

**UNIVERSIDAD NACIONAL AGRARIA  
LA MOLINA  
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN**



**“SEGMENTACIÓN DE CLIENTES DE UN CASINO UTILIZANDO  
EL ALGORITMO PARTICIÓN ALREDEDOR DE MEDOIDES  
(PAM) CON DATOS MIXTOS”**

PRESENTADO POR  
**RHONY MIGUEL ELGUERA VEGA**

TESIS PARA OPTAR EL TÍTULO DE  
**INGENIERO ESTADÍSTICO E INFORMÁTICO**

Lima – Perú  
2018

**UNIVERSIDAD NACIONAL AGRARIA  
LA MOLINA**

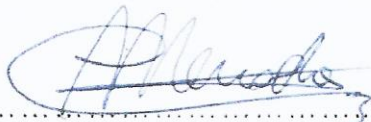
**FACULTAD DE ECONOMÍA Y PLANIFICACIÓN**

**“SEGMENTACIÓN DE CLIENTES DE UN CASINO UTILIZANDO  
EL ALGORITMO PARTICIÓN ALREDEDOR DE MEDOIDES  
(PAM) CON DATOS MIXTOS”**

**PRESENTADO POR  
RHONY MIGUEL ELGUERA VEGA**

**INGENIERO ESTADÍSTICO E INFORMÁTICO**

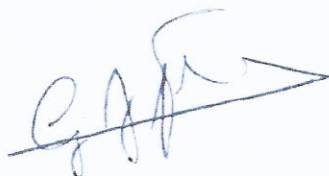
**SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO**



Dr. César Higinio Menacho Chiok  
Presidente



Mg. Jesús Walter Salinas Flores  
Asesor



MS. Grimaldo José Febres Huamán  
Miembro



MS. Carlos López de Castilla Vásquez  
Miembro



UNIVERSIDAD NACIONAL AGRARIA LA MOLINA  
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN



## ACTA DE SUSTENTACIÓN

Los miembros del Jurado que suscriben, reunidos para calificar la sustentación del Trabajo de Tesis, presentado por el Bachiller en Ciencias – Estadística e Informática, señor **RHONY MIGUEL ELGUERA VEGA**, titulado:

**“SEGMENTACIÓN DE CLIENTES DE UN CASINO UTILIZANDO EL ALGORITMO PARTICIÓN ALREDEDOR DE MEDOIDES (PAM) CON DATOS MIXTOS”**

Oídas las respuestas y las observaciones formuladas, declaramos: ... *Aprobado* .....

Con el calificado de: ... *Muy Buena* .....

En consecuencia, queda en condición de ser calificado APTO por el Consejo Universitario y recibir el Título Profesional de **INGENIERO ESTADÍSTICO E INFORMÁTICO** de conformidad con lo estipulado en el Artículo 89° del Estatuto de la Universidad Nacional Agraria La Molina (Resolución N° 01-2015-AE-UNALM, del 23/02/15), el Artículo 150°, inciso b) del Reglamento General de la UNALM 2017 (Resolución N° 001-2017-AU-UNALM), y el anexo N° 4, punto N° 23 (Modalidad para obtener el título profesional: 1. Sustentación de Tesis, según el Reglamento del Registro Nacional de Grados y Títulos.

Lima, 05 de enero del 2018

Dr. César Higinio Menacho Chio  
PRESIDENTE

Mg. Jesús Walter Salinas Flores  
ASESOR

MS. Grimaldo José Febres Huamán  
MIEMBRO

MS. Carlos López de Castilla Vásquez  
MIEMBRO

Ref.: TR. N°091-2017/FEP, del 17/03/17  
(Matricula N° 200702212)

*Mercader*

*A mis padres Rhony Elguera y Nélida Vega por estar a mi lado apoyándome y aconsejándome siempre, y por hacer de mí una mejor persona a través de sus consejos, enseñanzas y amor.*

## **AGRADECIMIENTOS**

Le agradezco a Dios por haberme acompañado y guiado a lo largo de mi carrera, por ser mi fortaleza en los momentos de debilidad y por brindarme una vida llena de aprendizajes, experiencias y sobre todo felicidad.

Al empresa de casinos por haberme dado la oportunidad de ampliar mis conocimientos en este rubro tan inmenso, y por haberme proporcionado los datos necesarios para llevar a cabo mi investigación.

A la Universidad Nacional Agraria La Molina por darme la oportunidad de estudiar y ser un profesional.

Al Ing. Mg. Jesús Walter Salinas Flores, asesor de la presente tesis, mi más profundo agradecimiento, por su relevante orientación y aporte durante el desarrollo del presente trabajo, así como su indispensable apoyo en la culminación del mismo.

Al Dr. César Higinio Menacho Chiok por las primeras correcciones de este ejemplar, su constante motivación y sus acertadas orientaciones y sugerencias.

A la Dra. Paola Manrique Holguín por su apoyo y aliento para la culminación de esta investigación.

A mi hermano Mario por su apoyo para la culminación de esta investigación.

A toda mi familia por darme todo su apoyo y por quererme por sobre todas las cosas.

A mi abuelita Mami Candy por tenerme siempre presente en sus oraciones.

A mis amigos por los gratos recuerdos compartidos y por su apoyo incondicional.

# ÍNDICE

DEDICATORIA

AGRADECIMIENTOS

RESUMEN

ABSTRACT

I.	INTRODUCCIÓN .....	1
II.	REVISIÓN DE LITERATURA .....	4
2.1	Análisis de agrupamiento o clustering.....	5
2.1.1	Los Métodos de agrupamiento.....	5
2.1.2	Los métodos de agrupamiento aplicados a la investigación de mercados .....	7
2.1.3	Métricas para los métodos de agrupamiento.....	9
2.1.4	Métodos para validar la calidad de los clúster .....	12
2.2	El método de partición k-Medoides.....	15
2.2.1	Determinación del número óptimo de clústers .....	15
2.2.2	Algoritmo de Partición Alrededor de Medoides (PAM).....	18
2.2.3	Ejemplo de aplicación del algoritmo PAM.....	20
III.	MATERIALES Y MÉTODOS .....	24
3.1	Materiales .....	24
3.2	Metodología.....	24
3.2.1	Tipo de investigación y formulación de hipótesis .....	24
3.2.2	Población y Muestra .....	25
3.2.3	Identificación de variables .....	25
3.2.4	Proceso para análisis de conglomerados.....	26
IV.	RESULTADOS Y DISCUSIÓN .....	28
4.1	Recopilación de datos.....	28
4.2	Pre procesamiento de datos.....	28
4.3	Análisis de agrupamiento con el algoritmo PAM.....	29
4.4	Validación del agrupamiento con el algoritmo PAM .....	31
4.5	Caracterización del agrupamiento con el algoritmo PAM .....	34
V.	CONCLUSIONES .....	39
VI.	RECOMENDACIONES.....	41
VII.	REFERENCIA BIBLIOGRÁFICA .....	42
VIII.	ANEXOS .....	43

## Índice de Figuras

	<b>Pág.</b>
Figura 1. Dendrograma	17
Figura 2. Gráfico de la silueta con el algoritmo PAM	30
Figura 3. Árbol de clasificación C5.0.	35
Figura 4. Distribución de monto de recargas para el clúster 1	37
Figura 5. Distribución de monto de recargas para el clúster 2	37
Figura 6. Distribución de montos de recargas para el clúster 3	38

## Índice de Cuadros

	<b>Pág.</b>
Cuadro 1. Muestra de 10 individuos recién nacidos	20
Cuadro 2. Cálculo de los parámetros para el cálculo de Gower	21
Cuadro 3. Matriz de distancias de Gower	21
Cuadro 4. Matriz de distancias de Gower (observaciones S1 y S2)	22
Cuadro 5. Matriz de distancias de Gower (observaciones S1 y S3)	23
Cuadro 6. Matriz de distancias de Gower (observaciones S4 y S7)	23
Cuadro 7. Cuartiles para el monto de recarga	29
Cuadro 8. Medoides finales con el algoritmo PAM	31
Cuadro 9. Distribución de número y porcentaje de clientes con el PAM	31
Cuadro 10. Análisis de Varianza para las variables agrupadas por clúster	32
Cuadro 11. Tabla de contingencia para estudiar la asociación entre sexo y la variable Clúster	32
Cuadro 12. Tabla de contingencia para estudiar la asociación entre nivel y la variable Clúster	33
Cuadro 13. Matriz de confusión considerando la clasificación como clúster	34
Cuadro 14. Reglas para los clústers según árbol de clasificación C5.0 con poda	34
Cuadro 15. Distribución de número y porcentaje de clientes con el algoritmo PAM y con el Árbol de clasificación C5.0	35
Cuadro 16. Promedios de las variables por clústeres	36
Cuadro 17. Distribución por Sexo según los clústers	36
Cuadro 18. Distribución por tipo de cliente según los clústeres	36

## RESUMEN

En la actualidad, la gran cantidad de datos que se almacenan de los clientes en las diferentes empresas y la capacidad de procesamiento que brindan las computadoras, han generado gran interés por investigar; así como, desarrollar métodos y algoritmos para el análisis de agrupamiento. Los métodos de agrupamiento dirigidos a la segmentación de clientes permiten a las empresas identificar los patrones y perfiles de compra o servicios, ayudando a tomar mejores decisiones de las estrategias de canales y publicidad para sus clientes. En la presente investigación se aplica el método de agrupamiento basado en las particiones de k-Medoides con el algoritmo PAM (Partición Alrededor de Medoides). El algoritmo PAM se basa en particionar el conjunto de datos en  $k$  grupos, donde  $k$  es conocido; es considerado más robusto ante datos atípicos y el ruido, se basa en minimizar la suma de disimilitudes entre un objeto y el Medoide (centro del grupo). El objetivo de la presente investigación es aplicar el algoritmo PAM para segmentar a los clientes de un casino con los datos obtenidos, a través del uso de tarjetas en el tragamonedas. El método de la silueta permitió identificar tres clústers como el número óptimo. El análisis de agrupamiento con el algoritmo PAM usando la medida de distancia Gower, resultó la segmentación de clientes para los tres clúster con porcentajes de 49.4%, 11.3% y 39.4% respectivamente. La agrupación fue validada, al obtener para las 6 variables cuantitativas todos los ANVAs significativos y con el árbol de clasificación C5.0 un 99.35% de precisión. Los resultados de la caracterización muestran que el clúster 1 son clientes con valores de los promedios para las 6 variables en un nivel intermedio, el 67.0% son hombres y 100% el tipo de tarjeta es classic. En el clúster 2 están los clientes con los valores más altos en los promedio de las 6 variables, el 59% son hombres y el 100% usan la tarjeta silver. En el clúster 3, se encuentran los clientes con los promedios más bajos, el 64% son hombres y el 100% usan tarjeta classic.

**Palabras claves.** Segmentación de clientes, Partición alrededor de medoides (PAM), Índice de silueta, Tipificación de clientes.



## ABSTRACT

Currently, the large amount of data stored by customers in different companies and the processing capacity provided by computers have generated great interest in research; as well as, develop methods and algorithms for grouping analysis. Clustering methods aimed at customer segmentation allow companies to identify patterns and profiles of purchase or services, helping them to make better decisions on channel and advertising strategies for their clients. In the present investigation the grouping method based on the partitions of k-Medoids with the PAM (Partition Around Medoids) algorithm is applied. The PAM algorithm is based on partitioning the data set into k groups, where k is known; is considered more robust to atypical data and noise, is based on minimizing the sum of dissimilarities between an object and the Medoid (center of the group). The objective of this research is to apply the PAM algorithm to segment the customers of a casino with the data obtained, through the use of cards in the slot machine. The silhouette method allowed to identify three clusters as the optimal number. The cluster analysis with the PAM algorithm using the Gower distance measure, resulted in the segmentation of clients for the three clusters with percentages of 49.4%, 11.3% and 39.4% respectively. The grouping was validated, obtaining all the significant ANVAs for the 6 quantitative variables and 99.35% accuracy with the C5.0 classification tree. The results of the characterization show that Cluster 1 are clients with averages values for the 6 variables in an intermediate level, 67.0% are men and 100% the card type is classic. In Cluster 2 there are the clients with the highest values in the average of the 6 variables, 59% are men and 100% use the silver card. In Cluster 3, customers with the lowest averages are found, 64% are men and 100% use classic cards.

**Keywords:** Customer segmentation, Partition around medoids (PAM), Silhouette index, Customer classification.

## I. INTRODUCCIÓN

El análisis cluster es el proceso de agrupar un conjunto de objetos (ejemplos) de tal manera, que los objetos dentro de un clúster tengan una similitud alta entre ellos (homogeneidad dentro de grupos) y baja con objetos de otros grupos (heterogeneidad entre grupos). El análisis de agrupamiento o análisis de cluster se considera una técnica descriptiva que corresponde a un aprendizaje no supervisado, a diferencia de las técnicas para el aprendizaje supervisado (hay una clase o variable que predecir) que están asociadas a los problemas de clasificación. En la actualidad, la gran cantidad de datos que se están almacenando y la potencia que brinda las computadoras han despertado gran interés por investigar y desarrollar métodos y algoritmos que permitan ser aplicados para el análisis de agrupamiento en diferentes sectores productivos; tales como, la industria, los negocios, el sector financiero, servicios y en las diferentes comunidades de investigación; como la Estadística, la minería de datos, aprendizaje de máquina, la inteligencia artificial, etc.

