UNIVERSIDAD NACIONAL DE CHIMBORAZO

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS Y COMPUTACIÓN

"Proyecto de Investigación previo a la obtención del título de Ingeniero en Sistemas y Computación"

TRABAJO DE TITULACIÓN

"ANÁLISIS DE EXACTITUD DE LOS ALGORITMOS DE CLUSTERING APLICADOS EN LA BASE DE DATOS DEL SISTEMA ACADÉMICO DE LA UNACH"

Autor:

Danny Daniel Cáceres Lobato

Tutor:

MsC. Lady Espinoza.

Riobamba - Ecuador

2019

PÁGINA DE ACEPTACIÓN

Los miembros del tribunal de graduación del proyecto de investigación de título: **"ANÁLISIS DE EXACTITUD DE LOS ALGORITMOS DE CLUSTERING APLICADOS EN LA BASE DE DATOS DEL SISTEMA ACADÉMICO DE LA UNACH"**, presentado por el Sr. Danny Daniel Cáceres Lobato y dirigida por: MsC. Lady Espinoza Tinoco.

Una vez escuchada la defensa oral y revisado el informe final del proyecto de investigación con fines de graduación escrito en el cual se ha constatado el cumplimiento de las observaciones realizadas, remite la presente para uso y custodia en la biblioteca de la Unach.

Para constancia de lo expuesto firman:

MsC. Lady Espinoza

Directora del proyecto **Firma**

PhD. Lida Barba

Miembro del Tribunal **Firma**

MsC. Jorge Delgado

Miembro del Tribunal **Firma**

DERECHOS DE AUTORÍA

La responsabilidad del contenido de este proyecto de graduación corresponde exclusivamente al Sr. Danny Cáceres bajo la dirección de la MsC. Lady Espinoza y al patrimonio intelectual de la Universidad Nacional de Chimborazo.

Autor:

Danny Daniel Cáceres Lobato

060457205-7

DEDICATORIA

Dedico este trabajo de investigación a Dios nuestro creador, quien con su misericordia y bondad nos brinda la vida y todo lo que tenemos a nuestro alrededor, lo dedico también a mis seres queridos que me han apoyado incondicionalmente en este duro transcurso para la obtención de un título profesional.

Danny Daniel Cáceres Lobato

AGRADECIMIENTO

En el presente trabajo de investigación quiero agradecer a Dios, por su amor y bondad que no tienen fin, que me permite sonreír ante todos mis logros que son resultado de su ayuda.

Agradezco a la Universidad Nacional de Chimborazo y a todo el elemento humano que forma parte de esta, por haber permitido formarme en ella principalmente a mis tutores de esta investigación MsC. Cristian Morales y MsC. Lady Espinoza.

Mi agradecimiento a mis padres Mercy Lobato, Danilo Cáceres por ser los principales promotores de mis sueños y confiar en mis expectativas, principalmente a mi mamá Mercy y abuelita Piedad Salcán por ser mujeres dedicadas y luchadoras por su familia, admirables personas que entregan todo por las personas que aman.

A mi esposa, hijo y hermanos por apoyarme en cada decisión y proyecto, gracias a la vida porque cada día me demuestra lo hermosa y justa que puede llegar a ser, a mis suegros Teresa Chisaguano y Carlos Paredes quienes estuvieron presentes incondicionalmente en todo este largo trayecto de la culminación de mis estudios.

Danny Daniel Cáceres Lobato

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

RESUMEN

La Universidad Nacional de Chimborazo (UNACH) dispone de varios sistemas informáticos uno de ellos es el Sistema de Control Académico (SICOA) mismo que contiene grandes cantidades de información académica, personal de estudiantes, docentes y producción científica, la cual se ha utilizado de forma limitada para la toma de decisiones en los procesos institucionales.

En el presente trabajo se hace uso de la información que contiene el sistema de control académico de la unach, correspondientes a información académica, personal de estudiantes, docentes y producción científica. El trabajo consistió en el análisis de exactitud de los algoritmos de clustering K-Means y K-Medoids de aprendizaje no supervisado aplicados a la información indicada, el proceso de minería se llevó a cabo por medio de la metodología CRISP-DM misma que permitió analiza, limpiar y construir los datos, adicionalmente, el proceso de exactitud fue evaluado por nueve modelos tres por cada tabla: estudiante, docente e investigación en cada uno de los modelos se utilizó la métrica del índice de validación de Davies Bouldin.

Como resultado se identificó que el algoritmo con mayor exactitud es K-Means, a través de este se generaron clústers de datos con característica similares y disimiles entre sí, estudiantes, docentes y producción científica. Los resultados obtenidos contribuirán al proyecto "Diseño de estrategias de mejoramiento continuo en la gestión académica e investigativa de la UNACH, utilizando minería de datos".

Palabras Clave: Clustering, K-Means, K- Medoids, Exactitud, Davies Bouldin.

ABSTRACT

The Universidad Nacional de Chimborazo (UNACH) has several computer systems, one of them is the Academic Control System, known in Spanish as "SICOA", or "ACS" in English. It contains large amounts of academic information, student staff, professors and scientific production, which has been used in a limited way for making decisions in institutional processes.

In this paper, the information contained in the UNACH Academic Control System, corresponding to academic information, student staff, professors and scientific production is used. The work was about the accuracy analysis of the K-Means and K-Medoids clustering algorithms of unsupervised learning applied to the indicated information, the mining process was carried out by means of the CRISP-DM methodology, this methodology allowed analyzing, cleaning and constructing the data. In addition, the accuracy process was evaluated by nine three models for each table: student, professor and research, in each of the models the Davies Bouldin validation index metric was used.

As a result, it was identified that the algorithm with greater accuracy is the K-Means, through which data clusters with similar and dissimilar characteristics, students, professors and scientific production were generated. The results obtained will contribute to the project "Designing strategies for continuous improvement in the academic and research management of UNACH, using data mining".

Keywords: Clustering, K-Means, K-Medoids, Accuracy, Davies Bouldin.

Reviewed by: Armas Geovanny, Mgs.
Linguistic Competences Professor

INTRODUCCIÓN

Las tecnologías de la información han generado el crecimiento exponencial de los datos a una velocidad impresionante (Mostafa, 2016), por lo general, los datos provienen de fuentes estructuradas como son las bases de datos relacionales y fuentes no estructurados provenientes de las redes sociales, aplicaciones móviles, mensaje de texto, audio, imágenes, video, entre otros. De esta manera, surge la necesidad de darle valor y utilidad a los datos de forma automática y eficiente, para ello, se utiliza Data Mining o minería de datos, lo cual es un conjunto de técnicas que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto (Sinnexus, 2019).

Las técnicas minería de datos extraen conocimiento y aportan información útil para la toma de decisiones en las organizaciones, las técnicas más comunes son: redes neuronales, árboles de decisión, clustering y reglas de asociación (Rygielski, Wang & Yen, 2002). El Clustering (agrupamiento) permite encontrar grupos de datos que están juntos entre ellos y separados del resto de datos, esta técnica posee una serie de algoritmos de aprendizaje no supervisados (Amat, 2017).

Los análisis de clustering se han dado en diferentes áreas, en la estimación de coste en el desarrollo de software que se realizó una comparación de los algoritmos de clustering Expectation Maximization (EM), K-Means y COBWEB y como resultado el algoritmo EM presento un mejor comportamiento debido a su naturaleza probabilística (Garre, Cuadrado & Sicilia, 2014), del mismo modo otro estudio realiza la segmentación de los contribuyentes que declaran IVA aplicando los algoritmos de Clustering Self-Organizing Maps (SOFM) y K-Means obteniendo como resultado el algoritmo SOFM como el que presentó mejores resultados obteniendo clústers claramente diferenciados entre sí.

En otro estudio se realiza el análisis de la base de datos de Moodle de una clase para principiantes en educación a distancia de una Universidad Federal utilizando los datos demográficos de los estudiantes y aplicando los algoritmos de clustering jerárquicos y particionales, obteniendo como resultado que las agrupaciones realizadas con los dos tipos de algoritmos presentaron características similares (Rodrigues, Gomes, Cavalcanti, Dantas & Sedraz , 2016).

Los algoritmos de clustering K-Means y K-Medoids fueron comparados con la finalidad de identificar el que posee las mejores características en cuanto a exactitud aplicando el índice de validación de Davies Bouldin. Para realizar el proceso de minería de datos, se utilizó la metodología CRISP-DM, la cual se constituye en una de las guías de referencia más utilizada en el desarrollo de proyectos de minería de datos. (Gordillo, Moine & Haedo, 2011)

El objetivo de esta investigación es analizar la exactitud de los algoritmos de clustering aplicados a la base de datos del sistema académico de la UNACH para apoyar al proyecto "Diseño de estrategias de mejora continua en la gestión académica e investigación de la UNACH, utilizando minería de datos".

El documento de investigación está organizado en los siguientes apartados: en el apartado I se presenta el planteamiento del problema y la definición de los objetivos de la investigación, en el apartado II se aborda la fundamentación teórica de la investigación, en el apartado III se define el proceso metodológico empleado en el desarrollo del proyecto y se realiza el análisis de resultados y las respectivas conclusiones en el apartado V.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1 Problema y Justificación

El crecimiento exponencial de los datos y la necesidad de convertirlos en información útil ha motivado a los centros de investigación y universidades del Ecuador, a utilizar datos históricos almacenados, para el análisis y la aplicación de técnicas de minería de datos que aportan con información y conocimiento para la toma de decisiones (Camana, 2016). Los campos de aplicación más destacados son: meteorología, elecciones presidenciales y educación (Camana, 2016).

En el ámbito educativo se tiene como ejemplo el estudio desarrollado por la Universidad Nacional de Loja, el mismo que sirvió para determinar las interacciones de los estudiantes del curso virtual del idioma inglés, modalidad de Estudios a Distancia (MED), con el que se logró aportar con conocimiento que apoyó la toma de decisiones en esta institución de educación superior (Camana, 2016).

La Universidad Nacional de Chimborazo dispone del sistema de control académico (SICOA), este posee una gran cantidad de datos de estudiantes, docentes y producción científica, sin embargo, existe un trabajo muy limitado en el uso de herramientas y técnicas de minería de datos que apoyen a los procesos académicos e investigativos y toma de decisiones.

En este contexto, se busca realizar un análisis de exactitud de los algoritmos de clustering, aplicado a los datos de información académica, personal de estudiantes, docentes y producción científica para generar agrupaciones homogéneas de datos que contribuyan con conocimiento para la toma de decisiones.

1.2. OBJETIVOS

1.2.1. Objetivo General

Identificar el mejor algoritmo de clustering aplicado en la base de datos del sistema académico de la UNACH por medio del análisis comparativo de la exactitud, para apoyar al proyecto "Diseño de estrategias de mejoramiento continuo en la gestión académica e investigación, utilizando minería de datos"

1.2.2. Objetivos específicos

- Utilizar la metodología CRISP-DM para el análisis, preparación y construcción de los datos académicos y personales de estudiantes, docentes y producción científica.
- Aplicar los algoritmos de K-Means y K-Medoids para establecer los clústers de datos de estudiantes, docentes y producción científica.
- Evaluar la exactitud de los algoritmos K-Means y K-Medoids mediante la validación del índice de Davies Bouldin para la definición del mejor algoritmo de clustering que apoye a los procesos de toma de decisiones.

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA

2.1. Minería de datos

La minería de datos es el proceso de descubrir automáticamente información valiosa en grandes volúmenes de datos, con el fin de extraer patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma decisiones (Pang-Ning , Steinbach & Kumar, 2006).

El objetivo principal de un proceso de minería de datos se basa en extraer la información de grandes cantidades de datos y transformarlos para su posterior uso (Hernández , Tomás, Felipe & Nuñez, 2013).

De acuerdo con (Logreira, 2011) la minería de datos tiene sus orígenes en 3 áreas:

- **Estadística Clásica:** Engloba conceptos de análisis de regresión, varianza, desviación estándar.
- **Inteligencia Artificial:** Se compone con heurísticas, aplica el pensamiento humano como el procesamiento a problemas estadísticos.
- **Aprendizaje Automático (machine learning):** Es la unión de estadística y la inteligencia artificial.

2.2. Aplicaciones de la minería de datos

La aplicación de la minería de datos abarca una gran cantidad de escenarios, entre ellos el de la educación (Rosado & Verjel , 2016). Según Peña (2014) la aplicación de la minería de datos en la educación tiene un lugar importante dentro de las investigaciones sobre la información que se almacena dentro de ámbito educativo.

Algunas de las tareas sustanciales de la minería de datos son la identificación de aplicaciones para las técnicas existentes, y desarrollar nuevas técnicas para dominios tradicionales o de

nueva aplicación, como el comercio electrónico y la bioinformática (Riquelme, Ruiz & Gilbert, 2006).

Existen varias áreas donde se aplica hoy en día la minería de datos como son geología, medicina, seguridad, detección de fraudes, astronomía, comercio y educación entre otros (Riquelme et al., 2006).

2.3. Metodología CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining), fue concebida alrededor del año 1999 (KDnuggets, 2014) y sugerida por SPSS, la cual garantiza una adecuada planeación y una mayor efectividad en los resultados de un proyecto de minería de datos (Chapman, 2007). Según (KDnuggets, 2014) en una encuesta realizada en el año 2014 afirma que CRISP-DM es la metodología más utilizada para proyectos de minería de datos, con un porcentaje del 43% motivo por la cual en este trabajo ha sido utilizada (ver tabla 1).

Tabla 1: Metodologías usadas en Minería de Datos

Metodología	Porcentaje de aplicación
CRISP-DM	43%
Propia	28%
SEMMA	9%
Otra, sin dominio específico	8%
KDD	6%
De la organización	4%
Otra, de dominio específico	2%
Ninguna	0%

Fuente: KDnuggets, 2014.

La metodología CRISP-DM, es de tipo jerárquica, organizada en cuatro niveles que van desde lo general a lo específico, siendo así que en el nivel más alto existen seis fases para el proceso de minería de datos y dichas fases se enuncian a continuación (Arancibia, 2009).

- Comprensión del negocio o problema
- Comprensión de los datos.
- Preparación de los datos.
- Modelado.
- Evaluación.
- Implementación.

2.4. Técnicas de la minería de datos

Las técnicas de Minería de Datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos (Perez & Santin, 2007, p.2).

En las técnicas de minería de datos se distinguen dos categorías, las técnicas predictivas (aprendizaje supervisado) clasificación, regresión y predicción, mientras que en las técnicas descriptivas (aprendizaje no supervisado) clustering y la asociación (Perez & Santin, 2007).

2.4.1. Técnicas predictivas o supervisadas

Estas técnicas informan sobre la relación entre los datos y sus características (Justicia, 2017).

- **Clasificación**: Es el proceso de dividir un conjunto de datos en grupos mutuamente excluyentes (Weiss & Indurkhya, 1998) **,** con el fin de que cada miembro de un grupo se encuentre lo más cerca posible de otros y los grupos diferentes estén lo más lejos posible de otros (Molina & García, 2006).
- **Regresión:** Tiene como objetivo pronosticar los valores de una variable continua a partir de la evolución sobre otra variable continua, comúnmente el tiempo. Ejemplo, se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costos, etc. a partir de los resultados de semanas, meses o años anteriores (Hernández, 2006).

- **Predicción:** Es un proceso que trata de descubrir el comportamiento de varios atributos en el futuro (Suárez, 2014), además que la predicción emplea valores conocidos de datos para predecir valores futuros basados en tendencias históricas y estadísticas (Maimon & Rokach, 2010).

2.4.2. Técnicas descriptivas o no supervisadas

Estas técnicas responden a preguntas formuladas acerca de los datos (Justicia, 2017).

- **Reglas de Asociación:** Se basan en el descubrimiento de reglas de asociación demostrando condiciones en los valores de los atributos que ocurren simultáneamente de forma frecuente en un determinado conjunto de datos (Hernández, 2006). Además, de descubrir hechos comunes dentro de un conjunto de datos, estas reglas son consideradas como una técnica poderosa de análisis de datos que aparecen comúnmente en minería de datos (Justicia, 2017).
- **Clustering o Agrupamiento:** Es una técnica fundamental de minería de datos la cual consiste en separar la información en grupos distintos, internamente los miembros de cada grupo poseen características similares unos de otros y características disimiles respecto a los miembros de los otros grupos. Los grupos o clústers pueden ser usados para clasificar nuevos datos (Hurtado, 2005)

2.5. Algoritmos de clustering

Los algoritmos de clustering o conglomerados buscan la generación de nuevos conjuntos a partir de datos analizados (Suárez, 2014). A estos algoritmos también se le conoce como clasificación no supervisada (unsupervised learning) (Garre et al., 2014), donde no se tienen clases de grupos predefinidas, sino que los grupos se van creando de acuerdo con las características de los datos (Hernández, 2006).

Los clústers o grupos, son un conjunto de objetos que son similares entre sí, pero diferentes a los objetos en otros grupos (Han, Kamber & Pei, 2012), por lo tanto, la agrupación en clúster es útil porque puede conducir al descubrimiento de grupos previamente desconocidos dentro de los datos.

2.5.1 Clustering

Las técnicas de agrupamiento o clustering se dividen en jerarquización y particionamiento (Jain, Aalam & Doja, 2010).

- **Algoritmos Jerárquicos:** Estos algoritmos de clustering crea una descomposición jerárquica formando un árbol o dendograma que divide la base de datos recursivamente en conjuntos cada vez más pequeños (Mamani , 2015). Además, este tipo de algoritmos no requieren que el usuario especifique de antemano el número de clústers.
- **Algoritmos Particionales:** Los algoritmos de clustering particional logran obtener una partición simple de los datos en vez de la obtención de la estructura del clúster tal como se produce con los dendograma de la técnica jerárquica (Jain, Murty & Flynn, 1999). La mayoría de los algoritmos de partición son iterativos y están basados en la distancia y requieren que el usuario especifique de antemano el número de clústers que se van a crear (Han et al., 2012).

2.5.1.1 Algoritmo K-Means

Se basa en el análisis de grupos o conglomerados, divide los datos recogidos en bloques, separados y agrupados por características similares (Gutiérrez & Molina, 2016).

El algoritmo K-Means clustering agrupa las observaciones en K grupos distintos, donde el analista determina el número de clústers (K). K-Means halla los K mejores clústers,

entendiendo como mejor clúster aquel cuya varianza interna (intra- clúster variation) sea lo más pequeña posible (Amat, 2017).

El algoritmo k-means consta de los siguientes pasos:

1. Especificar el número K de clústers que se quieren crear.
2. Seleccionar de forma aleatoria k observaciones del set de datos como centroides iniciales.
3. Asignar cada una de las observaciones al centroide más cercano.
4. Para cada uno de los K clústers recalcular su centroide.

Repetir los pasos 3 y 4 hasta que las asignaciones no cambien o se alcance el número máximo de iteraciones establecido.

Para Amat (2017) el algoritmo k-means presenta ciertas ventajas y desventajas que se detallan en la tabla 2:

Tabla 2: *Ventajas y desventajas del algoritmo K-means*

Ventajas	Desventajas
K-means es uno de los métodos de clustering más utilizados.	Requiere que se indique de antemano el número de clústers que se van a crear
Destaca por la sencillez y velocidad de su algoritmo	Las agrupaciones resultantes pueden variar dependiendo de la asignación aleatoria inicial de los centroides
	Presenta problemas de robustez frente a outliers.

Fuente: *Amat, 2017*

2.5.1.2. Algoritmo K-Medoids

El algoritmo de clustering K – medoids funciona como batch de k-means, con la diferencia de que no se utilizan los centros como prototipos de los grupos, más bien se utilizan los medoides. (Mirkin, 2019)

De acuerdo con Amat (2017) un medoid es un elemento dentro de un clúster cuya distancia (diferencia) promedio entre él y todos los demás elementos de este clúster es la menor posible.

El algoritmo k-medoids es más robusto que k-medias porque un medoid es menos influenciado al tener la presencia de ruido, valores atípicos u otros valores externos (Han et al., 2012).

El algoritmo k-medoids o también conocido como PAM (Mirkin, 2019) tiene los siguientes pasos:

1. Seleccionar K observaciones aleatorias como medoids iniciales.
2. Calcular la matriz de distancia entre todas las observaciones.
3. Asignar cada observación a su medoid más cercano.
4. Para cada uno de los clústers, comprobar si seleccionando otra observación como medoid se consigue reducir la distancia promedio del clúster, si esto ocurre, seleccionar la observación que consigue una mayor reducción como nuevo medoid.

Si un medoid ha cambiado en el paso 4, volver al paso 3, si no, se termina el proceso.

Amat (2017) afirma que a diferencia del algoritmo K-means, en el cual se minimiza la suma total de cuadrados intra - clúster (suma de las distancias al cuadrado de cada observación respecto a su centroide), el algoritmo K-medoid (PAM) minimiza la suma de las diferencias de cada observación respecto a su medoid.

Para Amat (2017) el algoritmo k- medoid presenta ciertas ventajas y desventajas que se detallan en la tabla 3:

Tabla 3*: Ventajas y desventajas del algoritmo K-medoids*

Ventajas	**Desventajas**
K-medoids es un algoritmo de clustering más fuerte que K-means y es más adecuado utilizarlo cuando el set de datos contenga outliers o ruido.	Necesita que se especifique el número de clústers que se van a crear. Para sets de datos considerables se necesitan muchos recursos computacionales.

Fuente: *Amat, 2017*

2.6. Exactitud en los algoritmos de clustering

Para León (2014) el clustering tiene como objetivo agrupar objetos similares en el mismo clúster y objetos diferentes ubicarlos en diferente clúster, por lo que existen métricas de validación de clústers internas y están basadas en dos criterios:

- **Cohesión:** El miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster.
- **Separación:** Los clústers deben tener una separación considerable entre ellos.

