004092
6	-0,0487860	-0,0239934	-0,0062555	0,6627443	-0,4865149	-0,0196580	-0,0010946	0,0013898	-0,0008498
13	-0,0240783	-0,0456001	-0,0124446	-0,7439837	-0,3587604	-0,0314463	-0,0028779	-0,0001846	0,0002598
10	0,1665758	-0,3653105	-0,2870655	0,0013244	-0,0020580	-0,0015435	0,0024423	0,2692185	-0,0381806
9	0,0411499	-0,0550342	-0,0575490	0,0047466	-0,0002490	-0,0187152	0,2347164	-0,6012174	0,3912370
8	-0,0203993	0,0827238	0,0495739	0,0029561	0,0016433	-0,0198178	0,1273733	0,5223295	-0,3809310
18	-0,0545104	-0,0375058	0,0173903	-0,0005905	0,0000683	-0,0149244	-0,0026413	0,0001055	0,0002763

Tabla 24 – Fuerza de las variables por componente principal

Por ultimo si analizamos cada una de las dos primeras componentes principales y las ordenamos en forma descendente, podemos observar (Tabla 25) que la primera componente recoge alternadamente información correspondiente a las variables que componen la serie blanca y roja.

#	Variable	Tipo	PC1	PC1 (abs)
16	LINFOCITOS	Serie blanca	0,3856560	0,3856560
7	NEUTROFILOS SEGMENTADOS	Serie blanca	-0,3815336	0,3815336
4	HEMOGLOBINA	Serie roja	0,3449529	0,3449529
3	HEMATOCRITO	Serie roja	0,3331425	0,3331425
15	EOSINOFILOS	Serie blanca	-0,3121421	0,3121421
2	HEMATIES RECUENTO	Serie roja	0,2864437	0,2864437
11	RDW	Serie roja	0,2599224	0,2599224
14	BASOFILOS	Serie blanca	-0,2456917	0,2456917
17	MONOCITOS	Serie blanca	0,2015411	0,2015411
5	LEUCOCITOS RECUENTO	Serie blanca	-0,1835377	0,1835377
1	GLUCOSA	Bioquímica básica	0,1645333	0,1645333
10	CHCM	Serie roja	0,1238361	0,1238361
12	NEUTROFILOS METAMIELOCITOS	Serie blanca	0,1198595	0,1198595
13	NEUTROFILOS EN CAYADO	Serie blanca	0,1166781	0,1166781
6	NEUTROFILOS MIELOCITOS	Serie blanca	0,1101253	0,1101253
9	HCM	Serie roja	0,0892663	0,0892663
8	VCM	Serie roja	0,0585475	0,0585475
18	CELULAS DE DOWNEY	Serie blanca	-0,0101446	0,0101446

Tabla 25 – Pesos asignados de la PC1

Pero si tomamos como fueron asignados los pesos para la componente principal número 2, vemos (Tabla 26) que prioriza los de la serie roja.

#	Variable	Tipo	PC2	PC2 (abs)
4	HEMOGLOBINA	Serie roja	0,4810571	0,4810571
3	HEMATOCRITO	Serie roja	0,4697957	0,4697957
2	HEMATIES RECUENTO	Serie roja	0,3745813	0,3745813
11	RDW	Serie roja	0,2881121	0,2881121
17	MONOCITOS	Serie blanca	-0,2163408	0,2163408
7	NEUTROFILOS SEGMENTADOS	Serie blanca	0,2158580	0,2158580
5	LEUCOCITOS RECUENTO	Serie blanca	0,2032384	0,2032384
15	EOSINOFILOS	Serie blanca	0,1609421	0,1609421
12	NEUTROFILOS METAMIELOCITOS	Serie blanca	0,1581692	0,1581692
6	NEUTROFILOS MIELOCITOS	Serie blanca	0,1548110	0,1548110
16	LINFOCITOS	Serie blanca	-0,1546096	0,1546096
13	NEUTROFILOS EN CAYADO	Serie blanca	0,1428302	0,1428302
9	HCM	Serie roja	0,1366231	0,1366231
8	VCM	Serie roja	0,1209531	0,1209531
10	CHCM	Serie roja	0,1201201	0,1201201
14	BASOFILOS	Serie blanca	0,1132473	0,1132473
1	GLUCOSA	Bioquímica básica	-0,1003685	0,1003685
18	CELULAS DE DOWNEY	Serie blanca	0,0061988	0,0061988

Tabla 26 – Pesos asignados de la PC2

9. Conclusiones

Si nos basamos en los objetivos planteados en el presente trabajo de investigación, podemos dividirlo en dos grandes grupos, el primero basado en la mejora de procesos y el control sobre ellos y el segundo en la implementación de análisis predictivos para anticiparse

a diagnósticos o patologías más severas y por ultimo a la investigación y nuevos descubrimientos en la ciencia de la medicina.