La clasificación y extracción de patrones y perfiles usando los datos de los clientes es muy importante para el soporte comercial y toma de decisiones oportuna de la aplicación de las nuevas tendencias emergentes de negocios para los procesos comerciales. En los últimos años, se ha reconocido que las técnicas de agrupación particionadas son adecuadas para agrupar un gran conjunto de datos debido a sus requisitos computacionales relativamente bajos. El análisis de agrupación con el propósito de la segmentación de los clientes, ayuda a las empresas a derivar iniciativas estratégicas al poder seleccionar el canal y la publicidad más apropiados de marketing para una campaña táctica (Sankar, R., 2011). La industria minorista recopila grandes cantidades de datos sobre ventas, historial de compras de clientes, bienes de transporte, consumo y servicio. Con una mayor disponibilidad y facilidad de uso de la moderna tecnología de la computación y el comercio electrónico, la disponibilidad y popularidad de tales negocios ha crecido rápidamente. Las técnicas de agrupación para segmentar perfiles de clientes para una tienda minorista, puede ayudar a identificar patrones y comportamientos de compras, mejorar el servicio y la satisfacción de los clientes produciendo el aumento y la retención (Pramod, P. & Latesh, G., 2011).

El objetivo de los métodos de agrupamiento es particionar el conjunto de objetos (observaciones) en diferentes grupos (clústeres). Existen una variedad de métodos para aplicar el análisis de clúster que son categorizados en: jerárquicos, basado en particiones, basado en densidades, basado en rejillas y basado en modelos. El agrupamiento basado en particiones son métodos de clúster no jerárquico, donde el número de clúster  $k$  a formarse es un valor

conocido y la conformación de los grupos se realiza por un proceso iterativo de  $k$  particiones, donde los objetos se distribuyen (moviéndose) en los  $k$  grupos, de tal manera que se va minimizando las distancias entre los objetos dentro del grupo con respecto a su *centroide* (la media, mediana, moda, medoides, etc.). Existen dos métodos basados en particiones que se están aplicando en la mayoría de los casos:  $k$ -medias y  $k$ -medoides. Estas técnicas se basan en particionar el conjunto de objetos en grupos, de tal manera que van minimizando su distancia (métrica para medir la similaridad) respecto a un centroide; que para el algoritmo de  $k$ -medias es la media y para el  $k$ -medoides algún dato que representa el punto medio dentro del grupo.

Dentro de las categorías de los métodos de agrupamiento basados en particiones, el algoritmo de mayor uso es el  $k$ -medias; sin embargo, tiene el inconveniente de ser computacionalmente muy extenso y la calidad de sus resultados del agrupamiento depende la elección del centroide inicial y la dimensión de los datos; más aún si existen datos atípicos y las variables son cuantitativas y cualitativas. Para resolver estos problemas se han desarrollado algoritmos que permiten lidiar con estas limitaciones, como el  $k$ -medoides (Batra, 2011). El método de  $k$ -medoides, es una técnica de partición de grupos que divide el conjunto de  $n$  objetos en  $k$  grupos (con  $k$  conocido a priori). En comparación del  $k$ -medias que usa como centroide la media del grupo, el  $k$ -Medoide usa el punto medio del grupo (Medoide). El  $k$ -Medoide es más robusto ante datos atípicos y el ruido; se basa en minimizar la suma de disimilaridades entre un objeto y el Medoide, a diferencia de  $k$ -medias que minimiza la suma de distancias euclidianas cuadradas. Entre los algoritmos para el método de  $k$ -Medoides se encuentran: PAM (Partitioning Around Medoids), CLARA (Clustering for Large Applications) y CLARANS (Clara based on Randomized Search). En la presente investigación se aplicó el algoritmo Partición Alrededor de Medoides (PAM), que fue desarrollado por (Kaufman & Rousseeuw, P., 1990), es más robusto puesto que es menos sensible a valores atípicos y al ruido, y puede trabajar con variables mixtas, siendo aplicado en muchos casos como una alternativa al  $k$ -medias. El algoritmo PAM, requiere conocer el número de clusters  $k$ . Un método útil para determinar el número óptimo de grupos es a través del índice de silueta, que es una métrica o indicador para evaluar el número óptimo de clúster a formarse. El PAM para trabajar con variables mixtas usa como métrica para evaluar la similaridad la medida de distancia conocida como Gower.

Los datos con los cuales se realizó la aplicación del algoritmo PAM fueron obtenidos de la base de datos de clientes de una empresa que cuenta con un casino en la ciudad de Trujillo. La finalidad fue usar los datos de los jugadores de tragamonedas que asisten al casino, para

obtener una segmentación de los clientes para conocer sus preferencias que apoyen en el área de marketing para mejorar la atención y ofrecer mejores productos y premios a sus jugadores más destacados, y principalmente, motivar a los jugadores ocasionales y de esta manera, incrementar el tráfico de clientes en la Sala de Juegos. Para el análisis de agrupamiento, se utilizó el programa estadístico R.

En la presente investigación, se presentó en primer lugar una generalización del análisis de agrupamiento mostrando una taxonomía de la categorización de los métodos, una revisión bibliográfica de las aplicaciones de los métodos de agrupamiento en el campo de investigación de mercados, las principales métricas que se aplican, las técnicas para determinar el número de clúster y la validación del agrupamiento. Se describe el método de agrupación con el algoritmo PAM. En los resultados y discusión se aplica un caso de estudio para la agrupación de los clientes de un casino.

Los objetivos de la presente investigación son:

1. Presentar la metodología para la aplicación del algoritmo Partición Alrededor de Medoides (PAM).
2. Realizar el análisis de agrupamiento para segmentar a los clientes de un casino con el algoritmo PAM.
3. Realizar la caracterización de los clientes del agrupamiento resultante.

## II. REVISIÓN DE LITERATURA

La agrupación es un proceso de partición de un conjunto de datos (objetos) en un conjunto de subclases significativas, llamadas agrupaciones o clústeres. El agrupamiento ayuda a los tomadores de decisiones a comprender la estructura del conjunto de datos. Un clúster, es una colección de objetos de datos que son "similares" entre sí y diferentes entre objetos de diferentes clústeres. El análisis de agrupamiento, pertenece a lo que se conoce como el aprendizaje no supervisado (no hay una variable clase) donde todas las variables son independientes a diferencia del aprendizaje supervisado (hay una variable clase) donde los datos están clasificados por una variable dependiente.

Existen una variedad de métodos de agrupamiento. Un buen método de agrupamiento debe generar clústeres de alta calidad por lo cual debe considerarse:

- Similitud intracase (dentro del clúster) debe ser alta y la disimilitud entre clases debe ser alta.
- La calidad de un resultado de agrupación también depende de la medida de similitud utilizada por el método y su implementación.
- La calidad de un método de agrupamiento también se mide por su capacidad de descubrir algunos o todos los patrones ocultos.
- La calidad de un resultado de agrupación también depende de la definición y representación del clúster elegido.

Los métodos de agrupación tienen amplias aplicaciones en los diversos campos del saber humano:

- Reconocimiento de patrones. Identificando comportamiento de perfiles de objetos.
- Análisis de datos espaciales. Creando mapas temáticos en SIG al agrupar espacios de características, detectando clústeres espaciales y los explica en minería de datos espaciales.
- Estudios de mercados. Ayuda a los especialistas en marketing a descubrir grupos distintos en sus bases de clientes y luego utilice este conocimiento para desarrollar programas de marketing específicos.
- Uso de la tierra. Identificación de áreas de uso similar de la tierra en una base de datos de observación de la tierra.
- Seguros. Identificación de grupos de titulares de pólizas de seguros de automóviles con un alto costo medio de reclamación.

- Planificación urbana. Identificación de grupos de casas de acuerdo con el tipo de casa, el valor y la ubicación geográfica.
- Estudios de terremotos. Ubicación de los epicentros de terremotos observados deben agruparse a lo largo de las fallas continentales

## 2.1 Análisis de agrupamiento o clustering

El Análisis de agrupamiento, conocido como Análisis de conglomerados, es una técnica estadística multivariada cuyo propósito es agrupar un conjunto de objetos, tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos.

### 2.1.1 Los Métodos de agrupamiento

En la literatura existen una variedad de métodos y algoritmos para la agrupación de objetos. La selección de un algoritmo de agrupamiento, depende generalmente de la cantidad y tipo de datos disponibles y el propósito de su aplicación. Una categorización de los algoritmos de agrupamiento son los siguientes:

1. **Métodos jerárquicos:** Se basan en un proceso secuencial para la formación de los grupos o clúster, a través de establecer jerarquías entre los objetos o individuos. Pueden ser métodos aglomerativos o divisivos. **Método aglomerativo:** Se inicia con un número de clúster igual al número total de objetos y se van uniendo en forma aglomerativa de acuerdo a una métrica de distancias; al final se forma un sólo clúster con todos los objetos. Las técnicas para la formación de los clúster son: *Enlace simple (vecino más cercano)*, *Enlace completo (vecino más alejado)*, *Enlace promedio*, *Enlace centroide*, *Enlace Ward*. **Método de divisiones:** Su proceso secuencial de formación de clúster es contrario al aglomerativo. Se inicia con un solo grupo o clúster, que contiene el total de objetos. Luego se van dividiendo (descendente) en subgrupos considerando los más alejados, hasta llegar a “n” grupos de un solo objeto. El más usado es: *Algoritmo de Howard-Harris*. .
2. **Métodos basados en particiones:** Es un método de clúster no jerárquico, donde el número de clúster a formarse  $k$  es un valor (parámetro) conocido. Los  $k$  grupos o clúster se construyen por un proceso iterativo de conformación de  $k$  particiones. El método consiste en distribuir (moviendo) los objetos en los  $k$  grupos, minimizando las distancias entre los objetos dentro del grupo con respecto a su *centroide* (la media, mediana, moda, medoides, etc.). Se inicia seleccionando  $k$  objetos aleatoriamente como los centroides

iniciales para cada clúster. En cada iteración, se asigna el objeto al clúster más similar (la menor distancia con respecto al centroide) y se calcula el nuevo centroide de cada clúster y se termina con la asignación de todos los objetos bajo un criterio de parada de convergencia. Los algoritmos basados en particiones son: k-medias, k-medianas, k-medoides (PAM y CLARA), etc.

3. **Métodos basados en modelo:** Se basan en hallar un modelo para cada clúster, que mejor se ajuste los datos a cada clúster. El método COBWEB pertenece a los métodos de aprendizaje conceptual o basado en modelos. Esto significa que cada clúster se considera como un modelo descriptivo. El COBWEB, es considerado un cluster jerárquico y su aprendizaje se representa por un árbol de clasificación (árbol conceptual jerárquico), donde cada nodo es un Cluster (concepto) que tiene una descripción probabilística del concepto que resume los objetos clasificados en él. El algoritmo se aplica a atributos cualitativos (COBWEB) y se extiende a numéricos usando la Normal (CLASSIST). COBWEB depende del orden de los objetos (se recomienda probar con los objetos en diferente orden).
4. **Métodos probabilísticos:** Se basan en hallar y usar funciones de densidades como medida de aproximación. El EM (Expectation Maximization), es conocido como el clustering probabilístico. El EM, busca el grupo de clústeres más probable dado un conjunto de ejemplos. Los ejemplos tienen cierta probabilidad de pertenecer a un grupo o cluster. Este clustering se basa en el modelo estadístico de mezcla de distribuciones. El EM, trata de obtener la función de densidad de probabilidades (FDP) desconocida para el conjunto de datos, haciendo una aproximación por una combinación lineal de las k distribuciones asociadas a cada cluster (mezcla de k distribuciones de probabilidades). La mezcla más sencilla se tiene cuando los atributos son numéricos con distribuciones gaussianas, determinándose k distribuciones normales con medias y variancias diferentes para cada cluster.
5. **Métodos basados en densidades:** Agrupan los objetos usando como vecindad la densidad de los objetos y evaluándola dentro de un umbral.

### **2.1.2 Los métodos de agrupamiento aplicados a la investigación de mercados**

La clasificación y extracción de patrones de los datos de los clientes es muy importante para el soporte de la toma de decisiones comercial. La identificación oportuna de las nuevas tendencias emergentes es muy importante en los procesos comerciales. Actualmente las empresas están generando y almacenando grandes volúmenes de datos como producto de sus procesos y transacciones comerciales, por lo cual necesitan aplicar técnicas de análisis de datos que le permitan explotarlas para extraer conocimientos relevantes sobre todo del comportamiento y perfiles de sus clientes. En el área de marketing, los métodos de agrupamiento se convierten en herramientas potentes que permiten segmentar a los clientes para identificar sus patrones y sus comportamientos de los productos o servicios que ofrecen las empresas.