2.6.1. Davies-Bouldin index (DB)

El índice de Davies-Bouldin apunta a identificar conjuntos de grupos que son compactos y bien separados (Bolshakova & Azuaje, 2003).

El índice de validación de Davies-Bouldin, DB, se define como:

$$DB = \frac{1}{k} \sum_{i=1, i \neq j}^{k} \max \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

Figura 1. Índice Davis Bouldin (Bolshakova & Azuaje, 2003)

En función a la separación entre grupos, valores pequeños para el índice DB indica clústers compactos, y cuyos centros estén bien separados los unos de los otros. Consecuentemente el número de clústers que minimiza el índice DB se toma como el óptimo (Bolshakova & Azuaje, 2003).

CAPÍTULO III

METODOLOGÍA

3.1. Tipo de investigación

En el desarrollo de la presente investigación se tiene una combinación de métodos y metodologías, el uso de estos depende de cada fase de la investigación, es así como para realizar el marco teórico referencial se empleó el enfoque de investigación cuantitativo propuesto por (Sampieri, 2010) con el cual se evaluó la literatura existente en las bases de datos científicas que sirvieron para definir los conceptos de minería de datos, metodologías para el desarrollo de minería de datos, técnicas de minería de datos, entre otros, empleando los métodos de investigación inductivo-deductivo y analítico-sintético (Navarro, 2014) .

La metodología CRISP-DM permitió llevar a cabo el proceso de minería de datos. A continuación, se describe el desarrollo del proceso:

1. **Recolección y análisis de los datos:** Los datos se obtuvieron de archivos en formato Excel provenientes de la base de datos del Sistema de Control Académico (SICOA) de la Universidad Nacional de Chimborazo, la vigencia de los datos es a partir del año 2012 hasta el año 2018, los datos proporcionan información personal, académica de estudiantes y docentes, además los proyectos de investigación científica realizados por los docentes y los resultados de la heteroevaluación aplicada a los mismos. Se utilizó el método analítico-sintético, debido a que en el estudio y análisis de la información proporcionada se simplificarán las variables.
2. **Comprensión de los datos:** Para el desarrollo de esta fase se empleó la plataforma de Talend Data Quality Online y se realizó una investigación descriptiva exploratoria de los datos, la cual permitió evaluar la calidad de los datos mediante tablas y gráficos, en función de los valores válidos, no válidos y nulos existentes.

3. **Preparación de los datos:** Se aplicó un método analítico deductivo, mismo que permitió corregir los errores identificados, limpiar y depurar los datos para posteriormente definir la combinación de las variables en el proceso de clustering.
4. **Implementación y evaluación de los modelos:** Se utilizó la herramienta RapidMiner misma que permitió implementar los modelos de clustering, adicionalmente, se empleó la herramienta R para el análisis del número óptimo de clústers para posteriormente, realizar la aplicación de los algoritmos K-Means y K-Medoids sobre el conjunto de datos elegido y luego realizar el análisis e interpretación de los grupos de datos o conglomerados obtenidos al momento de aplicar los algoritmos y finalmente analizar la exactitud de cada uno de estos mediante la métrica del Índice de Davies-Bouldin.
5. **Inducción científica:** En esta fase se aplicó los métodos inductivo-deductivo y analítico- sintético para identificar, evaluar y determinar los resultados obtenidos en el proceso de clustering que apoye al proyecto "Diseño de estrategias para mejorar la gestión académica e investigación, utilizando minería de datos".

3.2. Unidad de análisis

La unidad de análisis para esta investigación está basada en la base de datos del Sistema de Control Académico (SICOA) de la Universidad Nacional de Chimborazo específicamente en la información personal y académica de estudiantes, docentes y producción científica.

3.3. Técnicas de Análisis e interpretación de la información

El análisis de los datos se lo realizará mediante la técnica de clustering, aplicando los algoritmos K-Means y K-Medoids para obtener agrupaciones de datos homogéneas, además se aplicó la metodología CRISP-DM para el proceso de minería, además, la exactitud de los algoritmos se medirá mediante el índice de Davies Bouldin.

3.4. Aplicación de la metodología CRISP-DM

En este capítulo se detalla cada fase de la metodología CRISP-DM que se aplicó al proceso de minería de datos. A continuación, se detalla cada una de las fases.

3.4.1. Fase de comprensión del negocio o problema.

Determinar el objetivo del negocio.

En este proyecto el objetivo del proceso de minería de datos es analizar la exactitud de los de algoritmos de clustering aplicados en la base de datos del Sistema de Control Académico (SICOA) de la Universidad Nacional de Chimborazo además realizar agrupaciones o clústers de la información proporcionada por la base de datos antes mencionada. Dicho estudio se realizará para apoyar al proyecto "Diseño de estrategias de mejoramiento continuo en la gestión académica e investigativa de la institución, utilizando minería de datos".

Evaluación de la Situación: En esta tarea se realizó un análisis de los recursos hardware y software, además de los riesgos relacionados con la perdida de información útil para el desarrollo del proyecto de minería de datos ver anexo 1.

Determinar los objetivos de la minería de datos: Los objetivos de la minería de datos para esta investigación son:

- Realizar clústers o grupos de la información personal y académica de los estudiantes.
- Agrupar la información de los docentes a partir de datos personales, académicos y evaluación del docente.
- Realizar clústers de la información de los proyectos de investigación realizados por ellos docentes de la UNACH.
- Determinar la exactitud de los algoritmos de clustering.

Realizar el plan del proyecto: En esta tarea se desarrolló el plan de proyecto considerando los pasos a seguir y los métodos a emplear en cada paso como se lo puede observar en él (anexo 2).

3.4.2. Fase Comprensión de los datos

Recolectar datos iniciales: En esta tarea se recopiló la información de la base de datos del sistema de control académico SICOA proporcionada en archivos con formato .xlsx, información que se encuentra almacenada desde el año 2012 y los datos proporcionados son los siguientes.

- Estudiante Información Personal
- Estudiante Rendimiento
- Docente Información Personal
- Docente Información Académica
- Evaluación Docente
- Investigación

Descripción de los datos.: En esta tarea se realizó la descripción de los datos estudiante información personal, estudiante rendimiento, docente información personal, docente información académica, evaluación docente e investigación dentro de estas se puede observar el nombre, significado y tipo de dato. (ver anexo 3).

Exploración de los datos: Como se puede observar en el anexo 4 se realizó una exploración de los datos realizando pruebas estadísticas básicas que revelaron algunas propiedades de los datos utilizando RapidMiner Studio versión 9.3 edición educacional.

Verificar la calidad de los datos: En esta tarea se verificó la calidad de los datos, esto significa encontrar errores de codificación, errores en los datos, valores perdidos, valores nulos, valores válidos, valores inválidos ver anexo 5.

3.4.3. Preparación de los datos.

Selección de los datos: En esta tarea se seleccionó un subconjunto de datos, tomando en cuenta la calidad de los datos realizado anteriormente (ver anexo 5), de esta manera se eligieron los atributos relevantes para el estudio planteado.

Limpieza de los datos: En esta fase se realizó la depuración de todos los datos obtenidos para alcanzar el nivel de calidad requerido para la aplicación de las técnicas de data mining, se aplicaron varias técnicas de limpieza de datos como la normalización de los datos, discretización de campos numéricos, tratamiento con valores vacíos, etc. ver anexo 6

Construir los datos: En esta tarea se realizó la generación de nuevos atributos que se integraron a los registros ya existentes como por ejemplo ponderación promedio, foráneo, tiene hijos, etc. Ver anexo 7.

Integración de los datos: En este apartado se fusionaron las siguientes tablas:

La tabla Estudiantes que contiene datos personales y socioeconómicos de los estudiantes con la tabla Estudiante Rendimiento que contiene la información académica de los mismos.

Las tablas Docente con Docente Información Académica y Evaluación Docente, estas contenían información personal, académica y los resultados de las evaluaciones aplicadas a cada uno de los docentes.

Las tablas Investigación, Docente y Docente Información Académica mismas que contienen información personal, académica y producción científica de los docentes.

Formateo de Datos: En esta tarea se dio formato a algunos atributos de las tablas como género, facultad, estado civil, etnia, entre otros, debido a que los algoritmos aceptan solo campos con valores numéricos para el análisis, ver anexo 8.

Luego de realizar el proceso de comprensión y preparación de los datos se procedió a elegir los campos con cuales se aplicará el proceso de minería de datos estos se muestran a continuación.

Tabla 4: Datos del estudiante

Estudiantes	
Atributos	**Naturaleza**
Estudiante ID	Original
Estado Civil	Original
Etnia	Original
Género	Original
Promedio	Calculado
Tipo parroquia	Original
Número de Hermanos	Original
Número de Hijos	Original
Número integrantes Hogar	Original
Número Dependen Ingresos	Original
Total de Ingreso	Original
Total Ingresos del Estudiante	Original

Nota: En el anexo 7 se explica la procedencia de la variable promedio

Tabla 5: Datos del Docente

Docentes	
Atributo	**Naturaleza**
Cedula	Original
Estado Civil	Original
Etnia	Original
Género	Original
Nivel de Instrucción	Original
Resultado Final Evaluación Docente	Calculado
Horas de actividad Académica	Original
Horas clase	Original
Horas de Eventos Aprobados	Original
Horas de Eventos Asistidos	Original
No Eventos aprobados	Original
No Eventos Asistidos	Original
No Eventos Nacionales	Original
No Eventos Internacionales	Original
Facultad	Original

Nota: En el anexo 7 se explica la procedencia del atributo resultado final evaluación docente promedio

Tabla 6: Datos de Investigación

Investigación	
Atributo	**Naturaleza**
Cedula	Original
Nivel de Instrucción	Original
Pub Capitulo Libro	Original
Publicaciones Libro	Calculado
Horas Eventos Asistidos	Original
No Eventos aprobados	Original
No Eventos Asistidos	Original
No Eventos Nacionales	Original
Horas Eventos Aprobados	Original
No Eventos Internacionales	Original

Nota: En el anexo 7 se explica la procedencia del atributo Publicaciones libro

3.4.4. Modelado

Selección de técnica de modelado: De acuerdo con los objetivos de negocio y de minería de datos que se menciona en el apartado 4.1 (Comprensión del negocio) se eligió la técnica de clustering dentro de esta se encuentra los algoritmos K-Means y Medoids.

Generar el plan de prueba: Para evaluar la calidad y validez de los modelos además de la exactitud se utilizó la métrica del índice de Davies Bouldin, valores pequeños para el índice DB indica clústers compactos, y exactos.

Construir el modelo: En esta tarea se construyeron 9 modelos 3 por cada tabla estudiante, docente e investigación docente para citar un ejemplo en el primer modelo de evaluación se combinaron las siguientes variables: Estado Civil, Género, Etnia, Tipo Parroquia, Promedio. Ver anexo 9.

Evaluar el modelo: Para evaluar los modelos se tomó en cuenta los valores de la métrica índice de Davies Bouldin que se estableció en el apartado 4.4.2 (Generar el plan de prueba) además se determinó el algoritmo con mayor exactitud, estas evaluaciones se pueden observar en el anexo 10.

3.4.5. Evaluación

Evaluar los resultados: En esta tarea se realizó las agrupaciones de la información de las tablas estudiante, docente e investigación docente con el algoritmo que presento la mayor exactitud este fue el algoritmo K-Means ver anexo 11.

Revisar el Proceso

En esta tarea se revisó que el proceso de minería de datos y la aplicación de la metodología se haya realizado con normalidad, los resultados de exactitud y de la minería de datos obtenidos son los deseados.

Determinar los próximos pasos. El próximo paso es realizar la presentación de los resultados obtenidos de la minería de datos, de acuerdo con los objetivos de la minería de datos planteados.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. Resultados

Para determinar los resultados de esta investigación se utilizaron las tablas de estudiante, docentes e investigación mismas que se detallan a continuación:

Tabla estudiante: Posee 16009 registros desde el periodo académico septiembre 2012 - marzo 2013 hasta octubre 2018 - marzo 2019, luego de realizar la comprensión y preparación de los datos se utilizaron 15793 registros que corresponden al 98.65%, los campos utilizados son estudiante Id, estado civil, etnia, género, promedio, tipo parroquia, número de hermanos, número de hijos, número integrantes hogar, número dependen ingresos, total de ingreso y total ingresos del estudiante.

Tabla Docentes: Esta tabla cuenta con 826 registros desde el periodo académico septiembre 2012 - marzo 2013 hasta octubre 2018 - marzo 2019, luego de realizar la comprensión y preparación de los datos se utilizó el 54.24% es decir 448 registros, los campos empleados son cédula, estado civil, etnia, género, nivel de instrucción, resultado final evaluación docente, horas de actividad, académica, horas clase, horas de eventos aprobados, horas de eventos asistidos, número eventos aprobados, número eventos asistidos, número eventos nacionales, número eventos internacionales y facultad.

Tabla Investigación: En esta tabla se encuentra registros desde el periodo académico septiembre 2012 - marzo 2013 hasta octubre 2018 - marzo 2019 con un total de 2051 registros y luego de realizar la comprensión y preparación de los datos se utilizó el 64.41% equivalente a 1321 registros de publicaciones, los campos a utilizar en el análisis son cedula**,** nivel de instrucción**,** publicaciones capitulo libro**,** publicaciones libro**,** horas eventos asistidos**,** número

eventos aprobados**,** número eventos asistidos**,** número eventos nacionales**,** horas eventos aprobados y número eventos internacionales

Exactitud de los algoritmos en la tabla estudiante

En la tabla 7, se detallan los resultados del índice de Davies Bouldin obtenidos de la aplicación de los algoritmos K-Means y K-Medoids en la tabla estudiante.

Tabla 7: Davies Bouldin de la tabla estudiante

Algoritmo	Modelo	Davies Bouldin Index	Promedio Davies Bouldin Index (Exactitud)
K – Means	1	1.027	0.699
	2	0.369	
	3	0.701	
K – Medoids	1	1.045	1.680
	2	2.650	
	3	1.347	

Después de aplicar los algoritmos en la tabla estudiante, se calculó el promedio general índice de Davies Bouldin obteniendo como resultado que el algoritmo con mayor exactitud es K-Means con un valor de 0.699 a diferencia del algoritmo K-Medoids que presento un valor de 1.680.

Los resultados de la aplicación del algoritmo K-Means que presentó la mayor exactitud se muestran a continuación:

Tabla Estudiante Modelo 1

Conjunto de datos: Estudiantes

Campos analizar: Estado Civil, Etnia, Género, Promedio, Tipo Parroquia

Número máximo de iteraciones: 10

Número de clústers óptimo: 3

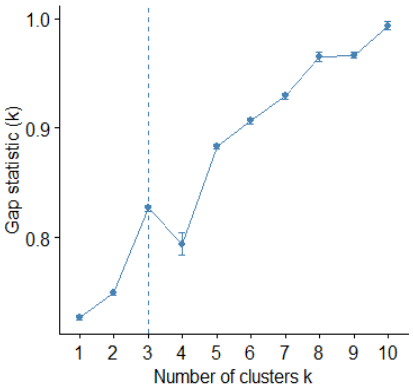

Figura 2. Número óptimo de clústers tabla estudiantes

La tabla 8 muestra los resultados de los clústers que generó este modelo.

Tabla 8: Clústers K-Means tabla estudiante

Clústers	Clúster 0	Clúster 1	Clúster 2
Estado Civil	-0.201	4.541	-0.190
Género	-0.003	0.176	-0.029
Etnia	-0.025	0.462	0.003
Tipo Parroquia	-0.471	-0.072	2.124
Promedio	-0.015	0.232	0.010
Número de instancias	12365	664	2764
Avg. Distancia al punto	3.012	5.323	3.064

Luego de realizar el análisis correspondiente se presenta la gráfica de los centroides de los clústers como se muestra en la figura 3.

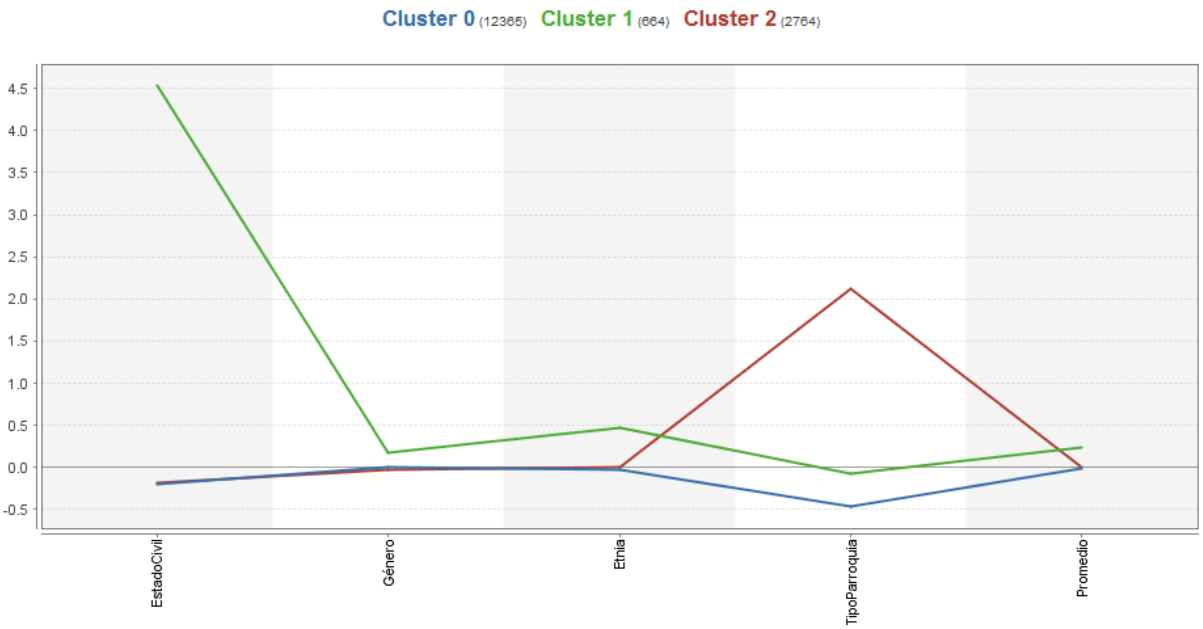

Figura 3. Modelo 3 tabla estudiante centroides de cada clúster

Análisis de la Gráfica

Al evaluar las variables de Estado Civil, Género, Etnia, Tipo Parroquia, Promedio se puede observar que hay una diferencia significativa entre los clústers 0,1 y 2, donde en el clúster 1 se diferencia el estado civil (casado, por ejemplo), mientras que en el clúster 2 se diferencia la etnia(mestizo) y finalmente el promedio obtenido (entre 8-10) del clúster 1 varia a diferencia de los demás.

Cabe mencionar que el promedio de los estudiantes agrupados en el clúster 1 varia en relación a los demás clústers, esto demuestra la diferencia significativa que hay entre este clúster y las demás agrupaciones. El clúster que posee la mayor cohesión es el clúster 0 integrado por 12365 estudiantes mientras que el clúster que posee la mayor separación es el clúster 1 conformado por 664 estudiantes.

Complementariamente se presenta los pesos de cada uno de los atributos que incidieron en este análisis.

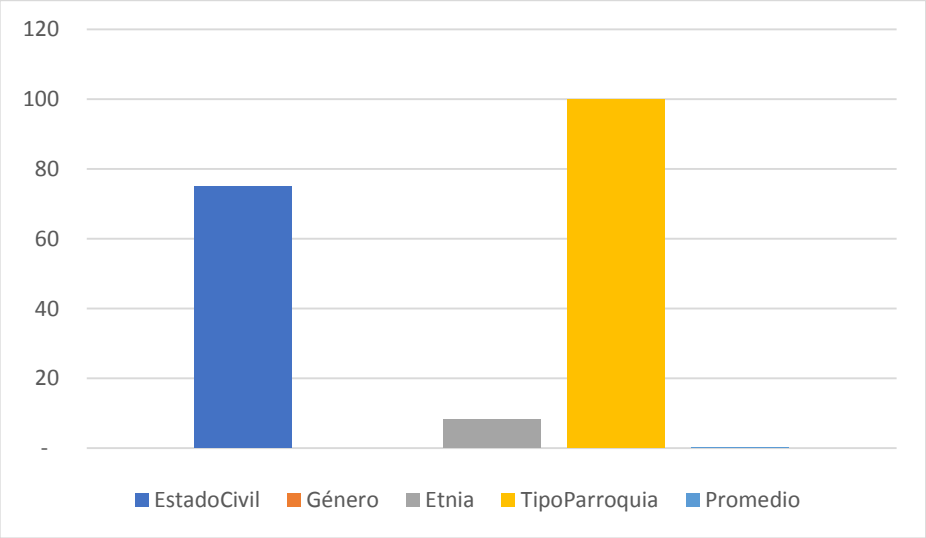

Figura 4: Análisis de los pesos de los atributos, modelo 1

Como se puede observar en la figura las variables que incidieron en este modelo son el tipo de parroquia y el estado civil del estudiante.

Exactitud de los algoritmos en la tabla docente

Como se puede observar en la tabla 9 se detallan los resultados del índice de Davies Bouldin obtenidos por los algoritmos K-Means y K-Medoids en la tabla Docente.

Tabla 9 *: Davies Bouldin de la tabla docente*

Algoritmo	Modelo	Davies Bouldin Index	Promedio Davies Bouldin Index (Exactitud)
K – Means	1	0.879	0.958
	2	0.978	
	3	1.018	
K – Medoids	1	1.586	1.551
	2	1.524	
	3	1.543	

Luego de aplicar los algoritmos en la tabla docente, se calculó el promedio general de exactitud que se obtuvo en cada modelo teniendo como resultado el algoritmo K-Means como el que mayor exactitud presentó con un valor de 0.958 a diferencia del algoritmo K-Medoids con un valor de 1.551.