En cuanto al primero, se planteó el objetivo de reducir el error en las solicitudes, ha quedado demostrado que hay un porcentaje de duplicidad en las solicitudes que puede ser eliminado mediante un control sistemático al solicitarlas y por otro lado, en un segundo análisis, se observó que en todas las variables analizadas, existen una cantidad de solicitudes que su valor es igual al anterior, por ello se deben analizar en mayor detalle para ver si son necesarias de realizar, como así también se demostró la cantidad de solicitudes realizadas en una frecuencia de extracción muy pequeña, de 1, 2, 3, etc. horas.

Para los análisis descriptivos, se observó que las variables de laboratorio, en su mayoría, no poseen una distribución normal y se detectaron la presencia de outliers que no necesariamente son valores erróneos, sino que esos valores pueden ser correctos. Como se mencionó anteriormente por lo general cada variable tienen una, dos o tres limites probables.

En el análisis de correlación entre las variables, se detectaron varias de ellas, con distinta fuerza, las cuales son correctas, esto motiva a futuros trabajos de investigación en donde se realice este procedimiento con otras variables o incluso sumarlas a las ya analizadas, formando patrones y documentándolos.

Luego se aplicó un modelo de clusterización, donde basado en las variables disponibles, no ha mostrado un resultado bien definido. Eso puede haber sucedido por las variables analizadas y sus pesos en los diagnósticos de los pacientes internados. Un siguiente paso sería agregar más variables al modelo, que determinen y definan mejor la clusterización o la utilización de otros algoritmos.

La utilización del algoritmo de componentes principales ha mostrado buenos resultados en la determinación de los pesos de cada variable, por lo que es un buen camino de investigación para los próximos trabajos, incluso en la predicción e incidencia de las variables en los diagnósticos.

Como futuros pasos a seguir, es continuar con la investigación y practica sobre los componentes principales a fin de desarrollar un modelo predictivo basado en el.

10. Referencias - Bibliografía

[1] L. McCay, C. Lemer, and A. W. Wu, “Laboratory safety and the WHO World Alliance for Patient Safety,” *Clin. Chim. Acta*, vol. 404, no. 1, pp. 6–11, Jun. 2009, doi: 10.1016/j.cca.2009.03.019.

[2] H. Pabón Martínez, P. A. Londoño Núñez, herbys.pabon@campusucc.edu.co, and paola.londono@campusucc.edu.co, “Plan de mejoramiento para disminuir los errores en la fase preanalítica en los análisis de laboratorio en la Clínica Regional de Occidente de la Policía Nacional seccional sanidad Valle de la ciudad de Santiago de

Cali," 2018.

[3] "Análisis de Sangre." [Online]. Available: https://www.tuotromedico.com/Guias/Analisis-de-Sangre/. [Accessed: 27-Jun-2020].

[4] "Transformación Box-Cox - Wikipedia, la enciclopedia libre." [Online]. Available: https://es.wikipedia.org/wiki/Transformación_Box-Cox. [Accessed: 15-Sep-2020].

[5] "Correlación - Wikipedia, la enciclopedia libre." [Online]. Available: https://es.wikipedia.org/wiki/Correlación. [Accessed: 15-Sep-2020].

[6] "¿Cómo interpretar y para qué sirve el coeficiente de correlación? - Rankia." [Online]. Available: https://www.rankia.cl/blog/mejores-opiniones-chile/4090045-como-interpretar-para-que-sirve-coeficiente-correlacion. [Accessed: 15-Sep-2020].

[7] "K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks | by Imad Dabbura | Towards Data Science." [Online]. Available: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. [Accessed: 14-Jan-2021].

[8] "¿Qué es el análisis de clúster? - Analytics Lane." [Online]. Available: https://www.analyticslane.com/2018/12/17/que-es-el-analisis-de-cluster/. [Accessed: 16-Sep-2020].

[9] F. Villena and J. Dunstan, "Obtención automática de palabras clave en textos clínicos: una aplicación de procesamiento del lenguaje natural a datos masivos de sospecha diagnóstica en Chile," 2019.

[10] "Understanding Boxplots. The image above is a boxplot. A boxplot… | by Michael Galarnyk | Towards Data Science." [Online]. Available: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51. [Accessed: 16-Jan-2021].

[11] "A Step-by-Step Explanation of Principal Component Analysis." [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis. [Accessed: 14-Jan-2021].