En (Sankar, 2011), se aplica un algoritmo de agrupación a la información demográfica para identificar la agrupación de clientes. En fase 1, los datos del cliente se limpian y desarrollan patrones usando varios parámetros y posteriormente, en la fase 2, se perfilaron los datos, aplicando técnicas de conglomerados se identificaron a los clientes con bajo y alto riesgo. A partir de los resultados experimentales, se mostró que el enfoque propuesto genera un patrón más útil a partir de datos de gran tamaño. Se identifican tres segmentos de clientes, de alto beneficio, alto valor y bajo riesgo por una de las técnicas de agrupamiento IBM I-Miner. Se caracteriza el clúster de bajo riesgo y alto valor que representa el 10-20 por ciento de los clientes que produce el 80% de los ingresos.

En (Prasad & Latesh, G., 2011), se menciona la importancia del análisis de agrupación en la industria minorista con la finalidad de identificar a los clientes por sus hábitos de compras, patrones y comportamientos, mejorar el servicio al cliente para una mejor satisfacción del cliente y, por lo tanto aumentar la retención. Los resultados del agrupamiento con k-medias apoya a un sistema de inteligencia de negocios proporcionando un poderoso análisis multidimensional y herramientas de visualización, incluida la construcción de cubos de datos sofisticados según las necesidades del análisis de datos. El uso de una aplicación de inteligencia empresarial que incorpora este agrupamiento como mecanismo para administrar un negocio minorista proporcionará a los minoristas medios para segmentar clientes y comprender su comportamiento y necesidades de una mejor manera, tomar decisiones basadas en el conocimiento con el fin de proporcionar un servicio personalizado y eficiente al cliente.



En (Aroral, P. & Varshney, S., 2015), se aplica los dos algoritmos de agrupamiento basados en particiones que son los más populares K-Means y K-Medoids se evalúan en el conjunto de datos transaccional. Los resultados de la comparación muestran que el tiempo empleado en la selección de los valores iniciales y la complejidad espacial de la superposición del clúster es mucho mejor en K-Medoids que en K-Means. Además, K-Medoids es mejor en términos de tiempo de ejecución, no sensible a valores atípicos y reduce el ruido en comparación con K-Means, ya que minimiza la suma de las diferencias de los objetos de datos.

En (Arbin, N., Suhailayani, N., & Zafirah, N., 2012), se realizó el análisis K-Means y K-Medoids con varios conjuntos de datos. Se analizaron con diferentes parámetros y atributos de los datos. El análisis comparativo de ambos algoritmos con diferentes grupos de datos para diseñar las fortalezas y debilidades de ambos. Se realizaron estudios exhaustivos para determinar la correlación de los datos con los algoritmos para encontrar la relación entre ellos. Ambos enfoques implementados produjeron buen resultado con error cuadrático medio inferior al 3%. Sin embargo, en la mayoría de los conjuntos de datos, el K-Medoids resultó ser el mejor para agrupar los datos.

En (Tiwari & Singh, R., 201), se realiza un estudio comparativo entre los algoritmos de agrupamiento basados en particiones k-means y k-medoid y comparamos la evaluación del rendimiento de ambos con datos de IRIS sobre la base de la complejidad de tiempo y espacio. Los resultados indican que los métodos de agrupamiento basados en particiones los algoritmos k-means y k-medoids son adecuados para formar clústeres esféricos con conjunto de datos de tamaño pequeño a mediano. La ventaja de k-means es su bajo costo de cálculo y el inconveniente es sensible a los datos ruidosos, mientras que k-medoid tiene un alto costo de cálculo y no es sensible a los datos ruidosos. La complejidad de tiempo de k-medias es significa  $O(i * k * m * n)$  y la complejidad de tiempo de k-medoid es  $O(ik(n-k)^2)$ .

En (Saunders, 1980), se examina los procesos de análisis de conglomerados, los beneficios de la segmentación y las aplicaciones que surgen como nuevas direcciones de investigación de mercados. El estudio se basa en una encuesta a 200 estudiantes de sexto grado, agrupados por sus preferencias en las diferentes carreras. Se emplearon comparaciones de similitud usando medidas de distancia euclidianas con el algoritmo k-medias. Los resultados son utilizados para identificar los perfiles de los estudiantes para brindarles asesoría personalizada sobre la elección de su futura carrera.

### 2.1.3 Métricas para los métodos de agrupamiento

La herramienta básica para los métodos de agrupamiento son medidas que miden la similitud o la disimilitud entre los objetos a ser agrupados. Para medir la cercanía entre los objetos, se toma como referencia otros objetos o alguna medida estadística de tendencia central (media, mediana, moda, medoide etc.). La decisión de la medida a usar en cualquier aplicación es a menudo una cuestión de una elección muy cuidadosa en la que resulta no sólo la importancia del tipo de variable (cuantitativo o cualitativo) sino del conocimiento previo del experto sobre el dominio de estudio. Existen dos tipos de medidas o índices de proximidad: los que miden la cercanía entre los objetos o similitud o los que miden la distancia entre los objetos o la disimilitud. Ambos índices son complementarios, lo cual quiere decir que cuanto mayor sea el índice de similitud entre dos objetos, menor sería su índice de disimilitud.

#### Medidas de similaridad

Las métricas para medir la similaridad, se conocen como distancia y se utiliza como sinónimo de métrica. Un valor mayor de distancia, indica que los objetos están más alejados. Generalmente, antes de aplicar estas medidas de distancia se realiza un proceso de estandarización o tipificación de los datos previamente, con la finalidad de eliminar la influencia de las unidades de medidas de las variables en el análisis. Sean  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$  y  $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$  dos vectores de observaciones de  $p$ -dimensional de variables, entonces algunas medidas de similitud son las siguientes:

#### Para variables cuantitativas:

- **Distancia euclidiana:** Es la más conocida y utilizada. Mide la distancia geométrica entre los vectores. Se define como:  $d(x_i, x_j) = \sqrt{\sum_{l=1}^p (x_{il} - x_{jl})^2}$
- **Distancia Manhattan:** Es la sumatoria en valor absoluto de las distancias entre las componentes de los vectores. Se define como:  $d(x_i, x_j) = \max |x_{ik} - x_{jk}|$ .
- **Distancia de Minkowski:** Se generaliza las dos anteriores, las cuales se obtienen, respectivamente, haciendo  $m = 2$  y  $m = 1$ . Se puede utilizar para todo valor real  $m \geq 1$ . Se

define como: 
$$d(x_i, x_j) = \left[ \sum_{k=1}^p (x_{ik} - x_{jk})^m \right]^{\frac{1}{m}}$$

- **Distancia Cheychev:** Es la sumatoria en valor absoluto de las distancias entre las componentes de los vectores. Se define como:  $d(x_i, x_j) = \max |x_{ik} - x_{jk}|$ .

- **Distancia de Mahalanobis:** Es la sumatoria en valor absoluto de las distancias entre las componentes de los vectores ponderando por la matriz de covariancias. Se define como:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^p (\bar{x}_i - \bar{x}_j)' S^{-1} (\bar{x}_i - \bar{x}_j)}$$

### Para variables cualitativas

- **Variables binarias.** Simétricas (ambos tienen el mismo peso). Se define como:

$$d(i, j) = \frac{r + s}{q + r + s + t}, \text{ donde: } q = \text{número de valores que son 1 en las dos, } r = \text{número de}$$

valores que son 1 en i y 0 en j, s = número de valores que son 0 en i y 1 en j, t = número de valores que son 0 en las dos.

No simétricas (el más importante y más raro vale 1). Se conoce como el coeficiente

Jaccard. Se define como:  $d(i, j) = \frac{r + s}{q + r + s}.$

- **Variables nominales.** Se define como:  $d(i, j) = \frac{p - m}{p},$  donde: m = número de valores

iguales, p = número total de casos. Se puede incluir pesos para darle mayor importancia a m. A partir de las nominales, se pueden crear nuevas variables binarias asimétricas.

- **Variables ordinales.** Son como las nominales pero con un orden relevante. El orden es importante, pero no la magnitud. Se tiene: 1) Cambia el valor de cada variable por el ranqueo:  $r_{if} \in \{1, 2, \dots, M_f\},$  donde:  $M_f$  es el índice del valor más alto de la variable. 2)

Mapear el ranqueo entre 0 y 1 para darle igual peso:  $Z_{if} = \frac{r_{if} - 1}{M_f - 1}.$  3) Usar cualquiera de las

medidas numéricas anteriores.

### Medidas de disimilitud

Entre las medidas se consideran las siguientes:

- **Disimilitud euclidiana cuadrada:** Aunque no es una métrica tiene dos ventajas importantes. Conduce a los mismos minimizadores que la distancia euclidiana y es más suavizada. Se define como:  $\delta(x_i, x_j) = \sum_{l=1}^p (x_{il} - x_{jl})^2$

### Distancias para variables mixtas.

Cuando se tienen datos de individuos que se han evaluado tanto variables cualitativas como cuantitativas, se define la distancia Gower de la siguiente manera:

$$S_{ij} = \frac{\sum_1^{n_1} (1 - \frac{|x_l - y_l|}{R_l}) + a + \alpha}{n_1 + (n_2 - d) + n_3},$$
$$d_{ij} = (1 - S_{ij})$$

donde:

$n_1$  es el número de variables cuantitativas continuas

$n_2$  es el número de variables binarias

$n_3$  es el número de variables cualitativas (no binarias)

$a$  es el número de coincidencias (1,1) en las variables binarias

$d$  es el número de coincidencias (0,0) en las variables binarias

$\alpha$  es el número de coincidencias en las variables cualitativas (no binarias)

$R_l$  es el rango (o recorrido) de la  $l$ -ésima variable cuantitativa

#### 2.1.4 Métodos para validar la calidad de los clúster

La calidad de los clústers formados pueden ser validados a través de calcular índices. Existen dos clases de validación: interna y externa.

**Validación interna:** La evaluación de los clústers se realiza sobre los mismos grupos generados. Entre los índices que se pueden obtener son:

- **Índice Davies-Bouldin:** El mejor algoritmo es el que produce el valor menor es el mejor.

$$\text{Se calcula por: } IDB = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \left( \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

Dónde:  $n$  = Número de clústers,  $c_x$  = Centro del clúster  $x$ ,  $\sigma_x$  = Distancia promedio de todos los elementos en el clúster  $x$  a su centroide  $c_x$  y  $d(c_i, c_j)$  = Es la distancia entre los centroides  $c_i$  y  $c_j$

- **Índice Dunn:** Busca identificar clúster densos y claramente separados. Grupos con

valores mayores del índice son mejores. Se calcula por:  $ID = \frac{\text{Min}_{1 \leq i < j \leq n} d(i, j)}{\text{Max}_{1 \leq k \leq n} d'(k)}$

Dónde:  $d(i, j)$  = Distancia entre clústers  $i$  y  $j$  (intra-clústers), que es la distancia entre centroides,  $d'(k)$  = Distancia intra-clúster del clúster  $k$ . Puede ser la máxima entre pares de elementos del clúster.

**Validación externa:** La evaluación se realiza entre grupos conocidos. La evaluación final generalmente la realiza una persona.

- **Índice de Pureza:** Mide en qué los clústers contienen una sola clase. Se calcula por:

$$IP = \frac{1}{N} \sum_{m \in M} \text{Max}_{d \in D} |m \cap d|$$

Dónde:  $M$  = número de clúster,  $D$  = Número de clases y  $N$  = Número de datos.

- **Índice de Rand:** Mide que tan parecidos son los clúster a las clases. Se calcula por:

$$IR = \frac{TP + TN}{TP + FP + TN + FN}$$

Dónde: TP=Verdaderos positivos, FP=Falsos positivos, TN=Verdaderos negativos y FN=Falsos negativos.

- **Índice de Jaccard.** Mide la similaridad entre dos grupos. Los elementos comunes entre dos grupo entre los elementos de los dos grupos. Se calcula por:

$$IJ(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}$$

Dónde: TP=Verdaderos positivos, FP=Falsos positivos, TN=Verdaderos negativos y FN=Falsos negativos.

- **Índice de Dice.** Similar al índice de Jaccars. Se calcula por:

$$ID(A, B) = \frac{2TP}{2TP + FP + FN}$$

### Árbol de clasificación C5.0

Los árboles de clasificación pertenecen a las técnicas para el aprendizaje supervisado. El árbol de clasificación se presenta como una estructura jerárquica de nodos para asociar una variable y ramas que representan las divisiones de los valores de las variables. El árbol de clasificación muestra en el primer nivel un solo nodo llamado “raíz”, un conjunto de nodos intermedios y nodos terminales llamadas “hojas” que refleja algún valor de del atributo clase y todos los nodos están interconectadas con ramas que representan decisiones o evaluaciones que permiten la división.

Existen muchos algoritmos desarrollados para construir árboles de clasificación, por lo cual su uso depende del tipo de variables y las especificaciones. Entre los más utilizados se tiene:

#### 1) Algoritmo ID3

El ID3 (Iterative Dichotomiser 3), es un algoritmo de árbol de decisión desarrollado por (Quinlan J. R., 1986). El ID3, es el algoritmo más simple y potente para construir un árbol, trabajando sólo con atributos nominales. El ID3 utiliza la **Ganancia de Información (GI)**, como la medida para generar los nodos (raíz, intermedio y hoja) del árbol, seleccionando aquel atributo que proporcione la mayor ganancia de información (menor entropía y menor incertidumbre del atributo). El ID3, utiliza la Entropía (menor valor, menor incertidumbre y mayor información proporciona el atributo para la clasificación) y la Ganancia de Información (ganancia por usar el atributo) como medidas de la bondad de los atributos.