Los resultados de la aplicación del algoritmo K-Means que presento la mayor exactitud se muestran a continuación:

Tabla Docente Modelo 2

Conjunto de datos: Docente

Campos analizar: No. Eventos Aprobados, No. Eventos Asistidos, Horas Eventos Aprobados, Horas Eventos Asistidos, No. Eventos Nacionales, No. Eventos Internacionales, Horas Actividad Académica, Horas Clase, Resultado Final Evaluación Docente, Nivel Instrucción.

Número máximo de iteraciones: 10

Número de clústers óptimo: 3

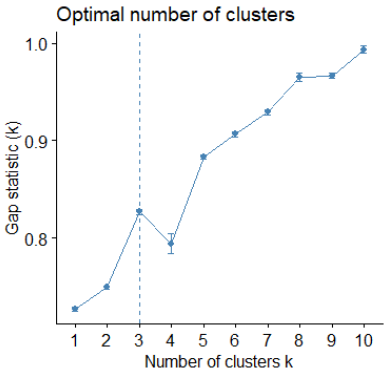

Figura 5. Número óptimo de clústers tabla docentes

En la tabla 10 se muestran los resultados de los clústers que genera este modelo.

Tabla 10*: Clústers K-Means tabla docente*

Clústers	**Clúster 0**	**Clúster 1**	**Clúster 2**
Nivel Instrucción	0.105	-0.307	-0.204
No. Eventos Aprobados	-0.362	0.810	4.100
No. Eventos Asistidos	-0.344	0.933	1.677
Horas Eventos Aprobados	-0.291	0.529	4.922
Horas Eventos Asistidos	-0.319	0.905	1.301
No. Eventos Nacionales	-0.394	0.910	4.043
No. Eventos Internacionales	-0.245	0.640	1.524
Horas Actividad Académica	-0.132	0.356	0.659
Horas Clase	0.030	-0.049	-0.584
Resultado Evaluación Docente	-0.107	0.294	0.465
Número de Instancias	332	108	8
Avg. Distancia al punto 0	5.730	11.01	22.60

Luego de realizar el análisis correspondiente se presenta la gráfica de los centroides de los clústers como se muestra en la figura 6.

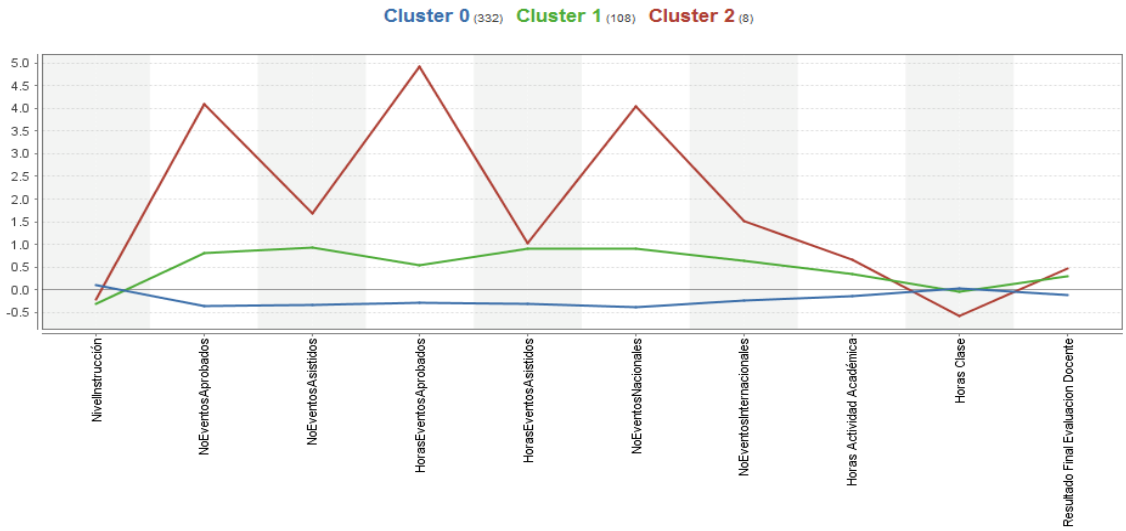

Figura 6. Modelo 3 tabla docente centroides de cada clúster

Análisis de la Gráfica

Del análisis a los resultados obtenidos se diferencian 3 clústers, de los cuales el clúster 2 se destaca por agrupar a docentes que participan en eventos académicos nacionales e internacionales, además de contar con diferencias marcadas en el número de horas clase, esto determina que el grupo del clúster 1 obtenga el mejor resultado en la evaluación docente, sin embargo, el número de docentes que pertenecen a este clúster es el más bajo (8 docentes en total) en relación a los demás clústers. Hay un segundo grupo (clúster 1) que se posiciona en el segundo lugar del resultado obtenido en la evaluación docente y corresponde aquellos docentes que ocasionalmente participan en eventos académicos nacionales e internacionales y representa a uno de los dos grupos más significativos del clustering conformado por 108 docentes. Hay un tercer grupo (clúster 0) que concentra al mayor número de docentes, 332 en total y se caracteriza la escasa participación de los docentes en eventos académicos nacionales e internacionales lo que determina una diferencia significativa en relación a los demás clústers e infiere en el resultado en la evaluación docente obtenida. El clúster que posee la mayor cohesión es el clúster 0 integrado por 332 docentes mientras que el clúster que posee la mayor separación es el clúster 1 conformado por 8 docentes.

Complementariamente se muestran los pesos de las variables que incidieron en este análisis.

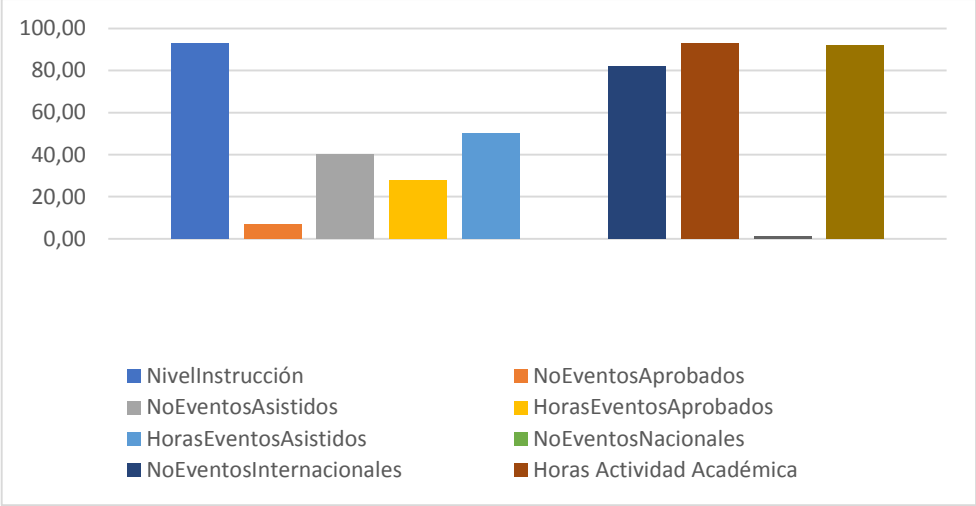

Figura 7. Análisis de los pesos de los atributos, modelo 2

De esta manera se puede evidenciar que las variables que más incidieron en este análisis son el nivel de instrucción, horas de actividad académica, resultado de evaluación docente, numero de eventos internacionales, corroborando así los resultados obtenidos en el clustering.

Exactitud de los algoritmos en la Tabla Investigación

Se observa en la tabla 11 los resultados del índice de Davies Bouldin obtenidos por los algoritmos K-Means y K-Medoids en la tabla investigación.

Tabla 11*: Davies Bouldin de la tabla investigación*

Algoritmo	Modelo	Davies Bouldin Index	Promedio Davies Bouldin Index (Exactitud)
K – Means	1	1.259	1.192
	2	1.215	
	3	1.104	
K – Medoids	1	1.525	1.999
	2	1.726	
	3	2.746	

Luego de aplicar los algoritmos en la tabla investigación, se calculó el promedio general del índice de Davies Bouldin mismos que se obtuvieron de cada modelo teniendo como resultado

el algoritmo K-Means con un valor de 1.192 que es el que mayor exactitud posee a diferencia del algoritmo K-Medoids que presento un valor de 1.999.

Los resultados de la aplicación del algoritmo K-Means que presento la mejor exactitud se muestran a continuación:

Modelo 3 Tabla Investigación

Conjunto de datos: Investigación

Campos analizadas: Nivel Instrucción, No. Eventos Aprobados, No. Eventos Asistidos, No. Eventos Nacionales, No. Eventos Internacionales, Total publicaciones

Número máximo de iteraciones: 10

Número de clústers óptimo:3.

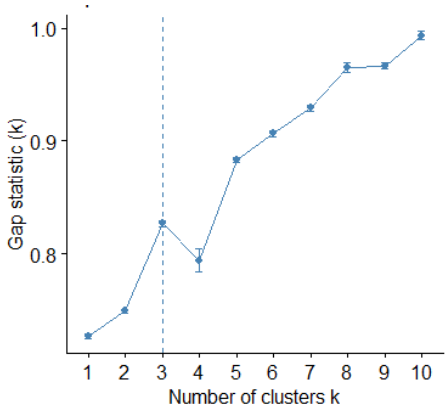

Figura 8. Número óptimo de clústers tabla investigación

En la tabla 12 se muestran los resultados de los clústers que genera este modelo.

Tabla 12: Clústers K-Means tabla investigación

Atributos	Clúster 0	Clúster 1	Clúster 2
Nivel de Instrucción	-0.446	0.143	-0.839
No. Eventos Aprobados	1.317	-0.394	0.657
No. Eventos Asistidos	0.899	-0.268	0.433
No. Eventos Nacionales	1.383	-0.399	-0.211
No. Eventos Internacionales	0.499	-0.207	3.914
Total publicaciones	0.509	-0.241	5.848
Número de Instancias	294	1010	16
Avg. Distancia al punto 0	8.088	2.286	29.912

Resultado de la gráfica de los centroides de cada clúster.

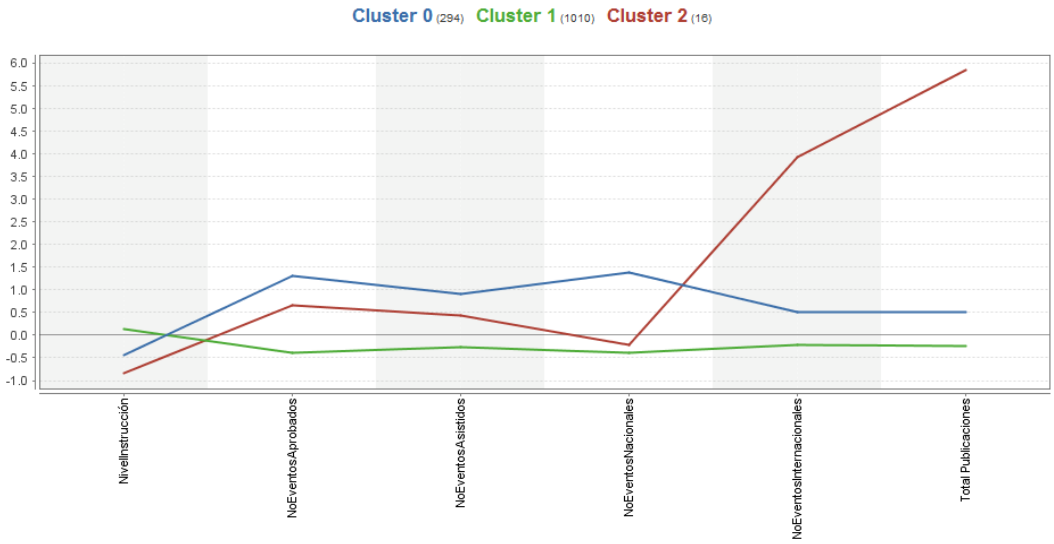

Figura 9. Modelo 3 tabla investigación centroides de cada clúster

Análisis de la Gráfica

Del análisis a los resultados obtenidos se diferencian 3 clústers, de los cuales el clúster 0 concentra a 294 docentes, este clúster se destaca por agrupar a docentes que tienen un determinado nivel académico y participan en eventos académicos nacionales, sin embargo, presentan una escasa participación en eventos internacionales.

El clúster 1 concentra al mayor número de docentes, 1010 en total y se caracteriza la escasa participación de los docentes en eventos académicos nacionales e internacionales lo que determina una diferencia significativa en relación a los demás clústers.

El clúster 2 conformado por 16 docentes, se caracteriza por el nivel académico del docente y una escasa participación del mismo en eventos académicos nacionales, sin embargo, presenta una elevada participación en eventos internacionales. El clúster que posee la mayor cohesión es el clúster 1 integrado por 1010 docentes mientras que el clúster que posee la mayor separación es el clúster 2 conformado por 16 docentes.

En conclusión, el nivel de instrucción y la participación en eventos nacionales e internacionales inciden en el total de publicaciones.

En la siguiente figura, se muestran los pesos de las variables que se utilizaron en este análisis.

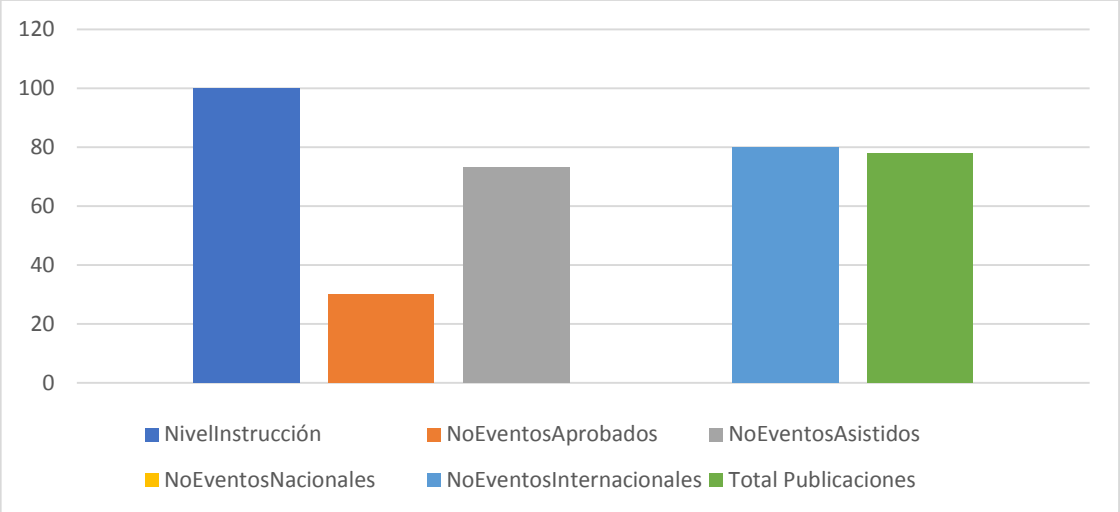

Figura 10. Pesos de las variables modelo 3 tabla investigación

Se observa en la figura 10 las variables que incidieron en este modelo son el nivel de instrucción el número de eventos internacionales el total de publicaciones y el número de eventos asistidos.

Resultado General del algoritmo con mayor exactitud

Para culminar, en la tabla 13 se muestra el promedio del índice de Davies Bouldin de los algoritmos de clustering misma que proviene de cada uno de los análisis de exactitud de las tablas estudiantes, docentes e investigación dando como resultado el algoritmo K-Means con el algoritmo que mayor exactitud presenta con un valor de 0.949.

Tabla 13*: Promedio índice de Davies Bouldin de los algoritmos*

Algoritmo	**Promedio Davies Bouldin Index (Exactitud)**
K – Means	0.949
K – Medoids	1.747

4.2. Discusión

Los resultados de exactitud obtenidos de la tabla estudiante, el algoritmo K - Medoids posee una exactitud (índice de Davies Bouldin) de 1.688 en comparación con la exactitud del algoritmo K - Means que es 0.699 por lo que se determina que el algoritmo que posee la mejor exactitud es el K -Means, mediante este algoritmo se realizó un análisis del número óptimo de clústers calculado por la herramienta R dando un valor de 3, las variables con las cuales se realizó el clustering son Estado Civil, Etnia, Género, Promedio, Tipo Parroquia además se realizó el cálculo de los pesos de cada una de estas dando como resultado que las variables que más influye en el clúster es el tipo parroquia y etnia.

En el análisis en la tabla docente, el algoritmo que posee la mejor exactitud es el K-Means con un valor de 0.958 con respecto al algoritmo K-Medoids con un valor de 1.551, mediante el algoritmo que presentó la mayor exactitud se realizó un análisis del número óptimo de clústers proporcionado por la herramienta R dando un valor de 3, el análisis del clustering se realizó con las variables, Nº. Eventos Aprobados, Nº. Eventos Asistidos, Horas Eventos Aprobados, Horas Eventos Asistidos, Nº. Eventos Nacionales, Nº. Eventos Internacionales, Horas Actividad Académica, Horas Clase, Resultado Final Evaluación Docente, Nivel Instrucción además se realizó el cálculo de los pesos de las variables que incidieron en este análisis teniendo como resultado las variables hora clase, resultado de evaluación docente, horas de actividad académica y numero de eventos internacionales.

De igual forma para el análisis de la tabla investigación, el algoritmo con mejor exactitud es el K-Means con un valor de 1.192 en relación con la exactitud del algoritmo K – Medoids que con un valor de 1.999. Se realizó un análisis del número óptimo de clústers con la herramienta R dando como resultado 3, las variables que intervinieron en este análisis fueron, Nivel Instrucción, Nº Eventos Aprobados, Nº Eventos Asistidos, Nº Eventos Nacionales, Nº Eventos Internacionales, Total publicaciones, de la misma manera se realizó un análisis de las variables

que más incidencia tuvieron teniendo como resultado el nivel de instrucción y el número de eventos internacionales.

Después de haber obtenido el análisis de exactitud de las tres tablas se procedió a sacar un promedio general del mismo obteniendo como resultado que el algoritmo con mayor exactitud es el K-Means con un promedio total de 0.983, cabe mencionar que (Garre, Cuadrado, & Sicilia, 2014) mencionan que el mejor algoritmo en su estudio es el EM (algoritmo probabilístico) y en segundo es el algoritmo K-Means.

5. CONCLUSIONES

Para el proceso de minería de datos se aplicó la metodología CRISP-DM, misma que consta de varias fases: compresión del negocio, comprensión de los datos, preparación de los datos, modelado y evaluación, siendo las fases de comprensión de los datos y preparación de los datos las más críticas porque en estas fases se realizó el análisis de los datos para verificar la calidad de los mismo determinando la cantidad de datos nulos existentes en cada una de las tablas estudiantes, docentes e investigación para luego proceder a realizar las tareas de construir, integrar y depurar los datos que se van a utilizar en este proceso, y es por esta razón que se requiere que las bases de datos cumplan con criterios de integridad, confiabilidad y normalización.

Para la aplicación de los algoritmos de clustering es necesario definir el número óptimo de clústers para que se realice una adecuada agrupación de los datos y esta sea compacta y exacta, mediante la aplicación de estos algoritmos se obtendrán agrupaciones que permitan diferenciar una agrupación con otra es así como en el análisis del modelo 1 de la tabla estudiantes se obtienen 3 clústers, el clúster 2 se diferencia por el estado civil (casado) con un total de 664 estudiantes, en el clúster 3 se diferencia la etnia (mestizo) con 2764 estudiantes y en el clúster 1 se diferencia por el promedio (8-10) con un total de 12365 estudiantes además del análisis de los pesos de las variables se pudo identificar cuáles son las que más inciden en este modelo obteniendo como resultados el tipo de parroquia, etnia y el estado civil del estudiante.

Para evaluar la exactitud de los algoritmos de clustering K-Means y K-Medoids se utilizó el criterio de validación del índice de Davies-Bouldin, valores pequeños para el índice DB indica clústers compactos y exactos, se utilizaron nueve modelos aplicados en las tablas: estudiante docente e investigación, los resultados de exactitud obtenidos en el algoritmo K-Means fue 0.949, mientras que en el algoritmo K-Medoids fue un valor de 1.747, por lo que se pudo determinar que algoritmo que presenta mayor exactitud es el K-Means.

6. RECOMENDACIONES

- Una de las fases imprescindibles de la metodología CRISP- DM es la de comprensión del negocio debido a que se debe tener en claro los objetivos de la minería de datos para seleccionar adecuadamente los atributos de la base de datos.
- Hacer un análisis exhaustivo de los datos siguiendo de manera ordenada cada una de las fases de la metodología CRISP-DM como los son la limpieza, construcción y formateo de los datos.
- Revisar la bibliografía actualizada sobre las plataformas más utilizadas en minería de datos permite determinar cuál es la plataforma más adecuada para desarrollar el proceso de minería.
- Para realizar el cálculo del número óptimo de clústers en los algoritmos se recomienda utilizar la plataforma R que facilita y disminuye considerablemente el tiempo al momento de realizar este cálculo.

7. REFERENCIAS BIBLIOGRÀFICAS

Amat, J. (09 de 2017). *Clustering y heatmaps: aprendizaje no supervisado*. Obtenido de rpub: https://rpubs.com/Joaquin_AR/310338

Bolshakova, N., & Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Cluster validation techniques for genome expression data.* Department of Computer Science, Trinity College Dublin, Irlanda.

Camana, R. (2016). Potenciales Aplicaciones de la Minería de Datos en Ecuador. *Revista Tecnológica ESPOL*, 14.

Chapman, C. (2007). CRISP-DM 1.0: Stepby-step data mining guide.

Garre, M., Cuadrado, J., & Sicilia, M. (2014). *Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software.* Madrid.

Gordillo, S., Moine, J., & Haedo, A. (2011). *Análisis comparativo de metodologías para la gestión de minería de datos.* Argentina.