## 2) Algoritmo C4.5

El algoritmo C4.5 fue propuesto por (Quinlan J. , 1993), siendo una extensión del ID3. El C4.5, utiliza como medida la **Razón de Ganancia** (RG) y criterio para ir seleccionando el atributo que dividirá el conjunto de entrenamiento que definirán los nodos del árbol. El C4.5 realiza la poda del árbol después de haberlo construido (post poda) posibilitando tener árboles más consistentes y evitando el problema de sobre ajuste. Puede trabajar con atributos nominales y continuos; pero los atributos continuos son convertidos en intervalos discretos (discretización automática) y puede construir árboles cuando existen datos faltantes (missing).

## 3) Algoritmo C5.0.

El algoritmo C5.0 es una extensión del algoritmo C4.5, que es también la extensión de ID3. El C5.0, puede ser aplicado a grandes volúmenes de datos (Big Data). El C5.0, supera al C4.5, en cuanto a la velocidad, la memoria y la eficiencia. El algoritmo C5.0, usa la máxima ganancia de información para dividir el conjunto de entrenamiento. El modelo C5.0 puede dividir las muestras en base a la información de campo de ganancia mayor. El subconjunto de la muestra que se obtiene de la primera división se dividirá después. El proceso continuará hasta que el subconjunto de la muestra no se puede dividir y por lo general es de acuerdo a otro campo. Por último, examinar la división de nivel más bajo, se rechazarán los subconjuntos de la muestra que no tienen notable contribución al modelo. C5.0 se maneja fácilmente el atributo de valor múltiple y falta un atributo de conjunto de datos. Entre las ventajas que ofrece el algoritmo C5.0 son:

- **Velocidad.** El C5.0 es significativamente más rápido que el C4.5 (varios órdenes de magnitud)
- **El uso de memoria.** El C5.0 es más eficiente el uso de la memoria que el C4.5
- **Precisión.** El conjunto de reglas de decisión obtenidas con el C5.0, presentan menores tasas de error que con el C4.5.
- **Árboles de decisión más pequeños.** El C5.0 obtiene resultados similares a C4.5 con árboles de decisión mucho más pequeños.
- **Soporte para Boosting.** Permite la aplicación de multclasificadores (ensambladores) como Boosting, mejorando el aprendizaje con el árbol y aumentando la precisión (tasa buena clasificación) del C5.0.
- **Ponderación.** El C5.0 permite ponderar los distintos casos y tipos de errores de clasificación.

- **Winnowing.** Es una opción automática del C5.0, que permite winnows los atributos para eliminar aquellos que pueden ser de poca relevancia.

## 2.2 El método de partición k-Medoides

El k-Medoides, es un método de agrupamiento basado en particiones que divide el conjunto de datos en grupos buscando minimizar la distancia entre los objetos que se van añadiendo a un grupo con respecto a un centroide. En comparación del k-medias que usa como centroide la media del grupo, el k-Medoides usa el punto medio del grupo (Medoide). El k-Medoides es una técnica de partición de grupos que divide los datos conformados por  $n$  objetos en  $k$  grupos (con  $k$  conocido de antemano). Es más robusto ante datos atípicos y el ruido y se basa en minimizar la suma de disimilaridades entre un objeto y el Medoide, a diferencia de k-medias que minimiza la suma de distancias euclidianas cuadradas. Un medoide puede ser definido como el objeto de un grupo cuya disimilaridad media a todos los objetos en el grupo es mínima. Es el punto ubicado más hacia el centro en todo el grupo. Entre los algoritmos para el método de k-Medoides se encuentran: PAM (Partitioning Around Medoids), CLARA (Clustering for LARge Applications) y CLARANS (CLARa based on RANdomized Search).

### 2.2.1 Determinación del número óptimo de clústers

- **Método de silueta promedio.** Para estimar el número óptimo de clústers, se usa el método de silueta promedio. La idea es calcular el algoritmo PAM utilizando valores diferentes de los conglomerados  $k$ . A continuación, la silueta de clústers promedio se dibuja según la cantidad de clústeres. La silueta promedio mide la calidad de una agrupación. Un alto ancho de silueta promedio indica un buen agrupamiento. El número óptimo de conglomerados  $k$  es el que maximiza la silueta promedio sobre un rango de valores posibles para  $k$  (Kaufman & Rousseeuw, 1990).

El índice de silueta es una métrica para evaluar el buen funcionamiento de los algoritmos de aprendizaje no supervisado. El objetivo de este índice es identificar el número óptimo de agrupamientos. En los algoritmos de aprendizaje no supervisado, el número de clústeres puede ser un parámetro de entrada del algoritmo (K-means) o determinado automáticamente por el algoritmo (DBSCAN). En el primer caso la determinación del número óptimo de clústeres tiene que ser realizado mediante alguna medida externa al algoritmo. El índice silueta es indicador del número ideal de clústeres. Un valor más alto



de este índice indica un caso más deseable del número de clústeres (Rousseeuw, 1987). El coeficiente de Silueta como una medida de cada objeto de la muestra,  $S(i)$ . El coeficiente de Silueta para un objeto se define:

$$S(i) = \frac{b-a}{\max(a,b)}$$

Dónde:

a = Es la distancia media entre el objeto y todos los otros objetos de la misma clase

b = Es la distancia media entre el objeto y todos los otros objetos del clúster más próximo.

El valor de  $S(i)$  puede ser obtenido combinando los valores de  $a(i)$  y  $b(i)$  como se muestra a continuación:

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{si } a(i) < b(i) \\ 0, & \text{si } a(i) = b(i) \\ \frac{a(i)}{b(i)} - 1, & \text{si } a(i) > b(i) \end{cases}$$

De la definición anterior se obtiene que el índice silueta es:  $-1 \leq s(i) \leq 1$ . Para que el valor de  $s(i)$  sea próximo a uno entonces  $a(i) \ll b(i)$ . Como  $a(i)$  es una medida de cómo de similar es  $i$  a su propio cluster, entonces esto implica que el objeto  $i$  es muy similar a los otros objetos de su clúster. Además se requiere que  $b(i)$  sea grande. Esto implica que el objeto  $i$  no es similar a los objetos del clúster más próximo. Un valor de  $s(i)$  cercano a cero indica que el objeto  $i$  está en la frontera de dos clusters. Por el contrario si el valor de  $s(i)$  es negativo, entonces dicho objeto debería ser asignado al cluster más cercano.

El valor promedio de  $s(i)$  sobre todos los datos de un clúster es una medida de cómo compactamente agrupados están los datos en el clúster. Así el promedio de  $s(i)$  sobre el conjunto de datos es una medida de cómo apropiadamente se han agrupado los datos. De forma que si el número de clústeres elegidos es demasiado bajo o alto, entonces el gráfico de silueta será más estrecho que el resto.

### **Medición para valores del índice:**

0.71-1.0, las estructuras encontradas son sólidas.

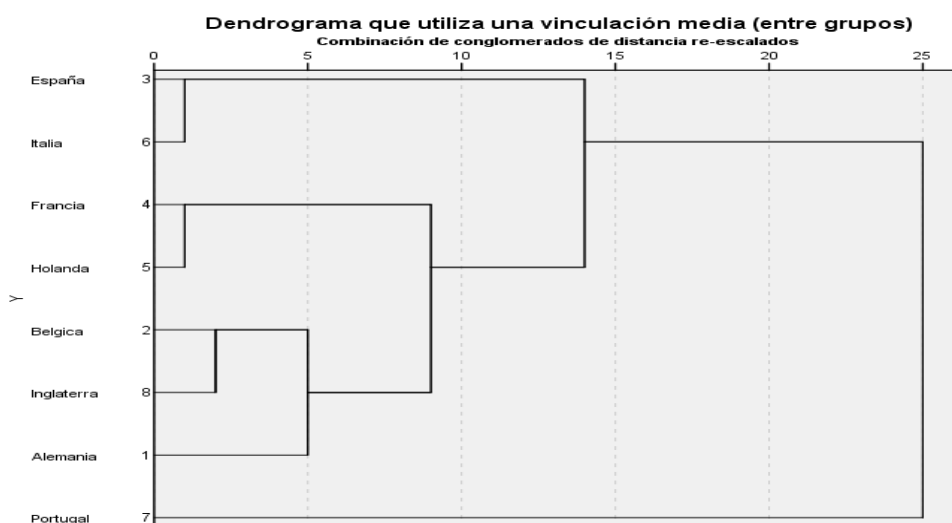
0.51-0.70, las estructuras encontradas con razonables.

0.26-0.50, las estructuras encontradas son débiles y tienden a ser artificiales. Se deberían intentar métodos alternativos para el análisis de los datos.

<0.25, no se encuentran estructuras

- **Dendrograma.** Es un método gráfico que muestra la secuencia de la formación de los grupos o clúster que se están conformando según la medida de distancia usada. Se basa en representar en una gráfica el número de clústeres que se observan en los distintos niveles del dendrograma frente a los niveles de fusión a los que los clústeres se unen en cada nivel. La presencia de una pendiente poco pronunciada sugiere que la siguiente unión de clústeres no aporta apenas información adicional sobre la aportada en el nivel anterior. Este método, por lo tanto, se basa en la existencia de pequeños saltos o discontinuidades en los niveles de fusión.

**Ejemplo.** Se cuenta con información sobre 8 países de la Comunidad Europea respecto a la Natalidad, Mortalidad y Mortalidad Infantil. El objetivo es agrupar a los países de Europa de acuerdo a las variables consideradas.



**Figura 1. Dendrograma**

- **Pseudo Estadística F (Beale, 1969).** Se propuso el uso de un contraste basado en la distribución F de Snedecor para probar la hipótesis de la existencia de  $C_2$  clústeres frente a la existencia de  $C_1$  clústeres, siendo  $C_2 > C_1$ . Para esto se consideran para para cada partición las desviaciones cuadráticas medias de los elementos de cada clúster a su centroide (DC1 y DC2):

$$DC_1 = \frac{1}{n - c_1} \sum_{i=1}^{c_1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad \text{y} \quad DC_2 = \frac{1}{n - c_2} \sum_{i=1}^{c_2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

donde  $n_1$  es el número de elementos del clúster  $C_1$ ,  $n_2$  del clúster  $C_2$  y  $n$  es el total de la muestra. El estadístico de prueba es:

$$F_c = \frac{DC_1 - DC_2}{DC_2} \approx F_{(p(c_2 - c_1), p(n - c_2))} \left[ \frac{\left( \frac{n - c_1}{n - c_2} \right) \left( \frac{c_2}{c_1} \right)^{\frac{2}{p}} - 1}{2} \right]$$

Si la prueba resulta significativa, indica que la división en  $C_2$  clústeres representa una agrupación mejor frente a la división en  $C_1$  clústeres.

### 2.2.2 Algoritmo de Partición Alrededor de Medoides (PAM)

Para aplicar el método de k-medoids, el algoritmo más usado se conoce como Partición Alrededor de Medoids (PAM). El PAM utiliza una búsqueda golosa que puede no encontrar la solución óptima, pero es más rápido que la búsqueda exhaustiva. Para encontrar k clusters (agrupamientos), el modelo PAM determina un objeto representativo para cada clúster. Este objeto representativo, llamado medoid, es el que se encuentra localizado más al centro dentro del clúster. Una vez que los medoids han sido seleccionados, cada objeto no seleccionado es agrupado con el medoid al cual es más similar. Para encontrar los k medoids, PAM comienza con una selección arbitraria de k objetos. En cada iteración, un intercambio entre un objeto seleccionado  $O_i$  y un objeto no seleccionado  $O_j$  es realizada, si y solo sí, el intercambio resulta en un incremento de la calidad del agrupamiento (clustering). El algoritmo puede ser descrito:

1. Inicialización. Seleccionar  $k$  objetos de los  $n$  puntos como el medoid inicial.
2. Sea  $O_j$  un objeto no seleccionado y  $O_i$  es un medoid (objeto seleccionado). Se calcula la medida de disimilaridad o distancia:  $d_{ij} = d(O_j, O_i)$
3. Se indica que  $O_j$  pertenece al clúster representado por  $O_i$ , si  $d(O_j, O_i) = \text{Min\_medoids}(O_j, O_e)$ , es valor el mínimo sobre todos los medoids  $O_e$
4. Se calcula la ganancia total obtenida seleccionando el objeto  $j$ :  $GT_j = \sum_i d_{ji}$ .
5. Si el costo total de la configuración aumentó, deshacer el intercambio e ir al paso 2). De lo contrario determinar los nuevos medoides, ir al paso 2).
6. Se termina cuando no hay más objetos que agrupar.

## **Ventajas y desventajas del algoritmo PAM**

Las principales ventajas que se puede indicar con el método PAM son las siguientes:

- El algoritmo PAM, es una alternativa robusta frente a los k-medias para dividir un conjunto de datos en grupos. (Kassambara, 2017)
- Los clusters formados por el algoritmo PAM tienen por medoide a una observación que tiene la menor distancia con respecto a las demás observaciones de su cluster y la mayor distancia posible con otros medoides, asegurando grupos más homogéneos.

Las principales desventajas que se puede indicar con el método PAM son las siguientes:

- PAM requiere que el usuario conozca los datos e indique el número apropiado de clústeres a producir.
- El algoritmo PAM cuando se trabaja con grandes conjuntos de datos puede necesitar demasiada memoria o demasiado tiempo de cálculo en la computadora. En este caso, la función CLARA es una mejor alternativa.

### 2.2.3 Ejemplo de aplicación del algoritmo PAM

En primer lugar se calcula la distancia con la métrica Gower para la aplicación del algoritmo PAM. En el Cuadro 1 se presenta los datos de 10 recién nacidos y 6 variables.

- Sexo: 1=niña, 0 = Niño
- Tiempo de gestación: 1=más de 35 semanas, 0=menos de 35 semanas
- Grupo Sanguíneo: 1=O, 2=A, 3=B, 4=AB
- Raza: 1=Blanca, 2=Negra, 3=Otros

**Cuadro 1. Muestra de 10 individuos recién nacidos**

Niño	Altura	Peso	Sexo	Tiempo de gestación	Grupo Sanguíneo	Raza
s1	49.3	4.2	1	1	1	1
s2	51.2	2.1	0	1	1	1
s3	52.3	4.1	0	1	3	2
s4	50.1	3.8	0	0	1	1
s5	47.6	2.9	1	1	2	2
s6	55.0	4.5	0	1	1	1
s7	51.3	4.0	1	1	1	2
s8	52.7	3.9	0	1	4	1
s9	49.4	2.9	1	0	2	1
s10	52.1	3.8	1	1	1	3

Fuente Elaboración propia

**Distancia Gower.** La distancia para variables mixtas se calcula con la métrica Gower. Las expresiones son:

$$S_{ij} = \frac{\sum_1^n (1 - \frac{|x_l - y_l|}{R_l}) + a + \alpha}{n_1 + (n_2 - d) + n_3},$$

$$d_{ij} = (1 - S_{ij})$$

dónde:

$n_1$  es el número de variables cuantitativas continuas

$n_2$  es el número de variables binarias

$n_3$  es el número de variables cualitativas (no binarias)

$a$  es el número de coincidencias (1,1) en las variables binarias

$d$  es el número de coincidencias (0,0) en las variables binarias

$\alpha$  es el número de coincidencias en las variables cualitativas (no binarias)

$R_l$  es el rango (o recorrido) de la l-ésima variable cuantitativa

Como ejemplo se calcula la distancia de  $d_{s1,s5}^2$  y  $d_{s5,s7}^2$ :

Se empieza por calcular los parámetros necesarios como se muestra en el Cuadro 2

**Cuadro 2. Cálculo de los parámetros para el cálculo de Gower**

	S1 S5 Altura	S1 S5 Peso	S5 S7 Altura	S5 S7 Peso
$R_l$	7.4	2.4	7.4	2.4
$ x_l - y_l $	1.7	1.3	3.7	1.1
$n_1$	2		2	
$n_2$	2		2	
$d$	0		0	
$n_3$	2		2	
$a$	2		2	
$\alpha$	0		1	

**Fuente: Elaboración propia**

Reemplazando:

$$S_{s1,s5} = \frac{(1 - \frac{|49.3 - 47.6|}{55 - 47.6}) + (1 - \frac{|4.2 - 2.9|}{4.5 - 2.1}) + 2 + 0}{2 + (2 - 0) + 2} = 0.5381006$$

$$d_{s1,s5} = (1 - 0.5381006) \text{ y } d_{s1,s5} = 0.4618994$$

$$S_{s5,s7} = \frac{(1 - \frac{|47.6 - 51.3|}{55 - 47.6}) + (1 - \frac{|2.9 - 4|}{4.5 - 2.1}) + 2 + 1}{2 + (2 - 0) + 2} = 0.6736111$$

$$d_{s5,s7} = (1 - 0.6736111) \text{ y } d_{s5,s7} = 0.3263889$$

De esta manera, se calcula el resto de distancias. En el Cuadro 3 se presenta la matriz de distancias con la métrica Gower para los 10 individuos.

**Cuadro 3. Matriz de distancias de Gower**

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
<b>s1</b>	0.0000									
<b>s2</b>	0.3553	0.0000								
<b>s3</b>	0.5745	0.5964	0.0000							
<b>s4</b>	0.3791	0.3714	0.6845	0.0000						
<b>s5</b>	0.4619	0.6366	0.5225	0.7855	0.0000					
<b>s6</b>	0.3159	0.3027	0.5063	0.3908	0.7778	0.0000				
<b>s7</b>	0.2256	0.4675	0.3628	0.5409	0.3264	0.4514	0.0000			
<b>s8</b>	0.4307	0.3905	0.4275	0.4786	0.6843	0.3122	0.5385	0.0000		
<b>s9</b>	0.4259	0.5961	0.8153	0.4939	0.3739	0.7372	0.6192	0.6438	0.0000	
<b>s10</b>	0.2575	0.4717	0.5253	0.5450	0.4972	0.4473	0.1986	0.5205	0.6233	0.0000

**Fuente: Elaboración propia con R-Studio, Paquete StatMatch**

**Demostración del método Partición alrededor de medoides PAM**

Una vez que se obtuvieron las distancias para las 10 observaciones de los recién nacidos se resolvió el algoritmo del método Partición alrededor de medoides PAM, el cual es un método iterativo que se explica a continuación:

**Paso 1.** Se estableció dividir las 10 observaciones en 2 grupos ( $k=2$ ) y ya que se cuenta con la matriz de Gower podemos fijar como medoides iniciales *S1* y *S2* que vendrían a ser las observaciones 1 y la observación 2, con las cuales se construyó un primer agrupamiento comparando el menor valor de distancia que resultó de comparar las dos columnas seleccionadas para luego ser resaltadas en negrita determinando así el posible clúster a donde irá la observación. También se pudo calcular el Costo Total igual a 2.75507, el cual es explicado en el siguiente Cuadro 4

**Cuadro 4. Matriz de distancias de Gower (observaciones *S1* y *S2*)**

	<b>S1</b>	<b>S2</b>	S3	S4	S5	S6	S7	S8	S9	S10	Clúster
<b>S1</b>	<b>0.0000</b>	0.3553	0.5745	0.3791	0.4619	0.3159	0.2256	0.4307	0.4259	0.2575	1
<b>S2</b>	0.3553	<b>0.0000</b>	0.4970	0.3095	0.6366	0.2523	0.4675	0.3255	0.5961	0.4717	2
<b>S3</b>	0.5745	<b>0.4970</b>	0.0000	0.5704	0.5225	0.4219	0.3628	0.3562	0.8153	0.5253	2
<b>S4</b>	0.3791	<b>0.3095</b>	0.5704	0.0000	0.7855	0.3256	0.5409	0.3988	0.4116	0.5450	2
<b>S5</b>	<b>0.4619</b>	0.6366	0.5225	0.7855	0.0000	0.7778	0.3264	0.6843	0.3739	0.4972	1
<b>S6</b>	0.3159	<b>0.2523</b>	0.4219	0.3256	0.7778	0.0000	0.4514	0.2601	0.7372	0.4473	2
<b>S7</b>	<b>0.2256</b>	0.4675	0.3628	0.5409	0.3264	0.4514	0.0000	0.5385	0.6192	0.1986	1
<b>S8</b>	0.4307	<b>0.3255</b>	0.3562	0.3988	0.6843	0.2601	0.5385	0.0000	0.6438	0.5205	2
<b>S9</b>	<b>0.4259</b>	0.5961	0.8153	0.4116	0.3739	0.7372	0.6192	0.6438	0.0000	0.6233	1
<b>S10</b>	<b>0.2575</b>	0.4717	0.5253	0.5450	0.4972	0.4473	0.1986	0.5205	0.6233	0.0000	1
Costo	1.3709	1.3842									
Costo total	<b>2.7551</b>										

**Paso 2** A continuación se presenta el cuadro 5 la iteración de *S1* y *S3* con el algoritmo PAM con su respectivo Costo Total igual a 2.7774 y la nueva agrupación que se obtendría si estos fueran nuestros medoides finales.

**Cuadro 5. Matriz de distancias de Gower (observaciones S1 y S3)**

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Clúster
<b>S1</b>	<b>0.0000</b>	0.3553	0.5745	0.3791	0.4619	0.3159	0.2256	0.4307	0.4259	0.2575	1
<b>S2</b>	<b>0.3553</b>	0.0000	0.4970	0.3095	0.6366	0.2523	0.4675	0.3255	0.5961	0.4717	1
<b>S3</b>	0.5745	0.4970	<b>0.0000</b>	0.5704	0.5225	0.4219	0.3628	0.3562	0.8153	0.5253	2
<b>S4</b>	<b>0.3791</b>	0.3095	0.5704	0.0000	0.7855	0.3256	0.5409	0.3988	0.4116	0.5450	1
<b>S5</b>	<b>0.4619</b>	0.6366	0.5225	0.7855	0.0000	0.7778	0.3264	0.6843	0.3739	0.4972	1
<b>S6</b>	<b>0.3159</b>	0.2523	0.4219	0.3256	0.7778	0.0000	0.4514	0.2601	0.7372	0.4473	1
<b>S7</b>	<b>0.2256</b>	0.4675	0.3628	0.5409	0.3264	0.4514	0.0000	0.5385	0.6192	0.1986	1
<b>S8</b>	0.4307	0.3255	<b>0.3562</b>	0.3988	0.6843	0.2601	0.5385	0.0000	0.6438	0.5205	2
<b>S9</b>	<b>0.4259</b>	0.5961	0.8153	0.4116	0.3739	0.7372	0.6192	0.6438	0.0000	0.6233	1
<b>S10</b>	<b>0.2575</b>	0.4717	0.5253	0.5450	0.4972	0.4473	0.1986	0.5205	0.6233	0.0000	1
Costo	2.4212		0.3562								
Costo total	<b>2.7774</b>										

**Paso 3.** Después de realizar todas las posibles iteraciones para formar dos clústers obtuvimos el menor Costo Total igual a 2.5589 tomando como medoides las observaciones S4 y S7. Por lo tanto estos serán nuestros medoides finales con los que se construyeron los 2 clústers para el ejemplo de los 10 individuos recién nacidos, el cual se detalla en el siguiente Cuadro 6.

**Cuadro 6. Matriz de distancias de Gower (observaciones S4 y S7)**

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Clúster
<b>S1</b>	0.0000	0.3553	0.5745	0.3791	0.4619	0.3159	<b>0.2256</b>	0.4307	0.4259	0.2575	2
<b>S2</b>	0.3553	0.0000	0.4970	<b>0.3095</b>	0.6366	0.2523	0.4675	0.3255	0.5961	0.4717	1
<b>S3</b>	0.5745	0.4970	0.0000	0.5704	0.5225	0.4219	<b>0.3628</b>	0.3562	0.8153	0.5253	2
<b>S4</b>	0.3791	0.3095	0.5704	<b>0.0000</b>	0.7855	0.3256	0.5409	0.3988	0.4116	0.5450	1
<b>S5</b>	0.4619	0.6366	0.5225	0.7855	0.0000	0.7778	<b>0.3264</b>	0.6843	0.3739	0.4972	2
<b>S6</b>	0.3159	0.2523	0.4219	<b>0.3256</b>	0.7778	0.0000	0.4514	0.2601	0.7372	0.4473	1
<b>S7</b>	0.2256	0.4675	0.3628	0.5409	0.3264	0.4514	<b>0.0000</b>	0.5385	0.6192	0.1986	2
<b>S8</b>	0.4307	0.3255	0.3562	<b>0.3988</b>	0.6843	0.2601	0.5385	0.0000	0.6438	0.5205	1
<b>S9</b>	0.4259	0.5961	0.8153	<b>0.4116</b>	0.3739	0.7372	0.6192	0.6438	0.0000	0.6233	1
<b>S10</b>	0.2575	0.4717	0.5253	0.5450	0.4972	0.4473	<b>0.1986</b>	0.5205	0.6233	0.0000	2
Costo				1.4456			1.1134				
Costo total	<b>2.5589</b>										



### III. MATERIALES Y MÉTODOS

Los materiales y métodos que se usaron en la presente investigación son los siguientes.

#### 3.1 Materiales

- 1) Una computadora personal Intel® Core™ i7. CPU 3.5 GHz. RAM de 4.00 GB.
- 2) Programa estadístico R. Se descarga software R versión x64 3.3.4.

##### Los paquetes:

- Cluster. Para hallar las agrupaciones y la librería para aplicar el algoritmo PAM
- Factoextra. Para obtener el gráfico e índice de silueta.
- StatMatch. Para hallar la matriz de distancias de Gower

##### Las funciones:

- Realizar el método de silueta. fviz\_nbclust
- Ejecutar el algoritmo PAM. Pam
- Asignar columna de grupo de pertenencia
- Acceder a los medoides y cluster. pam.X\$medoids
- Calculo de distancia para datos mixtos. daisy

- 3) Una impresora inyectora HP.