Gutiérrez , J., & Molina, B. (2016). Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. *ONTARE*, 19.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (Vol. 3). Massachusetts, United States of America: Elsevier.

Hartigan, J. (1975). *Clustering Algorithms.* New York.

Hernández , E. (2006). *Algoritmo de clustering basado en entropía para descubrir grupos en atributos de tipo mixto.* Instituto Politecnico Nacional, México D.F, México.

Hernández , R., Tomás, V., Felipe, A., & Nuñez, F. (Julio de 2013). Universidad Autónoma del Estado de Hidalgo. *Identificación de estilos de aprendizaje en alumnos universitarios de computación de la Huasteca Hidalguense mediante técnicas de minería de datos*. México, México.

Hernández, R. (2014). Obtenido de Metodología de la Investigación: http://observatorio.epacartagena.gov.co/wp-content/uploads/2017/08/metodologia-de-la-investigacion-sexta-edicion.compressed.pdf

Hurtado, F. (2005). *Segmentación de clientes usando el algoritmo de clustering K-Mean.* Universidad Nacional Mayor de San Marcos, Lima.

Jain, A., Murty, M., & Flynn, P. (1999). Data Clustering: A Review. *CM Comput. Surv.*, 261-323.

Jain, S., Aalam, A., & Doja, M. (2010). K-MEANS CLUSTERING USING WEKA INTERFACE. *Computing For Nation Development*, 6.

Justicia, M. (2017). Nuevas Técnicas de Minería de Textos: Aplicaciones. *Nuevas Técnicas de Minería de Textos: Aplicaciones.* Universidad de Granada, Granada.

KDnuggets. (2014). *CRISP-DM, aún la mejor metodología para proyectos de análisis, minería de datos o ciencia de datos*. Obtenido de KDnuggets: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

León, E. (2014). Métricas para la validación de Clustering. *Métricas para la validación de Clustering.* Universidad Nacional de Colombia, Bogotá.

Logreira, C. (2011). *Minería de datos y su incidencia en la toma de decisiones empresariales en el contexto de CRM* (Vol. 7). Medellin.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook* (Vol. 2). New York: Springer.

Mamani , Z. (2015). *Aplicación De La Mineria De Datos Distribuida Usando Algoritmo De Clustering K-Means Para Mejorar La Calidad De Servicios De Las Organizaciones Modernas.* UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS, Lima.

Mirkin, B. (2019). *Core Partitioning: K-means and Similarity Clustering. Bondgraphen,*. Obtenido de http://sci-hub.tw/10.1007/978-3-030-00271-8_4

Molina, J., & García Jesús. (2006). APLICACIONES PRÁCTICAS UTILIZANDO MICROSOFT EXCEL Y WEKA. *TÉCNICAS DE ANÁLISIS DE DATOS.* Universidad Carlos III, Madrid.

Mostafa, A. (2016). Review of Data Mining Concept and its Techniques. *Innovative Technology*, 9.

Pang-Ning , T., Steinbach, M., & Kumar, V. (2006). *Intorduction to Data Mining.* New York: Addison Wesley.

Peña, A. (2014). Educational data mining: A survey and a data mining-based analysis ofrecent works. Expert Systems with Applications. *ELSEVIER*, 1432-1462.

Perez, C., & Santin, D. (2007). *Minería de datos. Técnicas y herramientas.* Madrid: Paraninfo.

RapidMiner. (2016). *Predictive Analytics, Reimagined*. Obtenido de RapidMiner: https://rapidminer.com/

Riquelme, J., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Inteligencia Artificial.Revista Iberoamericana de Inteligencia Artificial*, 11-18.

Rodrigues, R., Gomes, A., Cavalcanti, J., Dantas, R., & Sedraz , J. (2016). A Comparative Study between Clustering Methods in Educational Data Minin. *IEEE Latin America Transactions*, 8.

Rosado , A., & Verjel , A. (2016). APLICACIÓN DE LA MINERIA DE DATOS EN LA EDUCACION EN LINEA. *Revista Colombiana de Tecnologías de avanzada*, 7.

Rygielski, C., Wang, y.-C., & Yen, D. (2002). Data mining techniques for customer relationship management. *lsevier*, 20.

Sinnexus. (2019). Obtenido de Datamining (Minería de datos): https://www.sinnexus.com/business_intelligence/datamining.aspx

Suárez, L. (2014). *Técnicas de minería de datos para la detección y prevención del lavado de activos y la financiación del terrorismo (LA/FT).* Unidad de Información y Análisis Financiero, Bogota, Colombia.

Universidad Nacional De Chimborazo. (30 de Enero de 2019). Reglamento de Evaluación Integral De Desempeño Del Personal Académico de la Unach. Riobamba, Chimborazo, Ecuador.

Vargas, L., Farfán, J., Rodriguéz, M., Aramayo, F., Flores, H., & Lopez, V. (2016). Comparación de las principales herramientas de Data Mining y Análisis de Sábanas Telefónicas. *II JORNADA ARGENTINA DE TECNOLOGÍA, INNOVACIÓN Y CREATIVIDAD*, 12.

Weiss, S. M., & Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide.* San Fracisco, USA: Morgan Kaufmann Publishers Inc.

ANEXOS

ANEXO 1: EVALUACIÓN DE LA SITUACIÓN

Inventario de recursos

Tabla 14*: Recursos Software*

Software	Utilidad
Talend Data Quality Web	Proporciona herramientas para realizar el proceso de calidad de datos.
RapidMiner Studio 9.1	Proporciona herramientas para realizar las tareas de minería de datos.

Tabla 15*: Recursos Hardware*

Marca	Modelo	Procesador	Memoria RAM	Disco duro	Sistema Operativo
Toshiba	Harman / kardon	Intel Core I7 4ta 2.50 Gz	8 GB	1 TB	Windows 10 pro

Fuente de datos

La fuente de información es una base de datos del sistema académico de la UNACH (SICOA) en un archivo Excel que contiene información académica y personal de los estudiantes y docentes de la institución.

ANEXO 2: PLAN DE PROYECTO

Tabla 16*: Plan de Proyecto*

Etapa	Tiempo	Recursos
Comprensión del Negocio	2 semanas	Investigador
Comprensión de los Datos	2 semanas	Investigador
Preparación de los Datos	3 semanas	Asesor de minería de datos e investigador
Modelado	2 semanas	Asesor de minería de datos e investigador
Evaluación	1 semana	Asesor de minería de datos e investigador
Distribución	1 semana	Asesor de minería de datos e investigador

Evaluación inicial de herramientas y técnicas

Talend data Quality ayudará a realizar el proceso de calidad de los datos y RapidMiner Studio 9.1 para realizar el proceso de minería de datos, además de la base de datos entregada por parte de la institución.

En cuanto a las técnicas de minería de datos para la extracción del conocimiento, RapidMiner Studio 9.1 nos ofrece algoritmos de clustering como K-means, K-medoid, X-means, DBscan, EM entre otros.

ANEXO 3: DESCRIPCIÓN DE LOS DATOS

Descripción de la Tabla Estudiante

Tabla 17: *Atributos de la tabla estudiante*

Atributos	Tipo	Descripción
Estudiante ID	Numérico	Identificador único del estudiante
Fecha Nacimiento	Fecha	Fecha de nacimiento del estudiante
Estado Civil	Texto	Estado Civil del estudiante
Orientación Sexual	Texto	Atracción afectiva del estudiante.
Genero	Texto	Tipo de Género del estudiante.
Etnia	Texto	Etnia de la cual se considera el estudiante.
Nacionalidad Indígena	Texto	Indica si algún estudiante proviene de alguna nacionalidad indigna del país.
Institución Educativa	Texto	Institución secundaria de donde proviene el estudiante
Enfermedad Catastrófica Extraña	Texto	Indica si el estudiante posee alguna enfermedad.
Tipo Discapacidad	Texto	Indica si el estudiante tiene algún tipo de discapacidad.
Actividad Cultural	Texto	Indica si el estudiante realiza alguna actividad cultural.
Numero Integrantes Hogar	Numérico	Número de integrantes que existen en la familia del estudiante
País Nacimiento	Texto	País de nacimiento del estudiante.
Provincia Nacimiento	Texto	Provincia de nacimiento del estudiante
Cantón Nacimiento	Texto	Cantón de nacimiento del estudiante
País Procedencia	Texto	País de donde procede el estudiante
Provincia Procedencia	Texto	Provincia de donde procede el estudiante
Cantón Procedencia	Texto	Cantón de procedencia del estudiante.
País Dirección	Texto	Dirección del país donde reside el estudiante.
Dirección Provincia	Texto	Dirección de la provincia donde reside el estudiante.
Dirección Cantón	Texto	Dirección del cantón donde reside el estudiante.

Atributos	Tipo	Descripción
Parroquia	Texto	Parroquia donde reside el estudiante.
Tipo Parroquia	Texto	Tipo de parroquia de donde procede el estudiante (rural, urbana)
Numero Hermanos	Numérico	Numero de hermanos que tiene el estudiante
Ingresos Padre	Numérico	Ingresos mensuales del padre del estudiante.
Ingresos Madre	Numérico	Ingresos mensuales de la madre del estudiante.
Ocupación Madre	Texto	Ocupación de la madre del estudiante
Ocupación Padre	Texto	Ocupación del padre del estudiante
Total Ingresos Padres	Numérico	Total de ingresos mensuales de los padres del estudiante.
Número Dependen Ingresos	Numérico	Número de personas que dependen de los ingresos de los padres del estudiante.
Tipo Vivienda	Texto	Indica si la vivienda es de propiedad del estudiante o es en alquiler.
Tipo Construcción	Texto	Indica si la vivienda del estudiante es de construcción mixta, ladrillo o bloque.
Ocupación	Texto	Indica si el estudiante hace actividades extra aparte de estudiar.
Total Ingresos	Numérico	Ingresos mensuales del estudiante.
Numero Hijos	Numérico	Número de hijos que posee el estudiante
Ocupación Cónyuge	Numérico	Ocupación del conyugue del estúdiate.
Ingresos Cónyuge	Numérico	Ingresos mensuales del cónyuge del estudiante.
Total Ingresos Estudiante	Numérico	Total de ingresos mensuales del estudiante.
Personas Dependen Ingresos	Numérico	Número de personas que dependen de los ingresos del estudiante.

Descripción de la Tabla Estudiante Rendimiento.

Tabla 18: *Atributos de la tabla estudiante rendimiento*

Atributos	Tipo	Descripción
Estudiante ID	Numérico	Identificador del estudiante.
Facultad	Texto	Indica la facultad a la que pertenece el estudiante,
Carrera	Texto	Indica la carrera a la que pertenece el estudiante.
Situación Actual	Texto	Muestra si el estudiante se encuentra graduado o no.
Nivel	Texto	Semestre del estudiante.
Periodo	Texto	Periodo en el que se matriculo de determinado semestre.
Promedio	Numérico	Promedio general que tuvo en todo ese semestre.

Descripción de la tabla Docente.

Tabla 19: *Atributos de la tabla docente*

Atributos	Tipo	Descripción
Cedula	Texto	Número de cédula del docente.
País	Texto	País de procedencia.
Nacionalidad	Texto	Nacionalidad según el país de procedencia.
Fecha Nacimiento	Fecha	Fecha de nacimiento del docente.
Estado Civil	Texto	Estado civil del docente.
Sexo	Texto	Sexo del docente.
Etnia	Texto	Etnia del docente.
Tipo Sangre	Texto	Tipo de sangre del docente.
Grupo GLBTI	Texto	Grupo LGBTI al que pertenece el docente.
Nacionalidad Indígena	Texto	Nacionalidad indígena del docente.
País	Texto	País en el que radica actualmente.
Cantón	Texto	Cantón en el que radica actualmente.
Parroquia	Texto	Parroquia en la que radica actualmente.
Número Hijos	Numérico	Número de hijos que tiene el docente.
Nivel Instrucción	Texto	Tipo de título académico del docente.
País	Texto	País en el que se obtuvo el título.
Tiempo Estudio	Texto	Tiempo que demoró en obtener el título.
Modalidad	Texto	Modalidad de estudio.
Área	Texto	Área en la que obtuvo el título.
Subárea	Texto	Subárea en la que obtuvo el título.
Campo	Texto	Campo en el que obtuvo el título.
Está Cursando	Texto	Si se encuentra estudiando actualmente,

Institución Educativa	Texto	Institución educativa en la que se obtuvo el título.
Título	Texto	Título que obtuvo.
No Eventos Aprobados	Numérico	Numero de eventos aprobados del docente.
No Eventos Asistidos	Numérico	Numero de eventos asistidos por el docente.
Horas Eventos Aprobados	Numérico	Horas de eventos aprobados del docente.
Horas Eventos Asistidos	Numérico	Horas de eventos asistidos por el docente.
No Eventos Nacionales	Numérico	Numero de eventos nacionales del docente.
No Eventos Internacionales	Numérico	Numero de eventos internacionales docente.
Experiencia Privada	Texto	Si posee experiencia en entidades privadas.
Experiencia Pública	Texto	Si posee experiencia en entidades públicas.
Familiar Sustituto	Texto	Familiar sustituto en el trabajo.
Enfermedad Catastrófica	Texto	Si posee alguna enfermedad catastrófica.
Tiene Discapacidad	Texto	Si posee alguna discapacidad.
Gestación Lactancia	Texto	Si se encuentra en estado de gestación o lactancia.

Descripción de la Tabla Docente Información Académica.

Tabla 20: *Atributos de la tabla docente información académica*

Atributos	Tipo	Descripción
Numero Documento	Texto	Contiene el número de cedula del docente.
Facultad	Texto	Indica la facultad a la que pertenece el docente.
Carrera	Texto	Indica la carrera a la que pertenece el docente.
Periodo	Texto	Muestra el periodo en que dio clases en esa facultad y carrera el docente.
Actividad Académica	Texto	Actividad académica que realiza el docente en la institución.
Horas Actividad Académica	Numérico	Horas realizadas de actividad académica.
Horas Clase	Numérico	Horas de clase impartidas por el docente.

Descripción de la Tabla Evaluación Docente.

Tabla 21: *Atributos de la tabla evaluación docente*

Atributos	Tipo	Descripción
Usuario Evaluado	Texto	Contiene el número de usuario del docente evaluado.
Cedula	Numérico	Contiene el número de cedula del docente evaluado.
Tipo Evaluación	Texto	Muestra el tipo de evaluación que se le realizó al docente (autoevaluación, heteroevaluación, etc.)
Componente	Texto	Indica el componente (docencia, gestión, investigación) en el que fue evaluado el docente.
Resultado Final	Texto	Contiene la calificación del docente en cada una de las evaluaciones.
Periodo	Texto	Indica el periodo en el que fue evaluado el docente.

Descripción de la tabla Investigación

Tabla 22*: Atributos de la tabla investigación*

Atributos	Tipo	Descripción
Estado Publicación	Texto	Describe el estado actual de la publicación.
Tipo Publicación	Texto	Indica a que tipo pertenece la publicación realizada.
Titulo	Texto	Contiene el título del proyecto de investigación.
Revista	Texto	Nombre de la revista en la que fue publicada.
Cedula	Texto	Número de cedula del docente que realiza la investigación.
Rol Institución	Texto	Rol que cumple en la institución el docente que realizó la investigación.
Sexo	Texto	Sexo del docente.
Tipo Autor	Texto	Si el autor es docente o no.
Orden Autor	Numérico	Orden de autor en la investigación.
Nombres	Texto	Nombres del docente.
Apellido Materno	Texto	Apellido materno del docente.
Apellido Paterno	Texto	Apellido paterno del docente.
Área de Investigación	Texto	Área en la que se realizó la investigación.
Línea de Investigación	Texto	Línea en la que se realizó la investigación.
Campo Amplio	Texto	Campo amplio en la que se realizó la investigación.

Atributos	Tipo	Descripción
Campo Detallado	Texto	Campo detallado en la que se realizó la investigación.
Campo Especifico	Texto	Campo específico en la que se realizó la investigación.
Año	Fecha	Año en el que se realizó la investigación.
Año-mes de Publicación	Fecha	Año y mes en el que se realizó la investigación.
Año-mes de Registro	Fecha	Año y mes de registro de la investigación.
Año-mes Registro Mod	Fecha	Año y mes de registro de la investigación.
Fecha Aceptación	Fecha	Fecha en la que fue aceptada la investigación.
Fecha Actualización	Fecha	Fecha en la que fue actualizada la investigación.
Fecha de Registro	Fecha	Fecha en la que fue registrada la investigación.
Fecha Publicación	Fecha	Fecha en la que fue publicada la investigación.
Facultad	Texto	Facultad en la que se realizó la investigación.
Carrera	Texto	Carrera en la que se realizó la investigación.
Código Carrera	Numérico	Código de la carrera en la que se realizó la investigación.
Ciudad de Publicación	Texto	Ciudad en la que se realizó la publicación de la investigación.
Comité Científico u Organizador	Texto	Comité científico u organizador.
Comité Editorial o Experto	Texto	Comité editorial o experto.
Congreso o Seminario	Texto	Congreso o seminario.
Es Editorial de Prestigio	Texto	Indica si la editorial es de prestigio o no.
Es Editorial de Prestigio	Numérico	Indica si la editorial es de prestigio o no.
Existe Aprobación de comisión	Numérico	Indica si fue aprobado o no por una comisión.
Existe Comité Científico u organizador	Numérico	Indica si existe un comité científico u organizador.
Existe Comité Editorial	Numérico	Indica si existe un comité editorial.
Existe Procedimiento selectivo	Numérico	Indica si existe un procedimiento selectivo

Atributos	Tipo	Descripción
Existe Revisión por pares externos	Numérico	Indica si existe una revisión por parte de pares externos.
Forma Publicación	Texto	Señala las características de la publicación.
Forma Publicación en Articulo Completo	Texto	Indica si forma o no publicación en artículo completo.
Observaciones Autor	Texto	Muestran las observaciones del autor.
Observaciones de Comisión	Texto	Muestra las observaciones de la comisión.
Observaciones Generales	Texto	Muestra las observaciones generales.
Listado de Revistas Senescyt	Texto	Si se encuentra o no en las revistas del listado de la SENESCYT.
Observaciones Publicación	Texto	Señalan las observaciones realizadas en la publicación.
Doaj	Texto	Indica si la publicación se encuentra disponible en DOAJ.
Doi	Texto	Identificador digital de la publicación.
Ebsco	Texto	Indica si la publicación se encuentra disponible en EBSCO.
Estado Personal Académico	Numérico	Estado personal académico del docente.
Isbn	Texto	Código ISBN que lo identifica.
Isi webKnowledge	Texto	Indica si la publicación se encuentra disponible en isi web Knowledge.
Issn	Texto	Código ISNN que lo identifica.
Jstor	Texto	Indica si la publicación se encuentra disponible en JSTOR.
Latindex	Texto	Indica si la publicación se encuentra disponible en LATINDEX.
Libros o Capítulos de libros revisados por pares	Numérico	Indica el número de libros o capítulos revisados por pares.
Lilacs	Numérico	Indica si la publicación se encuentra disponible en LILACS.
Nacional o internacional	Texto	Señala si la publicación es de tipo nacional o internacional.
Oaji	Texto	Indica si la publicación se encuentra disponible en OAJI.
País	Texto	País en el que se realizó la investigación.

Atributos	Tipo	Descripción
Procedimiento selectivo	Texto	Señala la resolución final sobre la investigación.
Proquest	Texto	Indica si la publicación se encuentra disponible en PROQUEST
Redalyc	Texto	Indica si la publicación se encuentra disponible en REDALYC.
Revisión por pares externos	Texto	Resolución que tomaron los pares externos.
Scielo	Texto	Indica si la publicación se encuentra disponible en Scielo.
Scimago Journal Rank	Texto	Indica si la publicación se encuentra disponible en Scielo.
Sjr	Numérico	Indica el índice de impacto de la investigación.
Paginas	Texto	Número de páginas de la publicación.
Volumen	Numérico	Volumen de la revista en el que fue publicado.
Organismo de Afiliación	Texto	Organismo al que se encuentra afiliado el docente.

ANEXO 4: EXPLORACIÓN DE LOS DATOS

a) Tabla Estudiante

- La figura 11 muestra el número de estudiantes por género que existen en la Universidad Nacional de Chimborazo.

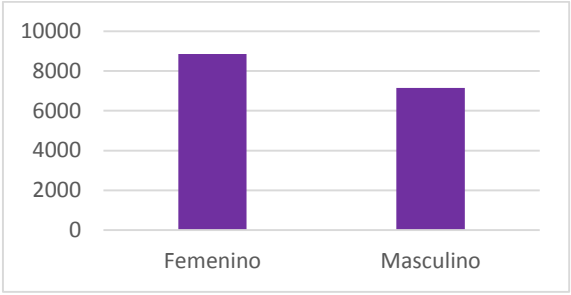

Figura 11. Número de estudiantes por género

Como se puede apreciar en la figura 14 en la UNACH existen 8861 estudiantes de género femenino y 7147 estudiantes de género masculino.

- La figura 12 muestra el número de estudiantes que existe en cada una de las facultades de la Universidad Nacional de Chimborazo, en la Facultad de Ciencias de la Salud existen 4984 estudiantes, en la Facultad de ciencias Políticas y Administrativas existen 3878 estudiantes, en la Facultad de Ciencias de la Educación, Humanas y Tecnológicas existen 3341 estudiantes y en la Facultad de Ingeniería 4054 estudiantes.