#### 3.2 Metodología

##### 3.2.1 Tipo de investigación y formulación de hipótesis

La investigación fué experimental con diseño descriptivo. Esto debido a que se obtuvieron los datos de los clientes de los casinos para un mes en específico (octubre) describiendo las distintas agrupaciones que hay de los clientes debido a las características que estos poseen.

## **Tipo de investigación.**

Investigación descriptiva, la cual consiste en llegar a conocer las situaciones, costumbres y actitudes predominantes a través de la descripción exacta de las actividades de los clientes del casino. Su meta no se limita a la recolección de datos, sino a la predicción e identificación de las relaciones que existen entre dos o más variables.

## **Formulación de hipótesis**

Existen grupos de clientes con características similares determinados por la edad, el tipo de cliente, número de visitas, sexo, etc. de modo que las estrategias de marketing puedan ser enfocadas de una manera adecuada a cada grupo, de esta manera la casa de juegos podrá tener una mejor llegada a sus jugadores ofreciéndoles mejores productos e incentivándoles a participar más del casino.

### **3.2.2 Población y Muestra**

**Población.** Todos los clientes de un casino de la ciudad de Trujillo que hicieron uso del juego de tragamonedas en el mes de octubre del año 2016.

**Muestra.** 2798 clientes de un casino de la ciudad de Trujillo que hicieron uso del juego de tragamonedas en el mes de octubre del año 2016

### **3.2.3 Identificación de variables**

Los datos con los cuales se realizó la aplicación del algoritmo de Partición Alrededor de Medoides (PAM) fueron obtenidos de la base de datos de clientes de una empresa que cuenta con un casino, específicamente de su casino en la ciudad de Trujillo.

La recopilación de la información fué mediante las tarjetas personalizadas y la obtención del conjunto de datos será realizada por el sistema informático WIGOS, que es un sistema con el cual cuenta el casino para el manejo de la información que es generada directamente de sus máquinas tragamonedas, ruletas y mesas de juego. Este sistema permitió obtener datos del jugador como:

- Datos personales
- Hábitos de consumo
- Frecuencia de visitas

- Nivel del jugador (Tipo de tarjeta personalizada)

### 3.2.4 Proceso para análisis de conglomerados

Las etapas propuestas para realizar el análisis de conglomerados a los datos de los clientes del casino fueron:

#### 1) Recopilación de datos

La recopilación de la información fue mediante las tarjetas personalizadas que tienen los jugadores a la hora de ir al casino y la obtención del conjunto de datos se realizó por el sistema informático WIGOS, que es un sistema con el cual cuenta el casino para el manejo de la información que es generada directamente de sus máquinas tragamonedas, ruletas y mesas de juego, este sistema permitió obtener datos del jugador.

#### 2) Pre procesamiento de datos

Con la finalidad de preparar los datos para aplicar el análisis de agrupamiento con el algoritmo PAM y obtener una segmentación de los clientes del casino, se aplicaron técnicas de pre procesamiento para la limpieza (manejos de datos atípicos y datos faltantes) y la transformación de los datos (discretización y recodificación de datos).

- **Manejo de datos faltantes.** Para los datos faltantes se procedió a eliminar aquellos datos missing y con respuesta No Sabe en las variables independientes.
- **Manejo de datos atípicos.** Se aplicó el procedimiento de diagrama de cajas. Se calcularon los límites de seguridad inferior y superior para la variable recargas y luego se eliminaron los datos outliers inferiores y superiores respectivamente.
- **Codificación de datos.** Para la variable sexo se le asignó 0=Masculino y 1=Femenino; para la variable Tipo de cliente se le asignó 1=Classic, 2=Silver y 3=Gold.

#### 3) Análisis de agrupamiento aplicando el algoritmo PAM

- Seleccionar la medida de distancia. Se usó una medida de distancia para datos mixtos. En este caso se usó la medida de Gower.
- Estimación del número óptimo de grupos (k). Se utilizó el índice de silueta.
- Se realizó el agrupamiento con el algoritmo PAM.

- Obtener los resultados. Identificación del clúster asignado a cada cliente, medidas de los medoides, etc.
- Se realizó el análisis y discusión de los resultados encontrados

#### **4) Validación de los resultados encontrados**

- Se realizó el análisis de variancia con los k grupos
- Aplicación de alguna técnica de clasificación para corroborar los grupos hallados

#### **5) Caracterización de los grupos formados**

- Obtener medidas estadísticas para cada segmento
- Realizar la tipificación de cada segmento

## **IV. RESULTADOS Y DISCUSIÓN**

En la presente investigación se aplicó el método de agrupamiento basado en partición denominado k-Medoides con el algoritmo PAM, con la finalidad de segmentar a los clientes de un casino usando los datos originados al usar sus tarjetas personalizadas en las máquinas de tragamonedas. Para obtener los resultados que satisfagan los objetivos propuestos se aplicó el proceso de análisis de agrupamiento mencionado en la sección 3.2.4. En el Anexo se presenta la sintaxis desarrollado con el paquete estadístico R para obtener los resultados mencionados.

### **4.1 Recopilación de datos**

La recopilación de los datos para la aplicación fueron obtenidos a través de las tarjetas personalizadas que usan los clientes para hacer uso de los juegos del casino (máquinas tragamonedas, ruletas y mesas de juego) y que son almacenados por el sistema informático WIGOS que es un sistema con el cual cuenta el casino para el manejo de la información que es generada diariamente. Inicialmente se recopiló los datos de 2798 clientes que jugaron en el mes de octubre del 2016. Se almacenó en una hoja de cálculo con 25 variables.

### **4.2 Pre procesamiento de datos**

Se aplicaron técnicas de pre procesamiento de datos para su limpieza (manejos de datos atípicos y datos faltantes) y la transformación de los datos (estandarización). Esto permitió obtener una base de datos depurada para aplicar el algoritmo PAM.

#### **▪ Manejo de datos faltantes**

Se procedió a eliminar aquellas variables que contengan datos faltantes, no se imputaron estas variables debido a que casi toda la variable estaba vacía; un ejemplo de esto se evidencia con las variables profesión, sueldo, dirección, salario las cuales al momento de inscribirse no eran un campo obligatorio. Por esta razón los clientes no se sentían obligados a brindar esta información y en muchos casos tampoco estaban dispuestos a brindar dichos datos. De las 20 variables originales se retuvieron 8 variables.

X1 Monto acumulado jugado (soles)

X2 Monto que recarga (soles)

X3 Número de visitas al mes

X4 Número de días sin actividad

- X5 Edad del cliente
- X6 Tiempo en sala
- X7 Nivel (Classic, Silver, Gold)
- X8 Sexo

▪ **Manejo de datos atípicos**

Se aplicó el procedimiento de diagrama de cajas. Se calcularon los límites de seguridad inferior y superior para la variable monto de recarga y luego se eliminaron los datos atípicos inferiores y superiores respectivamente. En el Cuadro 7, se presentan los cuartiles para la variable monto de recarga.

**Cuadro 7. Cuartiles para el monto de recarga**

Percentiles	25	190,00
	50	560,00
	75	1771.0

**Cálculo de los límites de seguridad inferior y superior:**

$$LI = 190.0 - 1.5 * (1771.0 - 190.0) = - 2438.0$$

$$LS = 1771.0 + 1.5 * (1771.0 - 190.0) = 4142.5$$

Aplicando el límite superior, se eliminaron los datos atípicos cuyos montos de recarga son mayores a S/. 4142.5. Se eliminaron 97 datos, quedando 2701.

▪ **Transformación de datos**

En los análisis de agrupamiento es recomendable que las variables sean estandarizadas, con la finalidad de eliminar sus unidades de medida que pueden influenciar en la métrica de agrupamiento. Todas las variables cuantitativas fueron estandarizadas conformando el nuevo conjunto de datos para aplicar el algoritmo PAM.

**4.3 Análisis de agrupamiento con el algoritmo PAM**

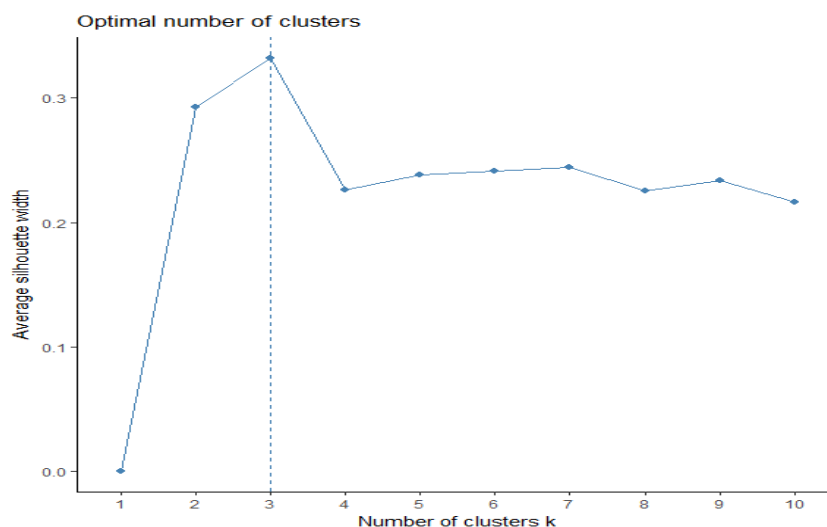
Se desarrolló el proceso de agrupamiento con el algoritmo PAM, con la finalidad de obtener una segmentación de los clientes del casino.

- **Seleccionar la medida de distancia**

En este caso, se usó la medida de Gower por tratarse de variables mixtas (cuantitativas y cualitativas) que es lo más recomendable. Se obtuvo una matriz de distancia la cual muestra las disimilitudes entre las observaciones.

- **Estimación del número óptimo de grupos (k)**

Para aplicar el algoritmo PAM se requiere conocer a priori el número de grupos (k) a formarse. Es un aspecto muy importante identificar el número de clúster más adecuado, puesto que de ello se puede obtener un agrupamiento consistente o no. Se han propuesto varios procedimientos o técnicas para seleccionar el número de clúster. En este caso, se utilizó el índice de silueta. Este índice es una métrica para evaluar el buen funcionamiento de los métodos de agrupamiento.



**Figura 2. Gráfico de la silueta con el algoritmo PAM**

Según la Figura 2 se muestra el resultado de aplicar índice de silueta con el algoritmo PAM. El índice de silueta identifica tres clúster (k=3) como el número óptimo de grupos que se pueden formar.

- **Realizar el agrupamiento con el algoritmo PAM**

Identificado como k=3, el número óptimo de grupos, se procedió a realizar el análisis de agrupamiento con el algoritmo PAM. La sintaxis en el programa R se puede ver en el Anexo1. Los resultados se presentan a continuación:

En el Cuadro 8 se presenta los valores de los medoides finales que fueron obtenidos con el algoritmo PAM para cada uno de los tres clúster, indicando el número de la observación correspondiente a la posición del medoide. Estos medoides corresponden al mejor (menor) costo total conseguido al momento de elegir los medoides con el PAM.

**Cuadro 8. Medoides finales con el algoritmo PAM**

Clúster	Nº Observación
1	863
2	734
3	2164

En el Cuadro 9, se presenta la distribución de los clientes del casino resultante del algoritmo PAM. En el clúster 1 se encuentra la mayor parte agrupada de los clientes, con un 49.4%, le sigue el clúster 2 con un 39.4% y último el clúster 3 que tiene el 11.3%.

**Cuadro 9. Distribución de número y porcentaje de clientes con el PAM**

Clúster	Número	Porcentaje
Clúster 1	1333	49.4
Clúster 2	304	11.3
Clúster 3	1064	39.4
Total	2701	100.0

#### **4.4 Validación del agrupamiento con el algoritmo PAM**

Con la finalidad de validar el agrupamiento de los clientes conseguidos con el algoritmo PAM, se propuso usar dos procedimientos (el análisis de variancia y matriz de confusión).

- **Realizar el análisis de variancia**

Se aplica en análisis de variancia considerando como factor la clasificación de los clúster. Esto es, los tres clúster pasan a ser los niveles del factor. Se espera bajo la hipótesis que las pruebas para las variables cuantitativas sean significativas, por lo cual se rechaza la hipótesis de una igualdad de medias, lo que indicaría que la variancia entre los grupos es mayor dentro de los grupos; que en términos del análisis de agrupamiento se está cumpliendo que existe una heterogeneidad entre grupos y homogeneidad dentro de grupos. En el cuadro 10 se presenta



los ANVAs para las 6 variables cuantitativas, se aprecia que todas las pruebas F resultaron significativas. Este resultado por lo antes mencionado, valida los tres clúster conseguidos con el algoritmo PAM y su respectiva consistencia.