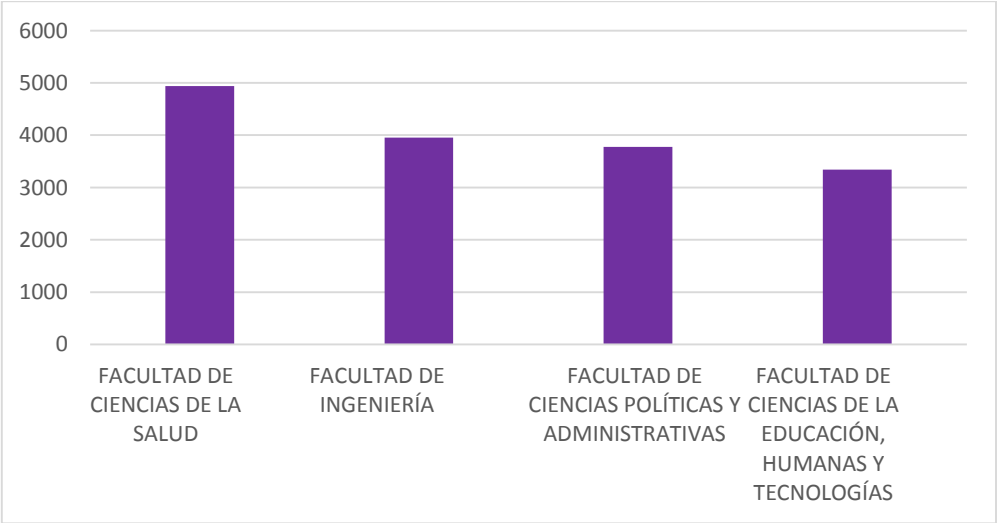

Figura 12. Número de estudiantes por facultad

- En la figura 13 se puede observar que el estado civil que predomina en los estudiantes de la universidad nacional de Chimborazo es soltero.

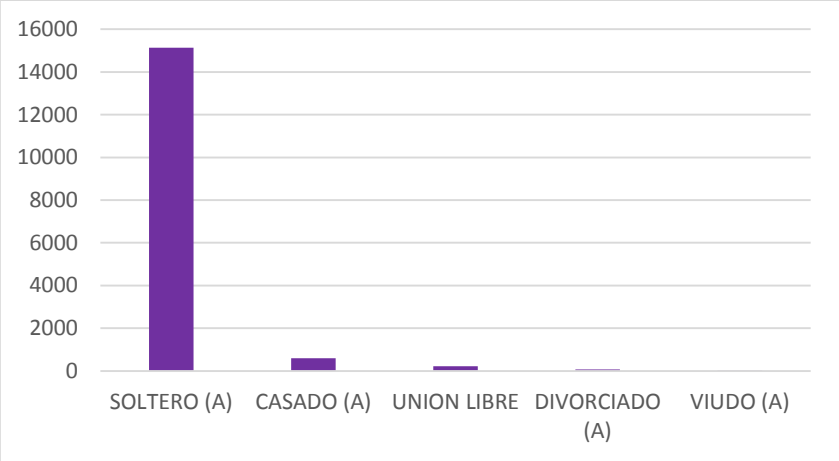

Figura 13. Número de estudiantes por estado civil

Se puede observar en la figura 13 que existen 15131 estudiantes de estado civil solteros, 589 casados, 219 que se encuentran en unión libre, 65 estudiantes divorciados y 4 estudiantes viudos.

b) Tabla Docentes

- La figura 14 muestra la distribución de los docentes por estado civil.

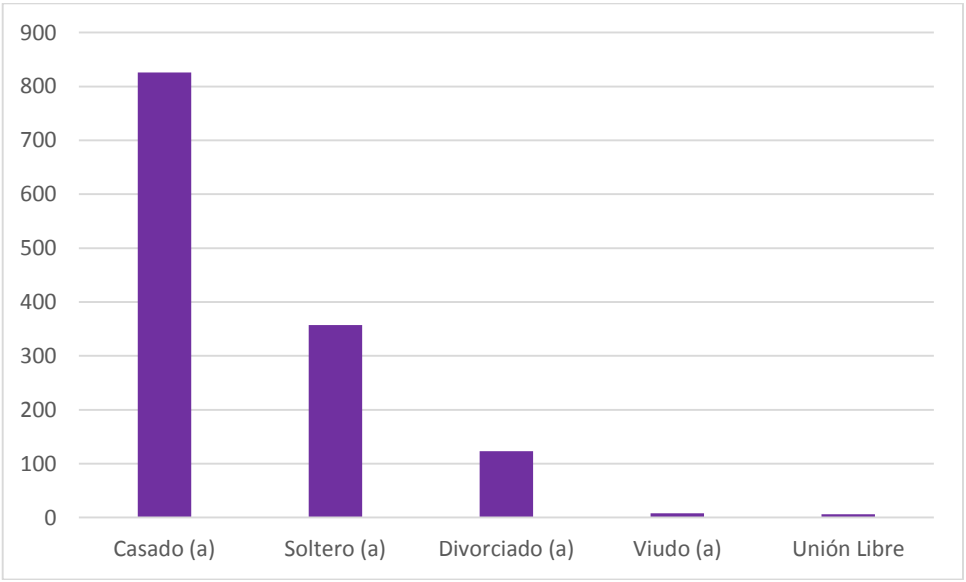

Figura 14. Número de docentes por estado civil

Se puede observar en la figura 14 que existen 826 docentes de estado civil casados, 357 solteros, 6 que se encuentran en unión libre, 123 docentes divorciados y 8 docentes viudos.

- La figura 15 muestra la distribución de los docentes por género.

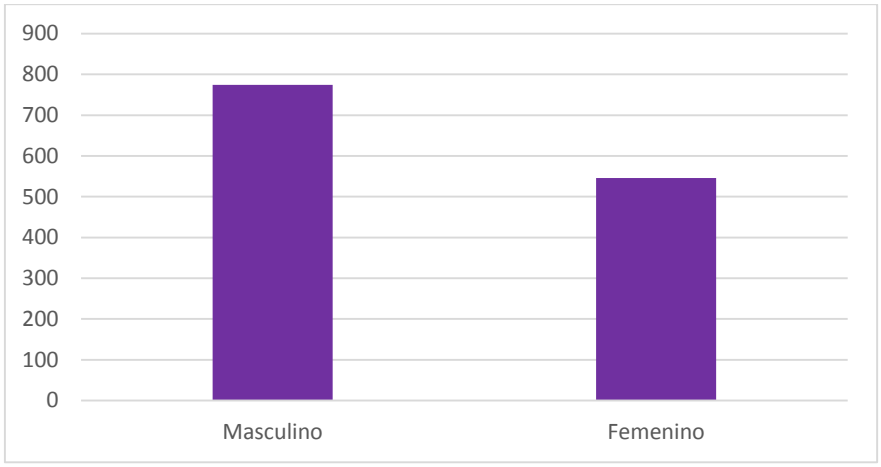

Figura 15. Número de docentes por género

Como se puede apreciar en la figura 15 en la UNACH existen 546 docentes de género femenino y 774 docentes de género masculino.

- La figura 16 muestra el número de estudiantes que existe en cada una de las facultades de la Universidad Nacional de Chimborazo, en la Facultad de Ciencias de la Salud existen 515 docentes, en la Facultad de ciencias Políticas y Administrativas existen 205 docentes, en la Facultad de Ciencias de la Educación, Humanas y Tecnológicas existen 266 y en la Facultad de Ingeniería 331 docentes.

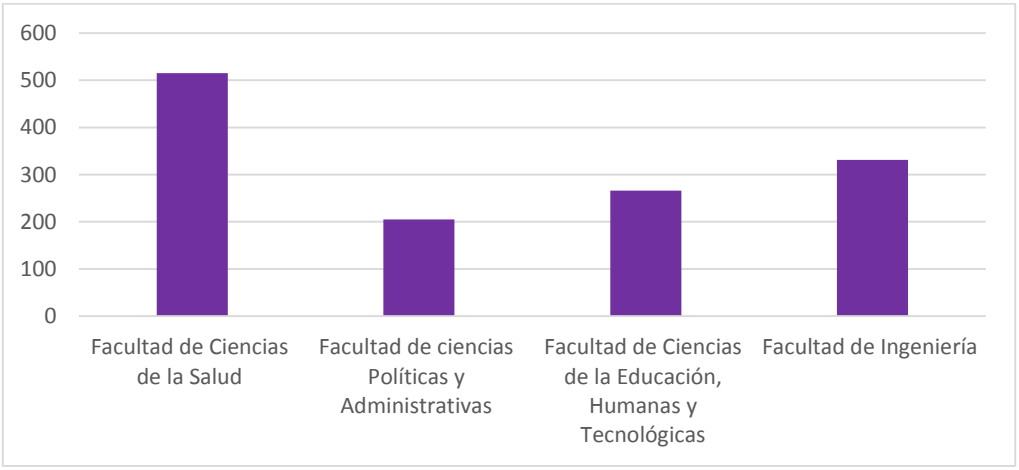

Figura 16. Número de docentes por facultad

c) Tabla Investigación

- La figura 17 muestra la distribución de los docentes por nivel de instrucción.

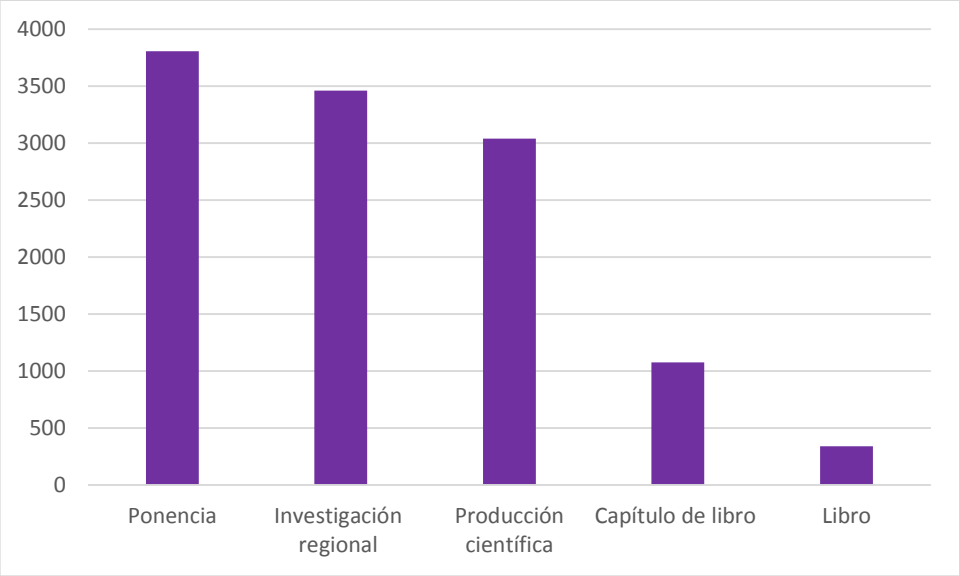

Figura 17. Tipo de Publicaciones

Se puede observar en la figura que existe un total de producción científica de 3040 artículos además de investigación regional 3463, publicaciones de capítulo de libro 1077 y 341 libros publicados

ANEXO 5: VERIFICAR LA CALIDAD DE LOS DATOS

La descripción de cada una de las tablas se muestra a continuación:

Tabla Estudiante

Tabla 23*: Calidad de datos tabla estudiante*

Campo	Valores nulos		Valores válidos		Valores Inválidos	
	Número	%	Numero	%	Numero	%
Estudiante ID	0	0,00	15766	98,49	242	1,51
Porcentaje Discapacidad	15901	99,33	107	0,67	0	0,00
Numero Integrantes Hogar	1308	8,17	14700	91,83	0	0,00
Numero Hermanos	66	0,41	15942	99,59	0	0,00
Ingresos Padre	0	0,00	16008	100,00	0	0,00
Ingresos Madre	0	0,00	16008	100,00	0	0,00
Total Ingresos Padres	0	0,00	16008	100,00	0	0,00
Numero Dependen Ingresos	0	0,00	16008	100,00	0	0,00
Valor Mensual Servicios	0	0,00	16008	100,00	0	0,00
Total Ingresos	0	0,00	16008	100,00	0	0,00
Numero Hijos	0	0,00	16008	100,00	0	0,00
Ingresos Cónyuge	0	0,00	16008	100,00	0	0,00
Total Ingresos Estudiante	0	0,00	16008	100,00	0	0,00
Personas Dependen Ingresos	0	0,00	16008	100,00	0	0,00
Fecha Nacimiento	0	0,00	16008	100,00	0	0,00
Estado Civil	0	0,00	16008	100,00	0	0,00
Orientación Sexual	9	0,06	15999	99,94	0	0,00
Sexo	0	0,00	16008	100,00	0	0,00
Género	0	0,00	16008	100,00	0	0,00
Etnia	9	0,06	15999	99,94	0	0,00
Nacionalidad Indígena	9	0,06	15999	99,94	0	0,00

Campo	Valores nulos		Valores válidos		Valores Inválidos	
Institución Educativa	0	0,00	16008	100,00	0	0,00
Tipo	0	0,00	16008	100,00	0	0,00
Enfermedad Catastrófica	0	0,00	16008	100,00	0	0,00
Tipo Discapacidad	9	0,06	15999	99,94	0	0,00
Actividad Deportiva	1560	9,75	14448	90,25	0	0,00
Actividad Cultural	1694	10,58	14314	89,42	0	0,00
País Nacimiento	0	0,00	15974	99,79	34	0,21
Prov. Nacimiento	0	0,00	16008	100,00	0	0,00
Cantón Nacimiento	5740	35,86	10268	64,14	0	0,00
País Procedencia	5519	34,48	10406	65,00	83	0,52
Provincia Procedencia	5629	35,16	10379	64,84	0	0,00
Cantón Procedencia	5629	35,16	10379	64,84	0	0,00
Tipo Parroquia	5416	33,83	10592	66,17	0	0,00
Ocupación	14896	93,05	16008	100,00	0	0,00

Tabla Rendimiento Académico

Tabla 24*: Calidad de datos tabla rendimiento académico*

Campo	Valores nulos		Valores válidos		Valores Inválidos	
	Número	**%**	**Numero**	**%**	**Numero**	**%**
Estudiante ID	0	0,00	85685	98,37	1420	1,63
Facultad	0	0,00	87105	100,00	0	0,00
Carrera	0	0,00	87105	100,00	0	0,00
Situación Actual	0	0,00	87105	100,00	0	0,00
Nivel	0	0,00	87105	100,00	0	0,00
Período	0	0,00	87105	100,00	0	0,00
Promedio	296	0,34	86809	99,66	0	0,00

Tabla Docentes

Tabla 25*: Calidad de datos tabla docentes*

Campo	Valores nulos		Valores válidos		Valores Inválidos	
	Número	%	Numero	%	Numero	%
Cédula	0	0,00	3253	79,40	844	20,6
País	13	0,32	4046	98,76	38	0,93
Nacionalidad	3	0,07	4094	99,93	0	0,00
Fecha Nacimiento	0	0,00	4097	100,00	0	0,00
Número Hijos	0	0,00	4097	100,00	0	0,00
Estado Civil	0	0,00	4097	100,00	0	0,00
Sexo	0	0,00	4097	100,00	0	0,00
Etnia	550	13,42	3547	86,58	0	0,00
Tipo Sangre	556	13,57	3541	86,43	0	0,00
Grupo GLBTI	0	0,00	4097	100,00	0	0,00
Nacionalidad Indígena	560	16,67	3537	86,33	0	0,00
Cantón	556	13,57	3541	86,43	0	0,00
Parroquia	556	13,57	3541	86,43	0	0,00
Nivel Instrucción	16	0,39	4081	99,61	0	0,00
Modalidad	1058	25,82	2725	66,51	314	7,66
Área	1847	45,08	2250	54,92	0	0,00
Subárea	1847	45,08	2250	54,92	0	0,00
Campo	1847	45,08	2250	54,92	0	0,00
Está Cursando	1058	25,82	2725	66,51	0	0,00
Institución Educativa	1	0,02	4096	99,98	0	0,00
Titulo	1	0,02	4096	99,98	0	0,00
Experiencia Privada	0	0,00	4097	100,00	0	0,00
Experiencia Pública	0	0,00	4097	100,00	0	0,00

Campo	Valores nulos		Valores válidos		Valores Inválidos	
Familiar Sustituto	0	0,00	4097	100,00	0	0,00
Enfermedad Catastrófica	2116	51,65	1981	48,35	0	0,00
Tiene Discapacidad	115	2,81	3982	97,19	0	0,00
Lactancia	3248	79,28	849	20,72	0	0,00
Tiempo Estudio	0	0,00	4097	100,00	0	0,00
N.º Eventos Aprobados	0	0,00	4097	100,00	0	0,00
N.º Eventos Asistidos	0	0,00	4097	100,00	0	0,00
Horas Eventos Aprobados	930	22,70	3167	77,30	0	0,00
Horas Eventos Asistidos	2191	53,48	1906	46,52	0	0,00
N.º Eventos Nacionales	0	0,00	4097	100,00	0	0,00
N.º Eventos Internacional	0	0,00	4097	100,00	0	0,00

Tabla Información Académica

Tabla 26*: Calidad de datos docente información académica*

Campo	Valores nulos		Valores válidos		Valores Inválidos	
	Número	**%**	**Numero**	**%**	**Numero**	**%**
Numero Documento	0	0,00	15479	80,00	3857	20,0
Facultad	0	0,00	19335	100,00	0	0,00
Carrera	2717	14,00	16619	86,00	0	0,00
Periodo	0	0,00	19335	100,00	0	0,00
Actividad Académica	2717	14,00	16619	86,00	0	0,00
Horas Actividad Académica	0	0,00	19335	100,00	0	0,00

Campo	Valores nulos		Valores válidos		Valores Inválidos	
Horas Clase	0	0,00	19335	100,00	0	0,00

Tabla Evaluación Docente

Tabla 27*: Calidad de datos tabla evaluación docente*

Campo	Valores nulos		Valores válidos		Valores Inválidos	
	Número	%	Numero	%	Numero	%
Usuario Evaluado	0	0,00	15276	100,00	0	0,00
Tipo Evaluación	0	0,00	15276	100,00	0	0,00
Componente	0	0,00	15276	100,00	0	0,00
Periodo	0	0,00	15276	100,00	0	0,00
Resultado Final	0	0,00	15276	100,00	0	0,00

Tabla Investigación

Tabla 28*: Calidad de datos tabla investigación*

Campo	Valores nulos		Valores válidos		Valores Inválidos	
	Número	%	Numero	%	Numero	%
Estado Publicación	0	0,00	12050	100,00	0	0,00
Tipo Publicación	0	0,00	12050	100,00	0	0,00
Revista	0	0,00	12050	100,00	0	0,00
Cédula	0	0,00	8270	68,63	3780	31,37
Rol Institución	4191	34,78	7859	65,22	0	0,00
Sexo	0	0,00	12050	100,00	0	0,00
Tipo Autor	0	0,00	12050	100,00	0	0,00
Orden Autor	0	0,00	12050	100,00	0	0,00
Nombres	0	0,00	12050	100,00	0	0,00
Apellido Materno	1408	11,68	10642	88,32	0	0,00
Apellido Paterno	0	0,00	12050	100,00	0	0,00
Área de Investigación	22	0,18	12028	99,82	0	0,00
Línea de Investigación	22	0,18	12028	99,82	0	0,00
Año	0	0,00	12050	100,00	0	0,00
Facultad	3949	32,77	8101	67,23	0	0,00

Campo	Valores nulos		Valores válidos		Valores Inválidos	
Carrera	3949	32,77	8101	67,23	0	0,00
Ciudad Publicación	183	1,52	11867	98,48	0	0,00
Es Editorial de Prestigio	2050	17,01	10000	82,99	0	0,00
Existe Aprobación de Comisión	9446	78,39	2604	21,61	0	0,00
Existe Comité Científico u Organizador	9446	78,39	2604	21,61	0	0,00
Existe Comité Editorial	9446	78,39	2604	21,61	0	0,00
Existe Procedimiento Selectivo	9446	78,39	2644	21,61	0	0,00
Existe Revisión por Pares Externos	9446	78,39	2604	21,61	0	0,00
Listado de Revistas SENESCYT	2050	17,01	10000	82,99	0	0,00
Estado Personal Académico	6077	50,43	5973	49,57	0	0,00
ISBN	7957	66,03	4093	33,97	0	0,00
ISSN	7066	58,64	4984	41,36	0	0,00
Nacional o Internacional	2050	17,01	10000	82,99	0	0,00
Organismo de afiliación	2050	17,01	10000	82,99	0	0,00
SJR	10855	90,08	1195	9,92	0	0,00
Volumen	6555	54,40	5495	82,99	0	0,00

ANEXO 6: LIMPIEZA DE LOS DATOS

En la tabla 31 se muestra los atributos que se descartaron para el análisis del proyecto planteado:

Tabla 29*: Limpieza de datos*

Tabla	Columnas
Estudiante	Fecha de Nacimiento, Orientación sexual, Institución Educativa, tipo, Enfermedad Catastrófica Extraña, Tipo Discapacidad, Porcentaje Discapacidad, País Nacimiento, Cantón Nacimiento, Parroquia, País Dirección, País Procedencia, Tipo Vivienda, Tipo construcción, Tiene vehículo, Ocupación Padre, Ocupación Madre, Servicio Agua Potable, Servicio electricidad, Servicio Teléfono, Servicio Internet, Servicio TV Pagada, Valor Mensual Servicios, Ocupación Conyugue, Ingresos Conyugue.
Estudiante Rendimiento	Estudiante ID, Carrera, Situación Actual, Nivel.
Docente	Fecha de Nacimiento, Tipo Sangre, Grupo GLBTI, País, Cantón, Parroquia, País, Tiempo Estudio, Modalidad, Área, Subárea, Campo, Está Cursando, Institución Educativa, Titulo, Experiencia Privada, Experiencia Pública, Familiar Sustituto, Enfermedad Catastrófica, Tiene Discapacidad, Gestación Lactancia.
Docente Inf Académica	Número de documento, Carrera, Actividad Académica.
Evaluación Docente	Usuario Evaluado, Tipo de Evaluación, Componente, Período
Investigación	Título, Cedula, Rol Institución, Sexo, Tipo Autor, Orden Autor, Nombres, Apellido Materno, Apellido Paterno, Año, Año Mes Publicación, Año mes Registro, ciudad de publicación, Existe Comité Científico u Organizador, Existe Comité Editorial, Existe Procedimiento Selectivo, Existe Revisión por Pares Externos, Listado de Revistas SENESCYT, Estado Personal Académico, Organismo de afiliación

Se puede observar que existe varios atributos eliminados de cada tabla esto se debe al estudio de calidad de datos realizado en el apartado 2.4 en donde se detectaron atributos con valores válidos, valores inválidos y espacios en blanco y es por esta razón que se procede a la eliminación de varios atributos.