**Cuadro 10. Análisis de Varianza para las variables agrupadas por clúster**

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
Jugado	Inter-grupos	188202440528	2	94101220264	2796.5	.000
	Intra-grupos	90787031405	2698	33649752		
	Total	278989471933	2700			
Recargas	Inter-grupos	5352995413	2	2676497707	2099.8	.000
	Intra-grupos	3438951014	2698	1274630		
	Total	8791946427	2700			
Visitas	Inter-grupos	62055	2	31027	550.5	.000
	Intra-grupos	152069	2698	56		
	Total	214123	2700			
Sin actividad	Inter-grupos	166048	2	83024	3598.4	.000
	Intra-grupos	62250	2698	23		
	Total	228299	2700			
Edad	Inter-grupos	3801	2	1901	9.6	.000
	Intra-grupos	534786	2698	198		
	Total	538587	2700			
Tiempo de visita	Inter-grupos	2504558	2	1252279	237.4	.000
	Intra-grupos	14231145	2698	5275		
	Total	16735703	2700			

▪ **Prueba Chi cuadrado para variables cualitativas**

Se propuso aplicar la prueba de chi-cuadrado para analizar la asociación de las variables cualitativas.

**Cuadro 11. Tabla de contingencia para estudiar la asociación entre sexo y la variable clúster**

	Clúster 1	Clúster 2	Clúster 3
Hombre	897	179	683
Mujer	436	125	381

Prueba Chi-cuadrado: p-valor= 0.01514

**Cuadro 12. Tabla de contingencia para estudiar la asociación entre nivel y la variable clúster**

	Clúster 1	Clúster 2	Clúster 3
Classic	1332	1	1062
Silver	1	303	2

Prueba Chi-cuadrado: p-valor= 2.2e-16

De acuerdo a los resultados del cuadro 11 y del cuadro 12, si hay asociación entre las variables sexo y nivel con la variable clúster respectivamente; es decir clúster y sexo están asociados al igual que clúster y nivel.

▪ **Aplicación de una técnica de clasificación**

Se propuso aplicar alguna técnica de clasificación para el aprendizaje supervisado como un procedimiento de validar los clústers obtenidos. Con la finalidad de validar el agrupamiento obtenido con el algoritmo PAM se aplicó el algoritmo de árbol de clasificación C5.0 con poda, en el cual se tuvo una data de entrenamiento, conformada por el 70% de la data original y una data de prueba, conformado por el 30% de la data original del casino. En el Cuadro 13 se presenta la respectiva matriz de confusión realizado con la data de prueba. En este cuadro se puede apreciar que el clasificador obtiene un 99.35% de correcta clasificación, lo cual se puede indicar que los clúster obtenidos son consistentes con los datos analizados.

**Cuadro 13. Matriz de confusión considerando la clasificación como clúster**

Observado	Pronosticado			
	Clúster 1	Clúster 2	Clúster 3	Total
Clúster 1	376	0	1	99.7%
Clúster 2	1	99	1	98.0%
Clúster 3	2	0	299	99.3%
Total	48.7%	12.7%	38.6%	99.35%

#### 4.5 Caracterización del agrupamiento obtenido con el algoritmo PAM

El propósito final de todo análisis de agrupamiento, es caracterizar o tipificar los grupos formados identificando las principales diferencias que se pueden encontrar entre los clientes que pertenecen a diferentes grupos (entre grupos) y las similitudes de los clientes que pertenecen al mismo grupo (dentro de grupos). Así mismo, esta caracterización permitirá a la empresa (casino) identificar patrones y perfiles sobre el comportamiento que realizan sus clientes en los diversos juegos y aplicar estrategias de marketing más efectivas.

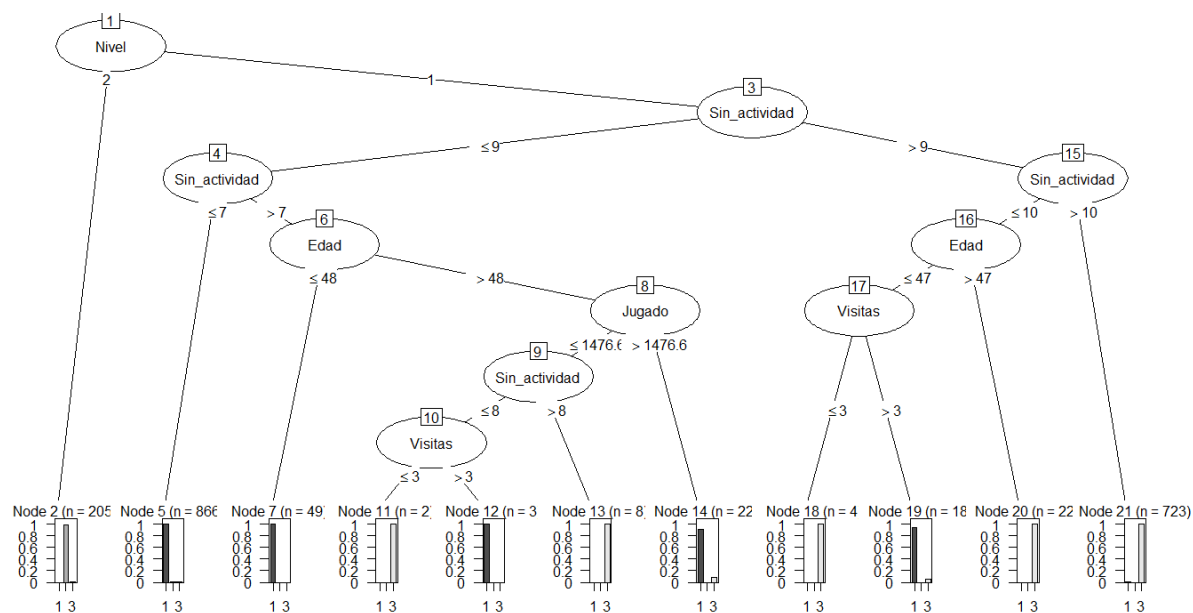
Con esta finalidad se elaboraron algunos cuadros con medidas estadísticas, que permitan la caracterización de los grupos obtenidos.

En el Cuadro 14 se muestran las reglas de decisión que se hallaron por medio del árbol de clasificación C5.0 con poda.

**Cuadro 14. Reglas para los clústers según árbol de clasificación C5.0 con poda**

Regla	Clúster 1	Clúster 2	Clúster 3
<b>1</b>	Nivel=1 & Sin actividad<=7 90.40% del clúster 1	Nivel = 2 100% del clúster 2	Nivel=1 & Sin actividad<=9 & Sin actividad>8 & Edad>48 & Jugado<=1476.6 1.05% del clúster 3
<b>2</b>	Sin actividad<=9 & Sin actividad>7 & Edad<=48 5.11% del clúster 1		Nivel=1 & Sin actividad>9 & Sin actividad <=10 & Edad>47 2.89% del clúster 3
<b>3</b>	Nivel=1 & Sin actividad>7 & Edad>48 & Jugado>1476.6 2.29% del clúster 1		Nivel=1 & Sin actividad>9 & Sin actividad >10 95.26% del clúster 3

Nivel=1: Classic      Nivel=2: Silver



**Figura 3. Árbol de clasificación C5.0.**

En la Figura 3 se puede observar como quedaron formados los nodos y cuantos clientes le corresponden a los últimos nodos para cada clúster.

Además, se puede observar en el Cuadro 15 la comparación de las proporciones de la data entrenamiento utilizada para obtener el modelo mediante el árbol de clasificación C5.0, la cual es el 70% de la data completa con respecto a la proporción de la data completa utilizada con el algoritmo PAM, las cuales resultan similares.

**Cuadro 15. Distribución de número y porcentaje de clientes con el algoritmo PAM y con el Árbol de clasificación C5.0**

Clúster	Data Completa		Data de Entrenamiento	
	Número	Porcentaje	Número	Porcentaje
Clúster 1	1333	49.4	958	49.8
Clúster 2	304	11.3	205	10.7
Clúster 3	1064	39.4	759	39.5
Total	2701	100.0	1922	100.0

En el Cuadro 16 se muestra los promedios para las variables cuantitativas y para cada uno de los tres clústeres. Se aprecia que respecto a la variable monto de recarga, el clúster 2 es el de mayor monto promedio con S/. 5195.7, seguido el clúster 1 con S/. 946.7 y el clúster 3 con S/. 565.0. En términos generales el clúster 2 está conformado por los clientes que tienen los mayores valores promedios para las variables monto jugado, monto de recarga, número de visitas, edad y tiempo en sala.

**Cuadro 16. Promedios de las variables por clústeres**

Clústers	Jugado	Recargas	Visitas	Sin Actividad	Edad	Tiempo sala
Clúster 1	4502.6	946.7	9	3	45	100.04
Clúster 2	29953.4	5195.7	20	4	48	191.87
Clúster 3	2656.4	565.0	4	19	45	91.66

En el Cuadro 17 se observa que la distribución de los clientes en los clústers por sexo. Los clústers 1 y 3 son los que tienen los mayores porcentajes de clientes que asisten al casino de sexo masculino con valores de 67% y 64% respectivamente.

**Cuadro 17. Distribución por Sexo según los clústers**

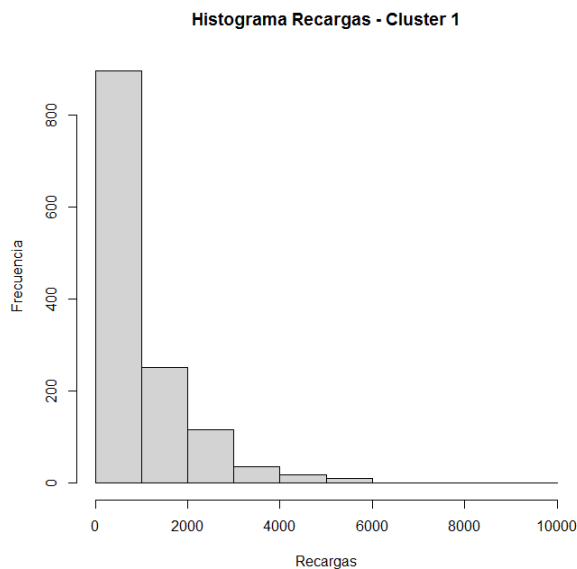
Clúster	Masculino	Femenino	Total Clúster
	Porcentaje	Porcentaje	
Clúster 1	67%	33%	100%
Clúster 2	59%	41%	100%
Clúster 3	64%	36%	100%

En el Cuadro 18 se presenta la distribución según los tres tipos de clientes del casino para cada uno de los clústeres. Se observa que cada clúster tiene el 100% de clientes de un solo tipo. En el clúster 1 y 3 el 100% de los clientes que asisten al casino son clientes con tarjeta Classic. En el clúster 2 el 100% de los clientes que asisten al casino son clientes con tarjeta Silver.

**Cuadro 18. Distribución por tipo de cliente según los clústeres**

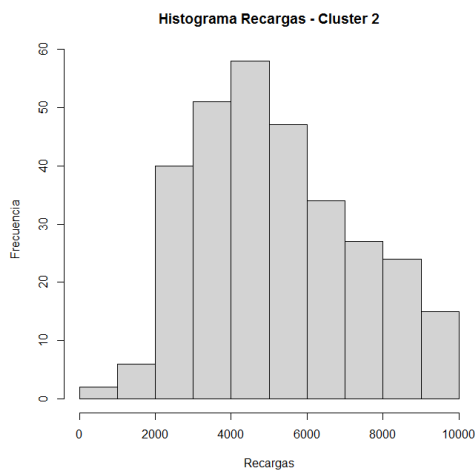
Clúster	Classic	Silver	Gold	Total Clúster
	Porcentaje	Porcentaje	Porcentaje	
Clúster 1	100%	0%	0%	100%
Clúster 2	0%	100%	0%	100%
Clúster 3	100%	0%	0%	100%

Con la finalidad de observar la distribución de los clientes en cada uno de los clúster formados, se elaboraron gráficos sobre la forma de la distribución. En la Figura 4 se muestra la distribución de los clientes respecto a los montos de recargas para el clúster 1. Se aprecia que los montos de recargas presentan una asimetría positiva, indicando que la mayoría de los clientes pertenecientes a este clúster muestran bajos montos de recargas en sus tarjetas de juego.



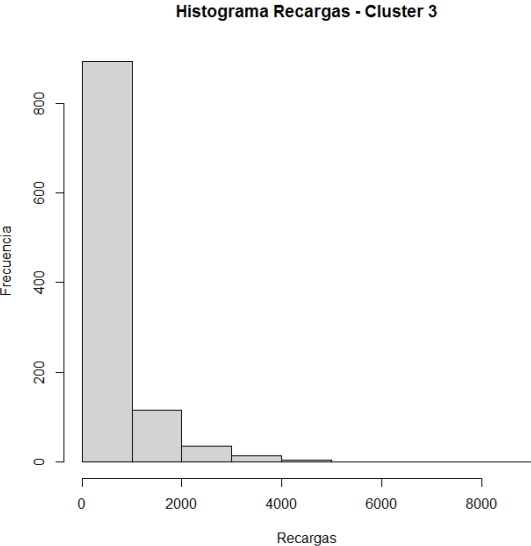
**Figura 4. Distribución de monto de recargas para el clúster 1**

En la Figura 5 se aprecia que los montos de recargas de los clientes que pertenecen al clúster 2 muestran una distribución simétrica, donde se puede indicar que los clientes pertenecientes a este clúster tienen recargas en sus tarjetas de juego cercanos a la media.