ANEXO 7: CONTRUIR LOS DATOS

Los nuevos campos creados se muestran y describen a continuación en la tabla 32.

Atributos Derivados

Tabla 30*: Atributos derivados*

Tabla	Campo	Descripción
Estudiantes	Promedio	Este atributo se generó sacando el promedio general de todos los niveles aprobados por cada estudiante.
	Foráneo	Ese atributo se generó asignando un valor de "Si" a estudiantes que pertenecen a la provincia de Chimborazo y "No" a los estudiantes que no pertenezcan a la provincia mencionada.
	Ponderación Promedio	En este campo se generó la ponderación del promedio general de los estudiantes como se muestra en la tabla 22.
Docente	Horas Clase	Este atributo contiene un promedio general de las horas clase que dicta el docente en las diferentes carreras de la institución.
	Horas Actividad Académica	Este atributo contiene un promedio general de las horas de actividad académica que tiene asignado el docente
	Resultado Final de la Evaluación	Este atributo es una combinación y un promedio general de las evaluaciones a los docentes como lo son heteroevaluación, coevaluación, autoevaluación.
	Equivalencia Calificación	Este atributo contiene la ponderación del resultado final de la evaluación misma que está basada en los valores que se detalla en la Tabla 23.
Investigación	Tiene Publicaciones	En este atributo se puede identificar si el docente tiene o no publicaciones científicas.
	Publicaciones Producción científica	Este atributo contiene el número de publicaciones de producción científica que tiene cada docente.
	Publicaciones regionales revista	Este atributo contiene el número de publicaciones regional revista que tiene cada docente.
	Publicaciones Libro	Este atributo contiene el número de publicaciones de libros que tiene cada docente.
	Publicaciones Ponencia	Este atributo contiene el número de publicaciones de ponencias que tiene cada docente.

A continuación, en las tablas 33 y 34, se describen ponderaciones de la calificación del estudiante y evaluación docente respectivamente:

Tabla 31*: Ponderación promedio estudiantes*

Rango de calificación	Ponderación
9.00 – 10.00	Excelente
7.00 – 8.99	Bueno
Menos 7.00	Insuficiente

Tabla 32: *Ponderación resultado final evaluación docente*

Rango Calificación	Ponderación
95.00 – 100.00	Excelente
90.00 – 94.99	Muy bueno
80.00 – 89.99	Bueno
70.00 – 79.99	Regular
Menos de 70	Insuficiente

Fuente: *(Universidad Nacional De Chimborazo, 2019)*

ANEXO 8: FORMATEO DE DATOS

. Formateo De Datos Tabla Estudiante

Tabla 33*: Formateo de dato estudiante*

Campo	Descripción	Valor Numérico
Estado Civil	Soltero(a)	1
	Unión libre	2
	Casado(a)	3
	Divorciado(a)	4
	Viudo (a)	5
Género	Masculino	1
	Femenino	2
Etnia	Mestizo/a	1
	Blanco/a	2
	Indígena	3
	Afroecuatoriano/a	4
	Montubio/a	5
	Mulato/a	6
	Negro/a	7
Actividad Deportiva, Actividad Cultural, Foráneo, Tiene Hermanos, Trabaja, Tiene Hijos	Si	1
	No	2
Tipo Parroquia	Urbana	1
	Rural	2

Campo	Descripción	Valor Numérico
Facultad	Facultad de Ciencias de la salud	1
	Facultad de Ciencias de la Educación, Humanas y Tecnologías	2
	Facultad de Ciencias Políticas y Administrativas	3
	Facultad de Ingeniería	4
Ponderación Promedio	Excelente	1
	Bueno	2
	Regular	3

Formateo De Datos Tabla Docente

Tabla 34*: Formateo de datos docente*

Campo	Descripción	Valor Numérico
País	Ecuador	1
	España	2
	Venezuela	3
	Australia	4
Estado Civil	Soltero(a)	1
	Unión libre	2
	Casado(a)	3
	Divorciado(a)	4
	Viudo (a)	5
Género	Masculino	1
	Femenino	2
Etnia	Mestizo/a	1
	Blanco/a	2
	Indígena	3
	Afroecuatoriano/a	4
	Montubio/a	5

Campo	Descripción	Valor Numérico
Nivel de Instrucción	Posgrado PhD	
	Posgrado Maestría	1
	Posgrado Especialidad	2
	Área Salud	3
	Especialidad	4
	Superior Universitaria	5
	Completa	
Facultad	Facultad de Ciencias de la salud	1
	Facultad de Ciencias de la Educación, Humanas y Tecnologías	2
	Facultad de Ciencias Políticas y Administrativas	3
	Facultad de Ingeniería	4
Ponderación Promedio	Excelente	1
	Muy Bueno	2
	Bueno	3
	Regular	4
	Insuficiente	5

Formateo De Datos Tabla Investigación

Tabla 35*: Formateo de datos investigación*

Campo	Descripción	Valor Numérico
Estado Civil	Soltero(a)	1
	Unión libre	2
	Casado(a)	3
	Divorciado(a)	4
	Viudo (a)	5
Género	Masculino	1
	Femenino	2

Campo	Descripción	Valor Numérico
Etnia	Mestizo/a	1
	Blanco/a	2
	Indígena	3
	Afroecuatoriano/a	4
	Montubio/a	5
Nivel de Instrucción	Posgrado PhD	1
	Posgrado Maestría	2
	Posgrado Especialidad Área Salud	3
	Especialidad	4
	Superior Universitaria Completa	5
Facultad	Facultad de Ciencias de la salud	1
	Facultad de Ciencias de la Educación, Humanas y Tecnologías	2
	Facultad de Ciencias Políticas y Administrativas	3
	Facultad de Ingeniería	4
Tiene Publicaciones, Publicación Producción Científica, Publicaciones Regional Revista, Publicaciones Libro, Publicaciones Ponencia	Si	1
	No	2

ANEXO 9: CONSTRUIR EL MODELO

La figura 18 muestra el modelo aplicado a los algoritmos de clustering, cabe mencionar que el componente clustering cambia por cada algoritmo aplicado (K-means, K- medoids).

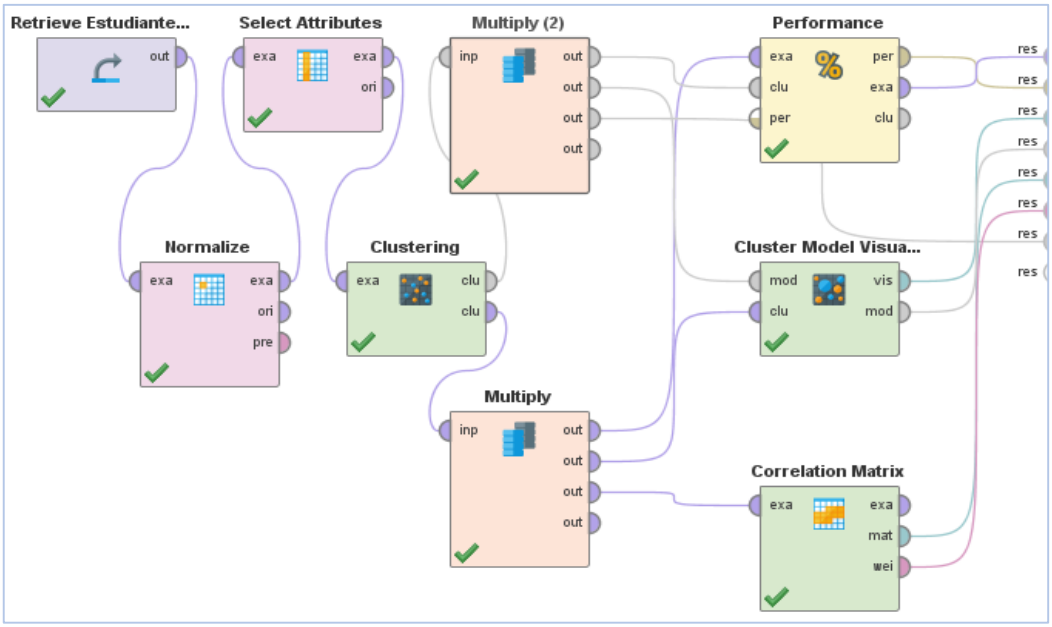

Figura 18.Modelo general clustering

Se definieron 3 objetivos para la minería de datos por lo tanto esta sección va a tener 3 apartados una por cada objetivo.

Objetivo 1

Realizar clústers o grupos de la información de los estudiantes a partir de variables categóricas y numéricas como datos demográficos, rendimiento académico, etc.

Modelo 1 Tabla Estudiante

Conjunto de datos: Estudiantes

Campos analizar: Estado Civil, Etnia, Género, Promedio, Tipo Parroquia

En las tablas 38 y 39 se muestran los resultados de los clústers que genera este modelo para los algoritmos k-means y k-medoids respectivamente:

Tabla 36*: Clústers K-Means tabla estudiante modelo 1*

Clúster	Estado Civil	Género	Etnia	Tipo Parroquia	Promedio	Número de instancias
Clúster 1	-0.201	-0.003	-0.025	-0.471	-0.015	12365
Clúster 2	4.541	0.176	0.462	-0.072	0.232	664
Clúster 3	-0.190	-0.029	0.003	2.124	0.010	2764

Tabla 37: *Clústers K-Medoids tabla estudiante modelo 1*

Clúster	Estado Civil	Género	Etnia	Tipo Parroquia	Promedio	Número de instancias
Clúster 1	-0.230	0.897	-0.270	2.124	-0.025	2866
Clúster 2	-0.230	0.897	-0.270	-0.471	1.165	6642
Clúster 3	-0.230	-1.114	-0.270	-0.471	-0.183	6285

Modelo 2 Tabla Estudiante

Conjunto de datos: Estudiantes

Campos analizar: número de hermanos, número hijos, número integrantes hogar, número dependen de ingresos, promedio, total ingresos, trabaja.

En las tablas 40 y 41 se muestran los resultados de los clústers que genera este modelo para los algoritmos k-means y k-medoids respectivamente:

Tabla 38*: Clústers K-Means tabla estudiante modelo 2*

Clústers	Clúster 1	Clúster 2
Número integrantes hogar	-0.000	-0.046
Numero de hermanos	-0.000	0.810
Numero dependen ingresos	0.000	-0.347
Trabaja	0.000	-0.747
Total de ingresos	-0.021	37.259
Número de hijos	-0.000	0.632
Promedio	-0.000	0.355

Tabla 39*: Clústers K-Medoids tabla estudiante modelo 2*

Clústers	Clúster 1	Clúster 2
Número integrantes hogar	-1.771	-1.771
Numero de hermanos	-0.734	0.424
Numero dependen ingresos	-0.564	1.144
Trabaja	0.238	0.238
Total de ingresos	-0.036	-0.008
Número de hijos	-0.245	-0.245
Promedio	1.165	0.782

Modelo 3 Tabla Estudiante

Conjunto de datos: Estudiantes

Campos analizar: Foráneo, numero de hermanos, número de hijos, total ingresos estudiantes, promedio

En las tablas 42 y 43 se muestran los resultados de los clústers que genera este modelo para los algoritmos k-means y k-medoids respectivamente:

Tabla 40: *Clústers K-Means tabla estudiante modelo 3*

Clústers	Clúster 1	Clúster 2
Foráneo	-0.006	0.203
Numero de hermanos	0.004	-0.127
Número de hijos	-0.156	4.975
Total ingresos estudiante	0	0
Promedio	-0.008	0.261

Tabla 41*: Clústers K-Medoids tabla estudiante modelo 3*

Clústers	Clúster 1	Clúster 2
Foráneo	0.839	-1.192
Numero de hermanos	-0.734	-0.734
Número de hijos	-0.245	-0.245
Total ingresos estudiante	0	0
Promedio	1.165	-0.813
Numero de instancias	8312	7481

Objetivo 2

Segmentar la información de los docentes a partir de datos demográficos, académicos y evaluación del docente.

Modelo 1 Tabla Docente

Conjunto de datos: Docente

Campos analizar: Estado civil, etnia, genero, nivel de instrucción, equivalencia calificación, resultado final evaluación docente.

En las tablas 44 y 45 se muestran los resultados de los clústers que genera este modelo para los algoritmos k-means y k-medoids respectivamente:

Tabla 42*: Clústers K-Means tabla docente modelo 1*

Clústers	Clúster 1	Clúster 2	Clúster 3
Estado civil	0.016	-0.173	-0.200
Genero	0.008	-0.166	-0.026
Etnia	-0.180	4.661	-0.018
Nivel instrucción	-0.031	-0.041	0.716
Resultado Eva Docente	0.191	-0.366	-3.846
Equivalencia calificación	-0.164	0.501	3.146
Numero de instancias	413	16	19

Tabla 43: *Clústers K-Medoids tabla docente modelo 1*

Clústers	Clúster 1	Clúster 2	Clúster 3
Estado civil	0.351	0.351	-1.743
Genero	1.429	-0.698	1.429
Etnia	-0.235	-0.235	-0.235
Nivel instrucción	-0.041	-0.041	-0.041
Resultado Eva Docente	-0.042	-0.264	0.089
Equivalencia calificación	0.235	0.235	0.235
Numero de instancias	110	265	73

Modelo 2 Tabla Docente

Conjunto de datos: Docente

Campos analizar: No. Eventos Aprobados, No. Eventos Asistidos, Horas Eventos Aprobados, Horas Eventos Asistidos, No. Eventos Nacionales, No. Eventos Internacionales, Horas Actividad Académica, Horas Clase, Resultado Final Evaluación Docente, Nivel Instrucción.

En las tablas 46 y 47 se muestran los resultados de los clústers que genera este modelo para los algoritmos k-means y k-medoids respectivamente:

Tabla 44*: Clústers K-Means tabla docente modelo 2*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel Instrucción	0.105	-0.307	-0.204
No. Eventos Aprobados	-0.362	0.810	4.100
No. Eventos Asistidos	-0.344	0.933	1.677
Horas Eventos Aprobados	-0.291	0.529	4.922
Horas Eventos Asistidos	-0.319	0.905	1.301
No. Eventos Nacionales	-0.394	0.910	4.043
No. Eventos Internacionales	-0.245	0.640	1.524
Horas Actividad Académica	-0.132	0.356	0.659
Horas Clase	0.030	-0.049	-0.584
Resultado Evaluación Docente	-0.107	0.294	0.465
Número de Instancias	332	108	8

Tabla 45*: Clústers K-Medoids tabla docente modelo 2*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel Instrucción	-0.041	-0.041	-0.041
No. Eventos Aprobados	-0.858	-0.981	1.305
No. Eventos Asistidos	-0.704	-0.320	2.375
Horas Eventos Aprobados	-0.636	-0.667	0.331
Horas Eventos Asistidos	-0.683	-0.396	1.725
No. Eventos Nacionales	-0.921	-1.154	1.982
No. Eventos Internacionales	-0.468	0.650	-0.189
Horas Actividad Académica	-0.387	-0.599	0.407
Horas Clase	2.544	-1.365	-0.212
Resultado Evaluación Docente	0.089	-0.264	0.175
Número de Instancias	254	103	91

Modelo 3 Tabla Docente

Conjunto de datos: Docente

Campos analizar: Facultad, Horas Actividad Académica, Horas Clase, Resultado Final Evaluación Docente, Nivel Instrucción.

En las tablas 48 y 49 se muestran los resultados de los clústers que genera este modelo para los algoritmos k-means y k-medoids respectivamente:

Tabla 46: *Clústers K-Means tabla docente modelo 3*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel Instrucción	-0.246	0.239	2.891
Facultad	0.103	-0.670	-0.966
Horas actividad académica	0.062	-0.180	-0.673
Horas clase	0.013	-0.149	-0.098
Resultado final evaluación	0.185	-4.528	-0.327
Número de Instancias	401	14	33

Tabla 47*: Clústers K-Medoids tabla docente modelo 3*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel Instrucción	-0.041	-0.041	-0.041
Facultad	1.179	-1.410	0.316
Horas actividad académica	-0.387	-0.978	-0.559
Horas clase	2.544	-0.149	-1.365
Resultado final evaluación	0.089	-0.042	-0.264
Número de Instancias	103	137	208

Objetivo 3

Realizar clústers de la información de los proyectos de investigación realizados por el personal de la Universidad.

Modelo 1 Tabla Investigación

Conjunto de datos: Investigación

Campos analizar: Nivel Instrucción, Pub. Capitulo libro, Publicaciones Libro, Publicaciones Ponencia, Publicaciones Regional Revista, Publicaciones Producción Científica, Tiene Publicaciones, Total publicaciones

Número máximo de iteraciones: 10

En las tablas 50 y 51 se muestran los resultados de los clústers que genera este modelo para los algoritmos k-means y k-medoids respectivamente:

Tabla 48: *Clústers K-Means tabla investigación modelo 1*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel de instrucción	0.167	-0.361	-0.600
Tiene publicaciones	0.637	-1.519	-1.519
Publicaciones producción Científica	0.266	-0.560	-1.013
Publicaciones Regional Revista	0.465	-1.085	-1.235
Publicaciones Libro	0.216	0.216	-4.134
Publicaciones Capitulo Libro	0.353	-0.749	-1.317
Publicaciones Ponencia	0.519	-1.201	-1.422
Total publicaciones	-0.386	0.638	2.383
Número de instancias	930	327	63

Tabla 49: *Clústers K-Medoids tabla investigación modelo 1*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel de instrucción	-1.240	-0.324	-0.324
Tiene publicaciones	-1.519	-1.519	0.658
Publicaciones producción Científica	0.266	0.266	.0266
Publicaciones Regional Revista	-2.148	-2.148	0.465
Publicaciones Libro	0.216	0.216	0.216
Publicaciones Capitulo Libro	-2.835	0.353	.0353
Publicaciones Ponencia	-1.927	-1.927	0.519
Total publicaciones	0.350	1.828	-0.389
Número de instancias	138	204	978

Modelo 2 Tabla Investigación

Conjunto de datos: Investigación

Campos analizar: Nivel Instrucción, No. Eventos Aprobados, No. Eventos Asistidos, No. Eventos Nacionales, No. Eventos Internacionales, Total publicaciones

En las tablas 51 y 52 se muestran los resultados de los clústers que genera este modelo para los algoritmos k-means y k-medoids respectivamente:

Tabla 50*: Clústers K-Means tabla investigación modelo 2*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel de Instrucción	-0.446	0.143	-0.839
No. Eventos Aprobados	1.317	-0.394	0.657
No. Eventos Asistidos	0.899	-0.268	0.433
No. Eventos Nacionales	1.383	-0.399	-0.211
No. Eventos Internacionales	0.499	-0.207	3.914
Total publicaciones	0.509	-0.241	5.848
Número de Instancias	294	1010	16

Tabla 51: Clústers K-Medoids tabla investigación modelo 2

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel de Instrucción	-0.324	-1.240	-0.324
No. Eventos Aprobados	-0.813	0.894	-0.516
No. Eventos Asistidos	-0.543	0.156	1.088
No. Eventos Nacionales	-0.833	-0.415	-0.137
No. Eventos Internacionales	-0.377	5.321	-0.092
Total publicaciones	-0.389	0.350	-0.389
Número de Instancias	817	43	460

Modelo 3 Tabla Investigación

Conjunto de datos: Investigación

Campos analizar: Nivel de Instrucción, Horas Actividad Académica, Horas Clase, Total Publicaciones.

En las tablas 54 y 55 se muestran los resultados de los clústers que genera este modelo para los algoritmos k-means y k-medoids respectivamente:

Tabla 52*: Clústers K-Means tabla investigación modelo 3*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel de Instrucción	-0.355	-0.688	2.261
Horas de Actividad Académica	-0.114	1.037	-0.173
Horas Clase	0.042	-0.094	-0.141
Total Publicaciones	-0.244	2.190	-0.352
Número de Instancias	983	141	196

Tabla 53*: Clústers K-Medoids tabla investigación modelo 3*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel de Instrucción	-0.324	-0.324	-1.240
Horas de Actividad Académica	-1.315	-0.143	-0.137
Horas Clase	-1.394	1.572	1.687
Total Publicaciones	-0.389	-0.389	0.350
Número de Instancias	482	666	172

ANEXO 10: EVALUAR EL MODELO

Los resultados obtenidos en el análisis para la tabla estudiante cada uno se detalla a continuación:

Análisis de exactitud de los algoritmos en la tabla estudiante

Luego de aplicar el modelo y realizar el análisis del mismo, se puede observar en las tablas 56,57 y 58 que el índice de Davies Bouldin para el algoritmo k-medoids es alto en comparación con el índice del algoritmo k-means por lo que para los modelos 1,2 y 3 de la tabla estudiante se elige el algoritmo k-means.