**Figura 5. Distribución de monto de recargas para el clúster 2**

La Figura 6, muestra que los montos de recargas de los clientes del clúster 3 tienen una asimetría positiva donde la mayoría de clientes pertenecientes a este clúster tienen bajos montos de recargas en sus tarjetas de juego.



**Figura 6. Distribución de montos de recargas para el clúster 3**

## V. CONCLUSIONES

Las conclusiones de la presente investigación son:

- 1) En el análisis de agrupamiento, en primer lugar aplicando el método de silueta pudo identificar tres como el número de clúster óptimo para el algoritmo PAM. Los resultados de aplicar el algoritmo PAM con la medida de distancia Gower, resultó con una distribución de la segmentación de los clientes del casino para cada uno de los tres clúster porcentajes de 49.4%, 11.3% y 39.4% respectivamente.
- 2) Se aplicaron dos métodos para la validación del agrupamiento. Los resultados de los 6 ANVAs para las variables cuantitativas, resultaron en todas las pruebas F significativas. Con lo cual se valida los tres clúster conseguidos con el algoritmo PAM y su respectiva consistencia. También se aplicó el algoritmo de árbol de clasificación C5.0 con poda, mostrando la tabla de confusión con un 99.35% de correcta clasificación; con lo cual se puede notar que las clasificaciones de los clientes a cada uno de los clúster es muy buena por medio del agrupamiento con el algoritmo PAM.
- 3) Respecto a la caracterización de los clústeres obtenidos con el algoritmo PAM.
  - **El clúster 1.** Según la caracterización que se obtuvo por medio del árbol de clasificación C5.0 con poda, sería la siguiente, que los clientes del clúster 1 tienen la variable sin actividad menor igual a 7 días, pertenecen al nivel de tarjetas classic, presentan edades menor igual a 48 y sus montos jugados acumulados a lo largo del mes es mayor igual a 1476.6 soles.  
Otra forma de observar cómo se comporta el clúster 1 es por medio del Cuadro 16, donde indica que son los clientes con los promedios intermedios para las 6 variables. Los clientes de este clúster tienen promedios de monto jugado de S/. 4502.6, monto de recargas de S/. 946.7 y tiempo en sala de 100.04 horas al mes. El 67.0% son de sexo masculino. El 100% el tipo de tarjeta es classic. Los montos de recargas presentan una distribución con asimetría positiva.
  - **El clúster 2.** Según lo obtenido mediante el árbol de clasificación C5.0 con poda, sería la siguiente, que los clientes del clúster 2 tienen montos jugados acumulados a lo largo del mes mayores a 12961.17 soles y pertenecen al nivel de tarjetas silver.  
Otra forma de entender este clúster es por medio del Cuadro 16 donde indica que son los clientes con los promedios mayores para las 6 variables. Tienen promedios de monto jugado de S/. 29953.4, monto de recargas de S/. 5195.7 y tiempo en sala de



191.87 horas. El 59.0% son de sexo masculino. El 100% el tipo de tarjeta es silver. Los montos de recargas presentan una distribución simétrica.

- **El clúster 3.** Según lo obtenido mediante el árbol de clasificación C5.0 con poda, sería la siguiente, que los clientes del clúster 3 tienen a la mayoría de sus integrantes con las siguientes características de las variables como nivel de tarjetas classic, sin actividad mayor a 10; también se puede observar que pertenecen a este clúster aquellos clientes que tienen actividad menor a 10 días y edad mayor a 47.

También podemos conocer el comportamiento de este clúster mediante el Cuadro 16 el cual indica que son clientes con los promedios menores para las 6 variables. Tienen promedios de monto jugado de S/. 2656.4, monto de recargas de S/. 565.0 y tiempo en sala de 91.66 horas al mes. El 64.0% son de sexo masculino. El 100% el tipo de tarjeta que usan es classic. Los montos de recargas presentan una distribución asimétrica positiva.

## **VI. RECOMENDACIONES**

- 1) Aplicar otros métodos de agrupamiento basados en particiones como el CLARA para observar la performance de los algoritmos.
- 2) Aplicar otros métodos de agrupamiento como los basados en probabilidades, como el algoritmo EM (Maximum Expectation) para observar la performance de los algoritmos.
- 3) Implementar estrategias de marketing según la caracterización encontrada para cada uno de los tres clústeres encontrados con el algoritmo PAM.
- 4) Usar información socio económica de los clientes para mejorar el agrupamiento

## VII. REFERENCIA BIBLIOGRÁFICA

- Arbin, N., Suhailayani, N., & Zafirah, N. (2012). Comparative Analysis between K-Means and K-Medoids for Statistical Clustering.
- Aroral, P., & Varshney, S. (2015). Analysis of K-Means and K-Medoids Algorithm For Big Data.
- Batra, A. (2011). Comparations Between Data Clustering Algorithms. *5ta. IEEE International Conference on Advanced Computing & Communication Technologies*, pp.274-279.
- Kassambara, A. (2017). *Multivariate Analysis I, Practical Guide To Cluster Analysis in R*. París: STHDA.
- Kaufman, L., & Rousseeuw, P. (1990). Finding Groups in Data. An Introduction to Cluster Analysis.
- Prasad, P., & Latesh, G. (2011). Generating Customer Profiles for Retail Stores Using Clustering Techniques. *International Journal on Computer Science and Engineering (IJCSE)*, Vol.3 No. 6.
- Quinlan, J. (1993). C4.5: Programs for Machine Learning. *Morgan Kaufmann*.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, Vol. 1., pp. 86-106.
- Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0), pp. 53-65.
- Sankar, R. (2011). Customer Data Clustering Using Data Mining Technique. *International Journal of Database Management Systems ( IJDBMS )* Vol.3, No.4, pp. 1-11.
- Saunders, J. (1980). Cluster Analysis for Market Segmentation. *European Journal of Marketing*, Vol. 14, Issue 7., pp. 422-435.
- Tiwari, M., & Singh, R. (201). Comparative Investigation of K-Means and K-Medoid Algorithm on Iris Data. *International Journal of Engineering Research and Development*. Vol. 4, Issue 8., pp. 69-72.

## VIII. ANEXOS

### ## Paquetes Requeridos

##

```
install.packages("StatMatch")
```

```
install.packages("cluster")
```

```
install.packages("factoextra")
```

```
install.packages("foreign")
```

```
install.packages("C50")
```

##

### ## Código para distancia de Gower

```
library(StatMatch)
```

```
x<-read.table(file.choose(),T)
```

```
#datos<- read.delim("clipboard")
```

```
x1<-cbind(x[,1:6],apply(x[,7:8],2,factor))
```

##

```
str(x1)
```

```
gower.dist(x3)
```

##

### ## Estandarizando las variables cuantitativas #####

```
x2<-scale(x[,1:6])
```

```
head(x2)
```

### ## Juntando con las variables cualitativas ###

```
x3<-cbind(x2,x1[,7:8])
```

```
head(x3)
```

##

### ## Método Silueta para saber el número óptimo de clústers Figura2.

```
library(cluster)
```

```
library(factoextra)
```

```
fviz_nbclust(x2, pam, method = "silhouette")+
```

```
  theme_classic()
```

##

### ## Método Partición Alrededor de Medoids (PAM)

```
library(cluster)
```

```
pamx <- pam(daisy(x3, metric = "gower"), 3, diss = TRUE)
```

```

pamx
summary(pamx) #separa nuestra data en la cantidad de cluster que hemos pedido
##
## Juntando la data con una columna que indica a que cluster pertenece
dd <- cbind(x3, cluster = pamx$clustering)
head(dd, n = 1100)
dd
##
## Preparando la data con columna de clúster para hallar los Cuadros
##
Jugado<-as.numeric(x[,2])
Recargas<-as.numeric(x[,3])
Visitas<-as.numeric(x[,4])
Sin_actividad<-as.numeric(x[,5])
Edad<-as.numeric(x[,6])
T_de_juego_promedio_x_visita<-as.numeric(x[,7])
Nivel<-as.factor(x[,8])
Sexo<-as.factor(x[,9])
Cluster<-as.factor(pamx$clustering)
##
## Data con la columna de clústers hallados agregada
x100<-
data.frame(Jugado,Recargas,Visitas,Sin_actividad,Edad,T_de_juego_promedio_x_visita,Nive
l,Sexo,Cluster)
##
head(x100)
attach(x100)
##
summary(x100)
##
## Cuadro 8. Medoides finales con el algoritmo PAM
##
pamx$medoids
##

```

```
## Cuadro 9. Distribución de número y porcentaje de clientes con el PAM.
```

```
##
```

```
library(foreign)
```

```
##
```

```
tabla<-table(x100[,9])
```

```
tabla
```

```
porcentajes<-prop.table(tabla)*100
```

```
porcentajes
```

```
##
```

```
Cuadro9<-cbind(tabla,round(porcentajes,1))
```

```
Cuadro9
```

```
##
```

```
## Cuadro 10. Análisis de Varianza para las variables agrupadas por clúster.
```

```
##
```

```
ANVA1<-aov(Jugado~Cluster)
```

```
summary(ANVA1)
```

```
##
```

```
ANVA2<-aov(Recargas~Cluster)
```

```
summary(ANVA2)
```

```
##
```

```
ANVA3<-aov(Visitas~Cluster)
```

```
summary(ANVA3)
```

```
##
```

```
ANVA4<-aov(Sin_actividad~Cluster)
```

```
summary(ANVA4)
```

```
##
```

```
ANVA5<-aov(Edad~Cluster)
```

```
summary(ANVA5)
```

```
##
```

```
ANVA6<-aov(T_de_juego_promedio_x_visita~Cluster)
```

```
summary(ANVA6)
```

```
##
```

```
##
```

```

## Cuadro 11. Tabla de contingencia para estudiar la asociación entre sexo y la
## variable clúster
Cluster1<-c(897,436)
Cluster2<-c(179,125)
Cluster3<-c(683,381)
cuadro11<-data.frame(Cluster1,Cluster2,Cluster3)
rownames(cuadro11)<-c("Hombres","Mujeres")
cuadro11
chisq.test(cuadro11)
##
## Cuadro 11. Tabla de contingencia para estudiar la asociación entre nivel y la
## variable clúster
cluster1<-c(1332,1)
cluster2<-c(1,303)
cluster3<-c(1062,2)
cuadro12<-data.frame(cluster1,cluster2,cluster3)
rownames(cuadro12)<-c("Classic","Silver")
cuadro12
chisq.test(cuadro12)
##
## Cuadro 13. Matriz de confusión considerando la clasificación como clúster.
##
#Árboles de clasificación C5.0
library(C50)

## Data prueba y data entrenamiento
##
set.seed(1234)
ind <- sample(2, nrow(x100), replace = TRUE, prob = c(0.7, 0.3))
data1 <- x100[ind == 1, ]
data2 <- x100[ind == 2, ]
##
dataentrenamiento<-data.frame(data1)
attach(dataentrenamiento)
##

```

```

datapueba<-data.frame(data2)
attach(datapueba)
##
head(x100,n=10)
head(dataentrenamiento,n=10)
##
modelo_c5.3<-C5.0( Cluster~ .,data = dataentrenamiento, control =
C5.0Control(noGlobalPruning = TRUE,CF=0.20))
print(modelo_c5.3)
##
summary(modelo_c5.3)
##
prediccion_c5.3 <- predict(modelo_c5.3,newdata=datapueba)
##
#### Matriz de confusión

# Matriz de confusión
tabla1 <- table(prediccion_c5.3, datapueba$Cluster)
tabla1
# % correctamente clasificados
100 * sum(diag(tabla1)) / sum(tabla1)
##
## Cuadro 14. Reglas para los clústers según árbol de clasificación C5.0 con poda
##
## Reglas para el árbol de clasificación C5.0 con poda

modelo_c5.4<-C5.0(Cluster ~ .,data = datapueba,rules=TRUE, control =
C5.0Control(noGlobalPruning = TRUE,CF=0.90))
##
summary(modelo_c5.4)
##
## Figura 3. Arbol de clasificacion C5.0.
##
plot(modelo_c5.3) #gráfico completo
##

```



```

## Cuadro 16. Promedios de las variables por clústers.
##
aggregate(x100[,1:6],list(Cluster),mean)
##
## Cuadro 17. Distribución por Sexo según los clústers
##
prop.table(table(Cluster,Sexo),1)
##
## Cuadro 18. Distribución por tipo de cliente según los clústeres
##
prop.table(table(Cluster,Nivel),1)
##
## Figura 4. Distribución de monto de recargas para el clúster 1.
##
cluster1<-subset(x100, Cluster == 1)
##
hist(cluster1[,2],main="Histograma Recargas - Cluster 1", xlab="Recargas",
ylab="Frecuencia",col="lightgray" )
##
##
## Figura 5. Distribución de monto de recargas para el clúster 1.
##
cluster2<-subset(x100, Cluster == 2)
##
hist(cluster2[,2],main="Histograma Recargas - Cluster 2", xlab="Recargas",
ylab="Frecuencia",col="lightgray" )
##
##
## Figura 6. Distribución de monto de recargas para el clúster 1.
##
cluster3<-subset(x100, Cluster == 3)
##
hist(cluster3[,2],main="Histograma Recargas - Cluster 3", xlab="Recargas",
ylab="Frecuencia",col="lightgray" )

```