Tabla 54*: Exactitud modelo 1 tabla estudiante*

Algoritmo	Clúster	Avg. Distancia al punto 0	Avg. Distancia promedio	Davies Bouldin Index
K – Means	1 2 3	3.012 5.323 3.064	3.118	1.027
K – Medoids	1 2 3	3.012 5.323 3.064	3.118	1.045

Tabla 55*: Exactitud modelo 2 tabla estudiante*

Algoritmo	Clúster	Avg. Distancia al punto 0	Avg. Distancia promedio	Davies Bouldin Index
K – Means	1 2	3.10 6.34	4.207	0.369
K – Medoids	1 2	3.10 6.34	6.207	2.650

Tabla 56*: Exactitud modelo 3 tabla estudiante*

Algoritmo	Clúster	Avg. Distancia al punto 0	Avg. Distancia promedio	Davies Bouldin Index
K – Means	1 2	3.188 4.308	3.222	0.701
K – Medoids	1 2	3.188 4.308	3.222	1.347

Análisis de exactitud de los algoritmos en la tabla docente

Luego de aplicar el modelo y realizar el análisis del mismo, se puede observar en las tablas 59, 60 y 61 que el índice de Davies Bouldin para el algoritmo k-medoids es alto en comparación con el índice del algoritmo k-means por lo que para los modelos 1, 2 y 3 de la tabla docente se elige el algoritmo k-means.

Tabla 57*: Exactitud modelo 1 tabla docente*

Algoritmo	Clúster	Avg. Distancia al punto 0	Avg. Distancia promedio	Davies Bouldin Index
K – Means	1	3.627	4.035	0.879
	2	7.825		
	3	9.710		
K – Medoids	1	3.396	4.892	1.586
	2	4.975		
	3	6.845		

Tabla 58*: Exactitud modelo 2 tabla docente*

Algoritmo	Clúster	Avg. Distancia al punto 0	Avg. Distancia promedio	Davies Bouldin Index
K – Means	1	5.730	7.305	1.278
	2	11.01		
	3	22.60		
K – Medoids	1	8.443	10.094	1.524
	2	8.017		
	3	17.05		

Tabla 59: *Exactitud modelo 3 tabla docente*

Algoritmo	Clúster	Avg. Distancia al punto 0	Avg. Distancia promedio	Davies Bouldin Index
K – Means	1	3.281	3.507	1.018
	2	6.644		
	3	4.914		
K – Medoids	1	3.179	4.952	1.543
	2	4.225		
	3	6.309		

Análisis de exactitud de los algoritmos en la tabla investigación

Luego de aplicar el modelo y realizar el análisis del mismo, se puede observar en las tablas 62, 63 y 64 que el índice de Davies Bouldin para el algoritmo k-medoids es alto en comparación con el índice del algoritmo k-means por lo que para los modelos 1,2 y 3 de la tabla investigación se elige el algoritmo k-means.

Tabla 60*: Exactitud modelo 1 tabla investigación*

Algoritmo	Clúster	Avg. Distancia al punto 0	Avg. Distancia promedio	Davies Bouldin Index
K – Means	1	1.163	3.902	1.259
	2	9.199		
	3	16.85		
K – Medoids	1	14.279	5.383	1.525
	2	14.822		
	3	2.159		

Tabla 61*: Exactitud modelo 2 tabla investigación*

Algoritmo	Clúster	Avg. Distancia al punto 0	Avg. Distancia promedio	Davies Bouldin Index
K – Means	1	8.088	3.840	1.215
	2	2.286		
	3	23.912		
K – Medoids	1	2.681	5.508	1.726
	2	23.686		
	3	8.829		

Tabla 62*: Exactitud modelo 3 tabla investigación*

Algoritmo	Clúster	Avg. Distancia al punto 0	Avg. Distancia promedio	Davies Bouldin Index
K – Means	1	1.888	2.387	1.104
	2	5.342		
	3	2.767		
K – Medoids	1	2.966	4.102	2.746
	2	3.499		
	3	9.621		

ANEXO 11: EVALUAR LOS RESULTADOS

Para obtener el número óptimo de clústers se utilizó la herramienta R y la plataforma de RapidMiner para realizar el clustering con el algoritmo K-means que presentó la mayor exactitud, a continuación, se tiene la evaluación de los resultados:

Modelo 2 Tabla Estudiante

Conjunto de datos: Estudiantes

Campos analizados: número de hermanos, número hijos, número integrantes hogar, número dependen de ingresos, promedio, total ingresos, trabaja.

Número máximo de iteraciones: 10

Número de clústers óptimo: 2

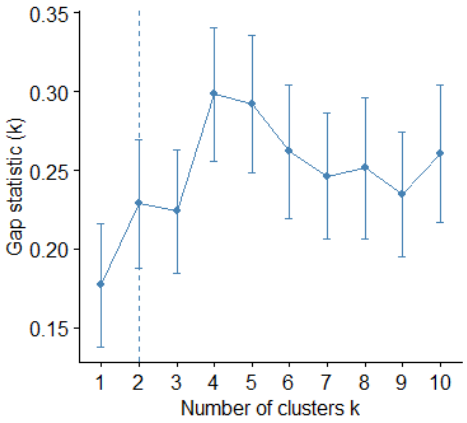

Figura 19. Número óptimo de clústers modelo 2 tabla estudiantes

En la tabla 66 se muestran los resultados de los clústers que genera este modelo.

Tabla 63*: Clústers K-Means tabla estudiantes modelo 2*

Clústers	Clúster 1	Clúster 2
Número integrantes hogar	-0.000	-0.046
Número de hermanos	-0.000	0.810
Número dependen ingresos	0.000	-0.347
Trabaja	0.000	-0.747
Total de ingresos	-0.021	37.259
Número de hijos	-0.000	0.632
Promedio	-0.000	0.355
Avg. Distancia al punto 0	6.103	189.34

Resultado de la gráfica de los centroides de cada clúster.

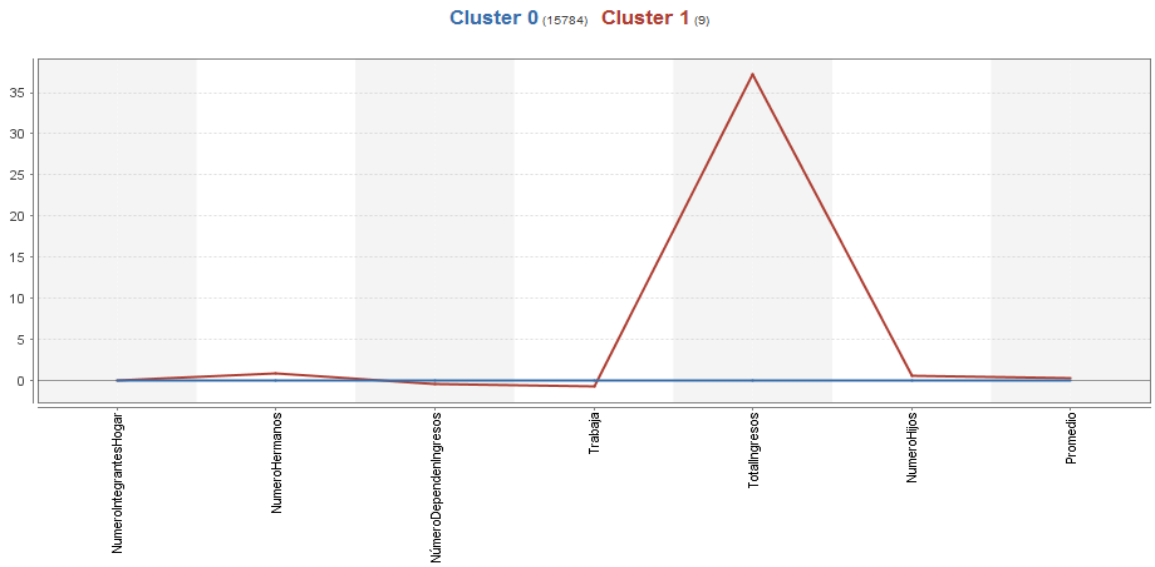

Figura 20. Modelo 2 tabla estudiantes centroides de cada clúster

Análisis de la Gráfica

Como se observa en la gráfica anterior, se tienen diferencias significativas entre el clúster 0 y el clúster 1, estas diferencias se manifiestan mediante las variables del total de ingresos, el número de hijos, el número de hermanos y si trabaja o no. El clúster que concentra al mayor número de estudiantes es el 0. La gráfica además muestra la distancia que existe entre el centroide de un grupo y otro. Si se analiza el promedio a pesar de que existen diferencias significativas entre el clúster 0 y el clúster 1, este no se ve afectado.

En la siguiente figura, se muestran los pesos de las variables que se utilizaron en este análisis.

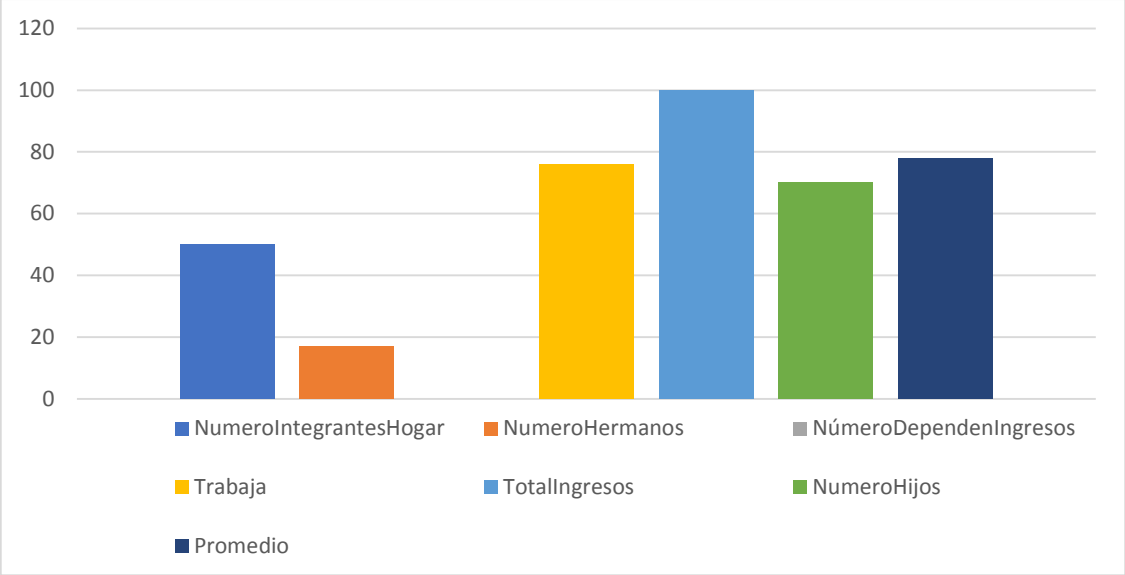

Figura 21. Pesos de las variables modelo 2 tabla estudiantes

Cotejando los pesos de las variables con los resultados del clustering, se ratifica la incidencia que tienen las variables: el número de hijos, si trabaja, el total de ingresos y el promedio general.

Modelo 3 Tabla Estudiante

Conjunto de datos: Estudiantes

Campos analizados: Foráneo, número de hermanos, número de hijos, total ingresos estudiantes, promedio

Número máximo de iteraciones: 10

Número de clústeres óptimo:2

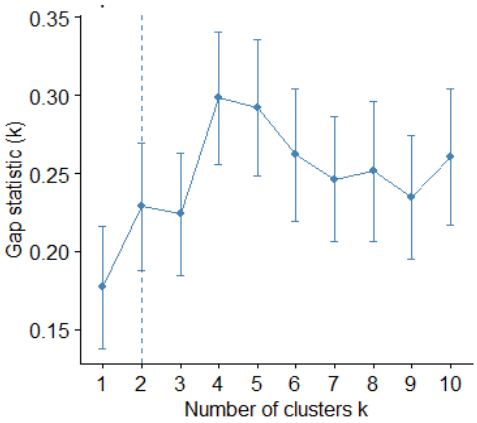

Figura 22. Número óptimo de clústers modelo 3 tabla estudiantes

En las tablas 67 se muestran los resultados de los clústers que genera este modelo.

Tabla 64*: Clústers K-Means tabla estudiantes modelo 3*

Clústers	Clúster 1	Clúster 2
Foráneo	-0.006	0.203
Número de hermanos	0.004	-0.127
Número de hijos	-0.156	4.975
Total ingresos estudiante	0	0
Promedio	-0.008	0.261
Avg. Distancia al punto 0	3.188	4.308

Resultado de la gráfica de los centroides de cada clúster.

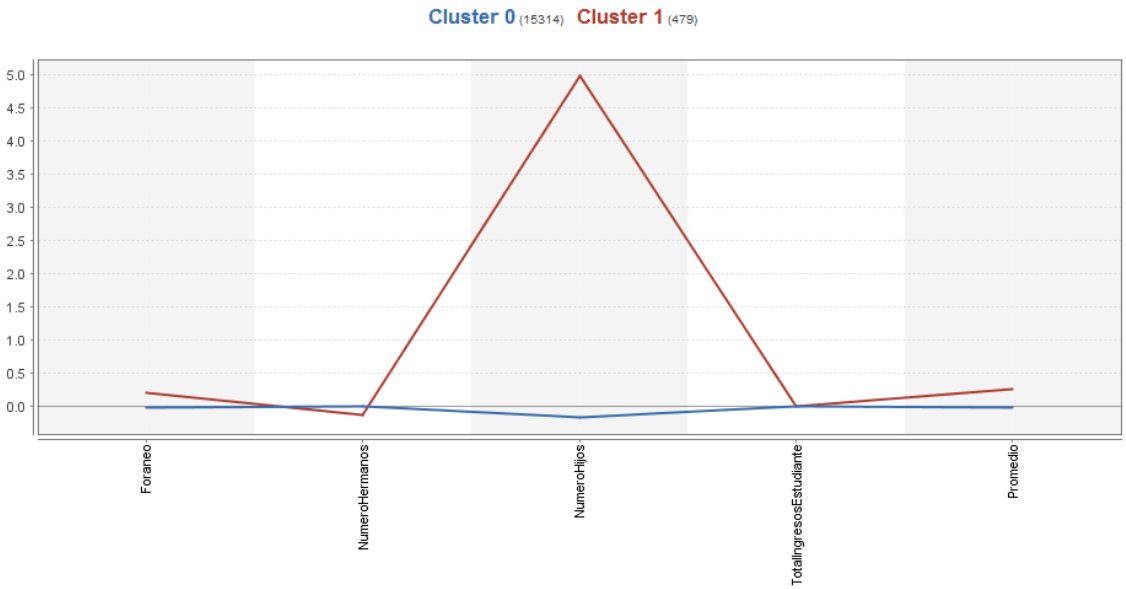

Figura 23 Modelo 3 tabla estudiantes centroides de cada clúster

Análisis de la Gráfica

En la gráfica se puede observar algunos detalles como son:

El clúster 0 concentra el menor número de estudiantes, 479 en total, esta agrupación se caracteriza por concentrar a los estudiantes que son foráneos y que tienen hijos, estos dos criterios muestran diferencias significativas en el promedio, lo que determina que estas dos condiciones influyen en el promedio del estudiante.

En el clúster 1 se concentra el mayor número de estudiantes, 15314 en total y estos se caracterizan por presentar una condición diferente al clúster 0, en el cual los estudiantes no son foráneos, puede ser que no tengan hijos y no poseen ingresos económicos, lo que si genera una diferencia significativa en el promedio obtenido, más aún si este se compara con el promedio de los estudiantes que muestra una mayor responsabilidad como lo es tener hijos.

En la siguiente figura, se muestran los pesos de las variables que se utilizaron en este análisis.

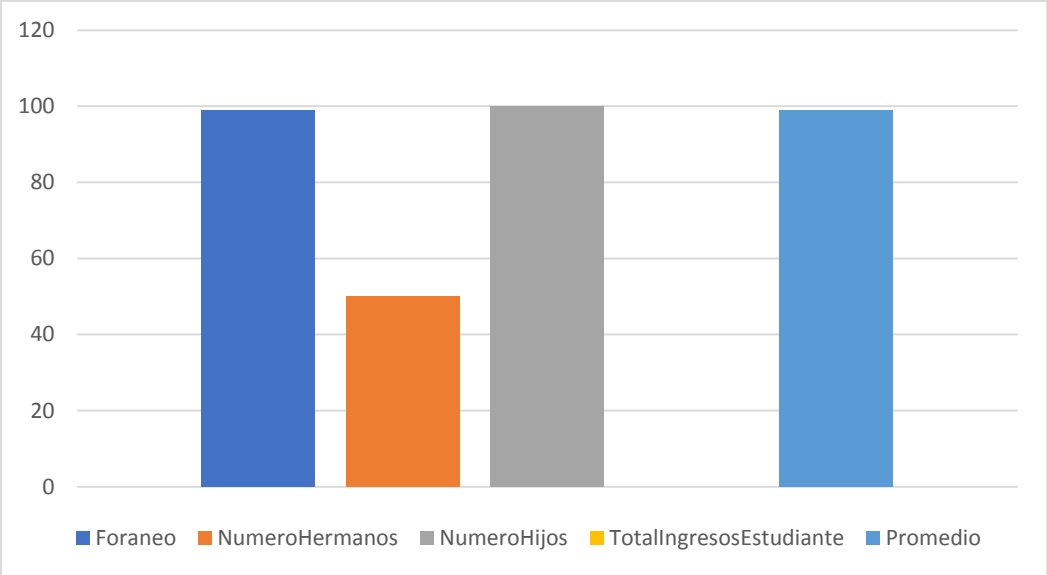

Figura 24. Pesos de las variables Modelo 3 Tabla Estudiantes

Se puede decir que las variables que más incidieron en este análisis son el número de hijos, foráneo y el promedio corroborando así los resultados obtenidos en el clustering.

Modelo 1 Tabla Docente

Conjunto de datos: Docente

Campos analizados: Estado civil, etnia, genero, nivel de instrucción, equivalencia calificación, resultado final evaluación docente.

Número máximo de iteraciones: 10

Número de clústers óptimo: 3

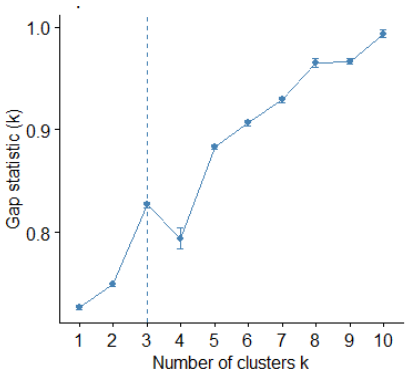

Figura 25. Número óptimo de clústers modelo 1 tabla docentes

En la tabla 68 se muestran los resultados de los clústers que genera este modelo.

Tabla 65*: Clústers K-Means tabla docente modelo 1*

Clústers	Clúster 1	Clúster 2	Clúster 3
Estado civil	0.016	-0.173	-0.200
Genero	0.008	-0.166	-0.026
Etnia	-0.180	4.661	-0.018
Nivel instrucción	-0.031	-0.041	0.716
Resultado Eva Docente	0.191	-0.366	-3.846
Equivalencia calificación	-0.164	0.501	3.146
Número de instancias	413	16	19
Avg. Distancia al punto 0	3.627	7.825	9.710

Resultado de la gráfica de los centroides de cada clúster.

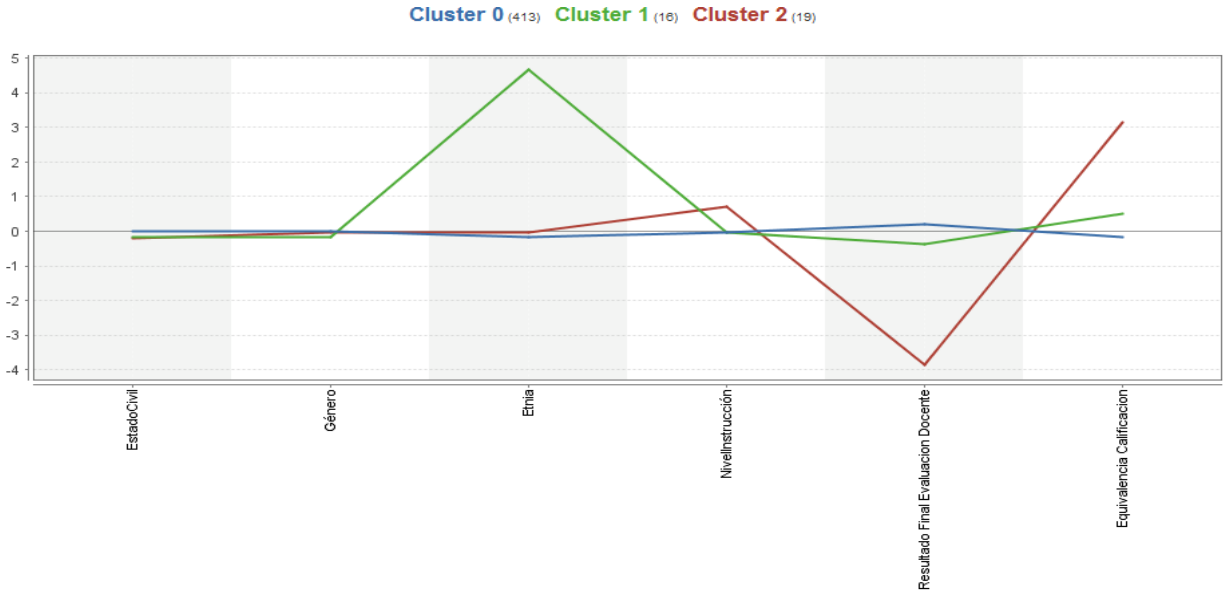

Figura 26. Modelo 1 tabla docente centroides de cada clúster

Análisis de la Gráfica

Del análisis a los resultados se puede obtener, 3 clústers entre los cuales hay diferencias significativas, el clúster 0 concentra el mayor número de docentes 413 en total, el comportamiento de este clúster se muestra equilibrado, el clúster 1 conformado por un total de 16 docentes se diferencia por la etnia la cual incide en el resultado de la evaluación docente y el clúster 3 conformado por 19 docentes se diferencia por el nivel de instrucción. Otros criterios que generan diferencias entre un clúster y otro es el estado civil y el género.

En la siguiente figura, se muestran los pesos de las variables que se utilizaron en este análisis.

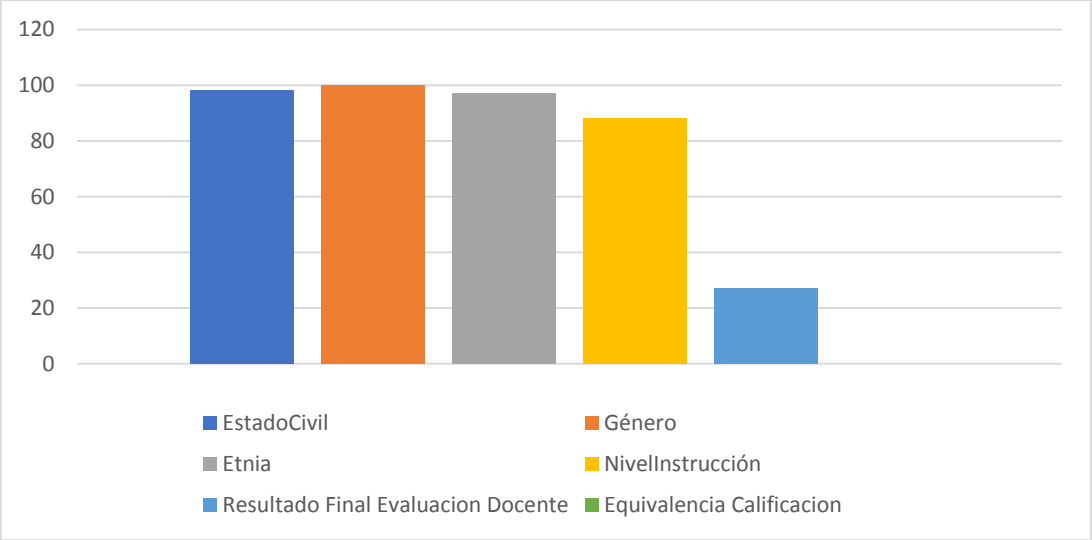

Figura 27. Pesos de las variables modelo 1 tabla docentes

Como se observa en la figura anterior, las variables que más incidieron en este análisis son género, estado civil, etnia y nivel de instrucción.

Modelo 3 Tabla Docente

Conjunto de datos: Docente

Campos analizados: Facultad, Horas Actividad Académica, Horas Clase, Resultado Final Evaluación Docente, Nivel Instrucción.

Número máximo de iteraciones: 10

Número de clústers óptimo:3

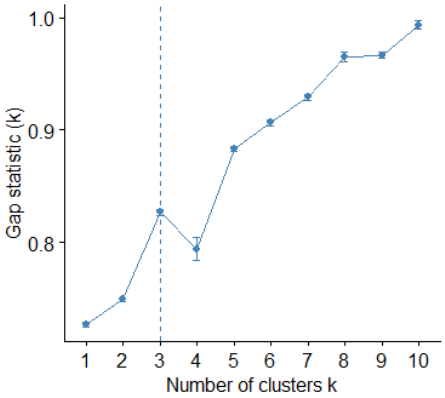

Figura 28. Número óptimo de clústers modelo 3 tabla docentes

En la tabla 70 se muestran los resultados de los clústers que se generaron en este modelo.

Tabla 66*: Clústers K-Means tabla docente modelo 3*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel Instrucción	-0.246	0.239	2.891
Facultad	0.103	-0.670	-0.966
Horas actividad académica	0.062	-0.180	-0.673
Horas clase	0.013	-0.149	-0.098
Resultado final evaluación	0.185	-4.528	-0.327
Número de Instancias	401	14	33
Avg. Distancia al punto 0	3.281	6.644	4.914

Resultado de la gráfica de los centroides de cada clúster.

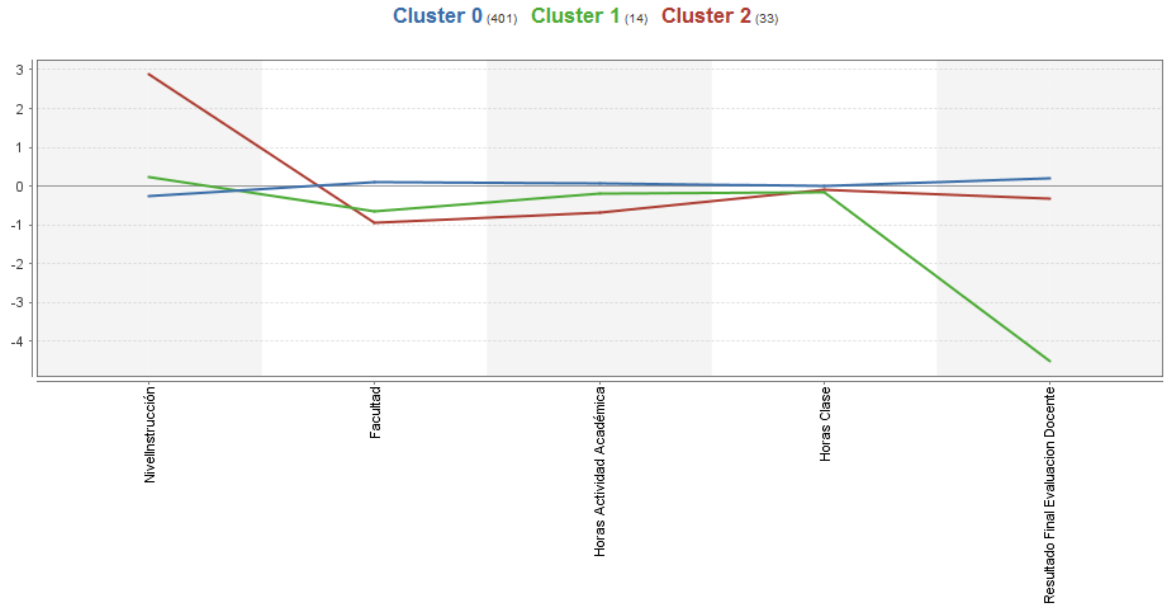

Figura 29. Modelo 3 tabla docente centroides de cada clúster

Análisis de la Gráfica

Del análisis a los resultados se puede obtener, 3 clústers entre los cuales hay diferencias significativas, las cuales se encuentran marcadas por el nivel de instrucción, la facultad, las horas de actividad académica y el resultado final de la evaluación docente, teniendo así en el clústers 0 docentes con el nivel de instrucción más alto (PhD) y el resultado final de la evaluación docente más alta con respecto a los clústers 1 y 2.

En la siguiente figura se muestran los pesos de las variables que se utilizaron en este análisis.

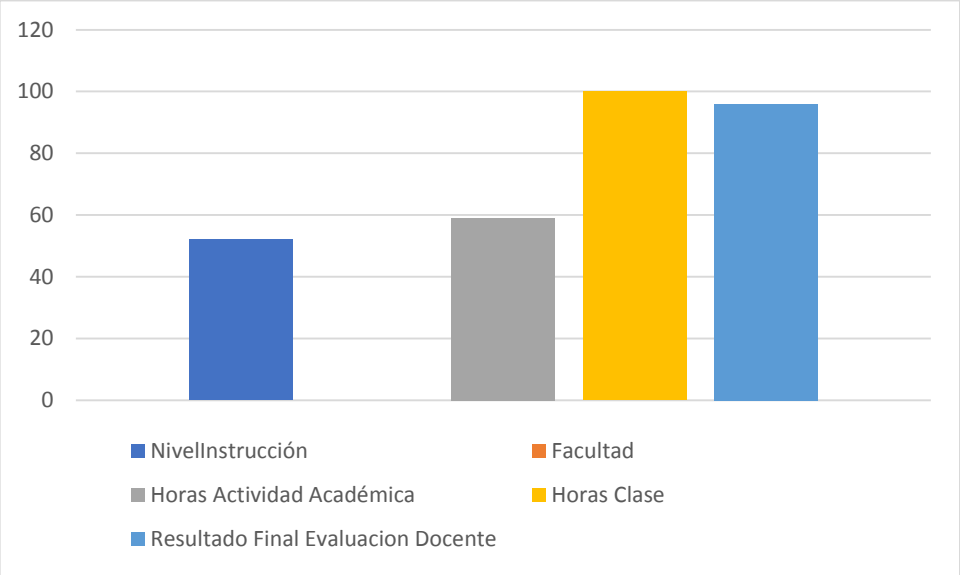

Figura 30 Pesos de las variables modelo 3 tabla docentes

Como se observa en la figura anterior, las variables que inciden en los resultados del clustering son: horas clase, horas actividad académica y Resultado final evaluación docente.

Modelo 1 Tabla Investigación

Conjunto de datos: Investigación

Campos analizados: Nivel Instrucción, Pub. Capitulo libro, Publicaciones Libro, Publicaciones Ponencia, Publicaciones Regional Revista, Publicaciones Producción Científica, Tiene Publicaciones, Total publicaciones

Número máximo de iteraciones: 10

Número de clústers óptimo:3

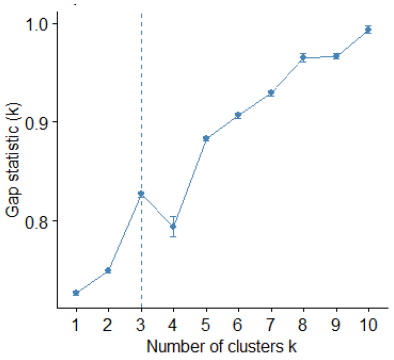

Figura 31. Número óptimo de clústers modelo 1 tabla investigación

En la tabla 71 se muestran los resultados de los clústers que genera este modelo.

Tabla 67*: Clústers K-Means tabla investigación modelo 1*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel de instrucción	0.167	-0.361	-0.600
Tiene publicaciones	0.637	-1.519	-1.519
Publicaciones producción Científica	0.266	-0.560	-1.013
Publicaciones Regional Revista	0.465	-1.085	-1.235
Publicaciones Libro	0.216	0.216	-4.134
Publicaciones Capitulo Libro	0.353	-0.749	-1.317
Publicaciones Ponencia	0.519	-1.201	-1.422
Total publicaciones	-0.386	0.638	2.383
Número de instancias	930	327	63
Avg. Distancia al punto 0	1.163	9.199	16.85

Resultado de la gráfica de los centroides de cada clúster.

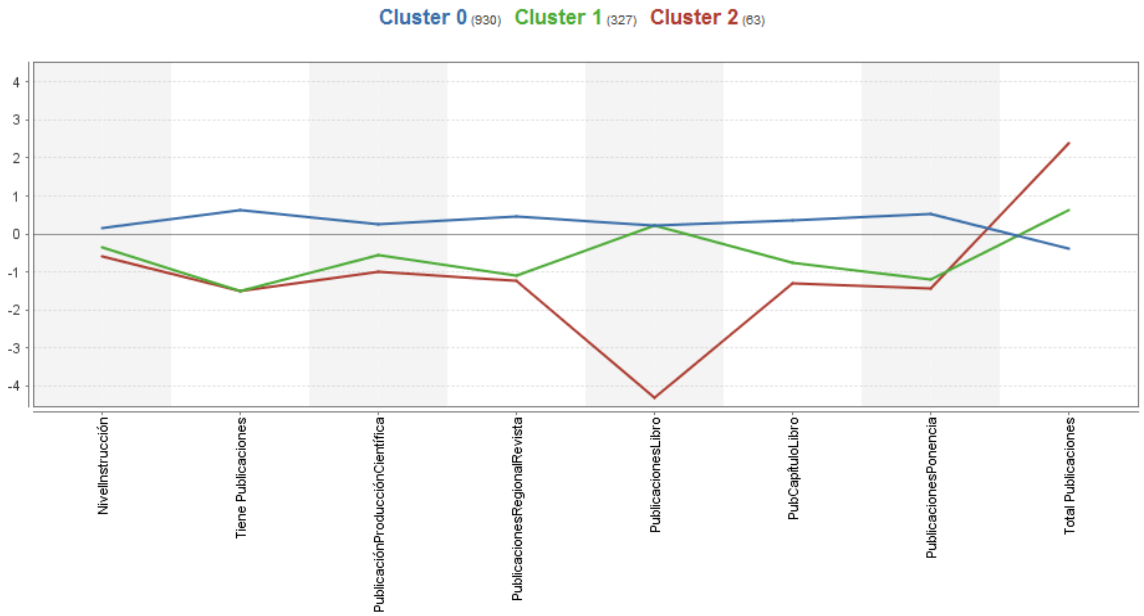

Figura 32 Modelo 1 tabla investigación centroides de cada clúster

Análisis de la Gráfica

Del análisis a los resultados se diferencian 3 clústeres. El clúster 0 concentra al mayor número de docentes 930 en total y se caracteriza por agrupar a docentes que han realizado ponencias, publicaciones regionales, libros y capítulos de libro. El clúster 1 concentra a 327 docentes y se caracteriza por agrupar a docentes con escasa producción científica.

En la siguiente figura se muestran los pesos de las variables que se utilizaron en este análisis.

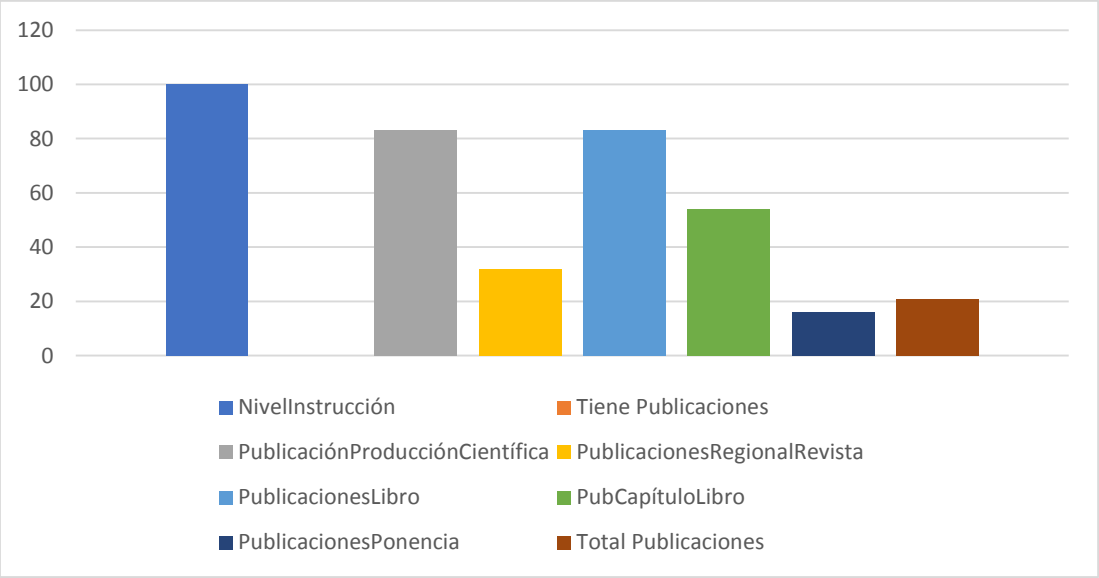

Figura 33. Pesos de las variables modelo 1 tabla investigación

Como se observa en la figura anterior, las variables que más incidieron en este análisis son nivel de instrucción, publicaciones producción científica, publicaciones capítulo de libro y publicaciones libro corroborando así los resultados obtenidos en el clustering.

Modelo 3 Tabla Investigación

Conjunto de datos: Investigación

Campos analizadas: Nivel de Instrucción, Horas Actividad Académica, Horas Clase, Total Publicaciones.

Número máximo de iteraciones: 10

Número de clústers óptimo: 3

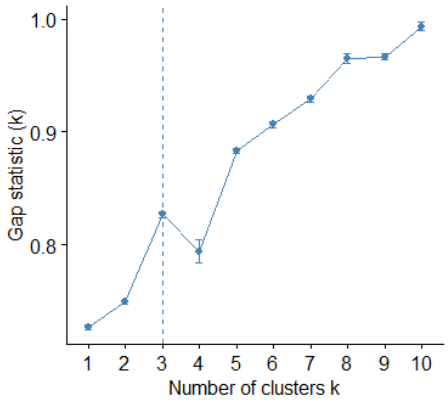

Figura 34 Número óptimo de clústers modelo 3 tabla investigación

En las tablas 73 se muestran los resultados de los clústers que genera este modelo para los algoritmos k-means y k-medoids respectivamente:

Tabla 68 *: Clústers K-Means tabla investigación modelo 3*

Clústers	Clúster 1	Clúster 2	Clúster 3
Nivel de Instrucción	-0.355	-0.688	2.261
Horas de Actividad Académica	-0.114	1.037	-0.173
Horas Clase	0.042	-0.094	-0.141
Total Publicaciones	-0.244	2.190	-0.352
Número de Instancias	983	141	196
Avg. Distancia al punto 0	1.888	5.342	2.767

Resultado de la gráfica de los centroides de cada clúster.

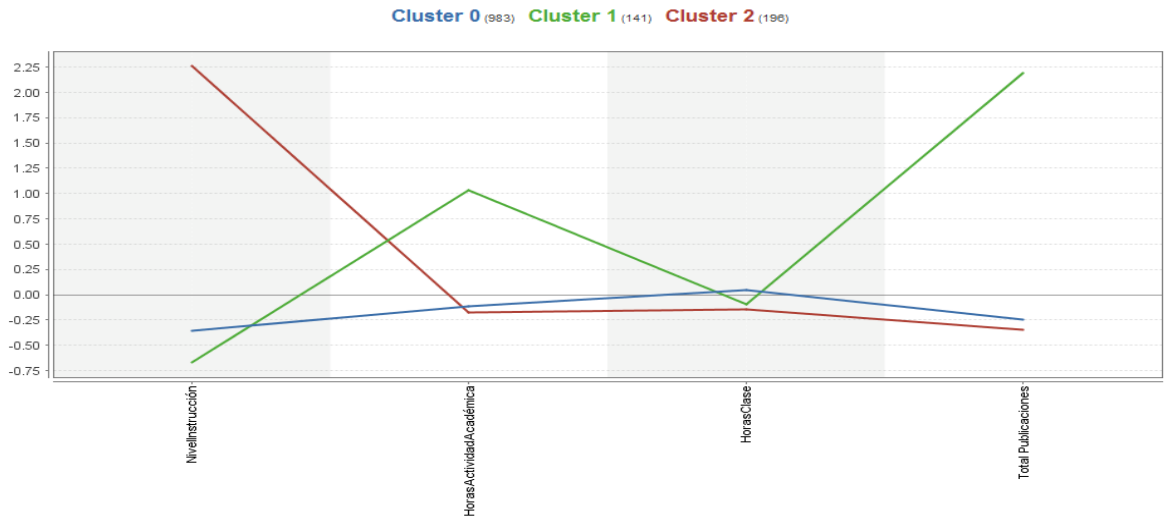

Figura 35 Modelo 3 tabla investigación centroides de cada clúster

Análisis de la Gráfica

Del análisis a los resultados, se tiene 3 clústeres. El clúster 0 concentra a 983 docentes, los cuales se caracterizan por concentrar un equilibrado comportamiento, sin embargo, no representa una elevada actividad académica, el clúster 1 conformado por 141 docentes se caracteriza por concentrar publicaciones y una elevada actividad académica y el clúster 3 conformado por 194 docentes los cuales se destacan por presentar un diferente nivel académico y presentar una reducida producción académica.

Esto determina que el nivel académico incide en el número de publicaciones y no necesariamente el nivel académico más elevado garantiza el número de publicaciones.

En la siguiente figura, se muestran los pesos de las variables que se utilizaron en este análisis.

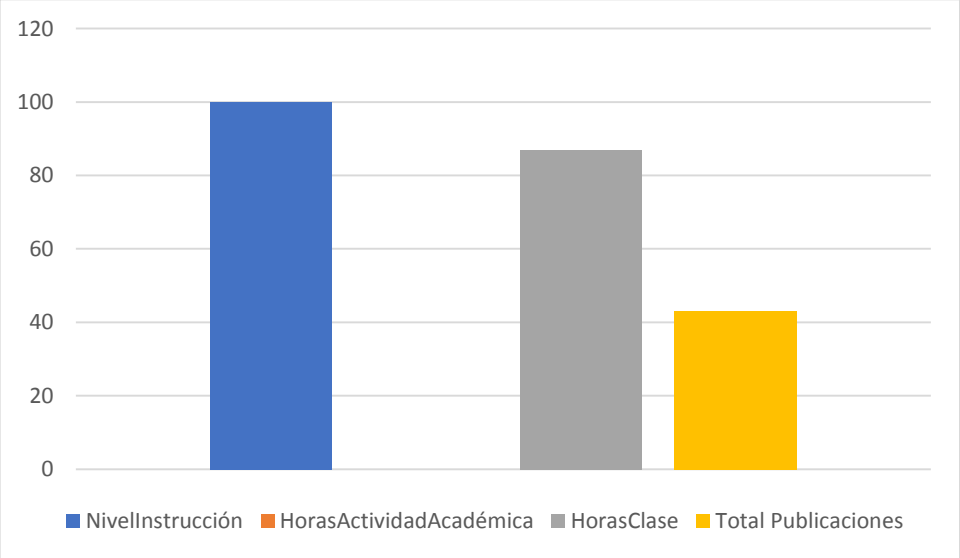

Figura 36 Pesos de las variables modelo 3 tabla investigación

De análisis a la figura anterior se tiene que las variables que inciden y generan diferencias significativas en el resultado del clustering son: Nivel de Instrucción y Horas Clase.