UNIVERSIDAD NACIONAL DE CHIMBORAZO

FACULTAD DE INGENIERÍA

CARRERA DE SISTEMAS Y COMPUTACIÓN

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

TRABAJO DE TITULACIÓN

ANÁLISIS CLÚSTER RELACIONADO CON EL ACCESO A LA EDUCACIÓN VIRTUAL POR PARTE DE LOS PROFESORES Y ESTUDIANTES DE LA UNACH.

AUTOR:
Rosa Virginia Tapia Estrada

TUTOR:
Ing. Lida Barba, PhD.

Riobamba - Ecuador:

Año 2021

PÁGINA DE ACEPTACIÓN

Los miembros del tribunal de Graduación del proyecto de investigación de título: **"ANÁLISIS CLÚSTER RELACIONADO CON EL ACCESO A LA EDUCACIÓN VIRTUAL POR PARTE DE LOS PROFESORES Y ESTUDIANTES DE LA UNACH."**, presentado por la estudiante Srta. Rosa Virginia Tapia Estrada, dirigido por la PhD. Lida Mercedes Barba Maggi.

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

Para constancia de lo expuesto firman:

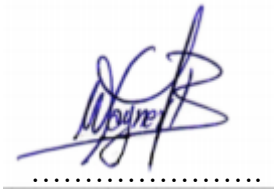
PhD. Lida Barba
Tutora del Proyecto

…….
Firma

XIMENA ALEXANDRA QUINTANA LOPEZ
Firmado digitalmente por XIMENA ALEXANDRA QUINTANA LOPEZ
Fecha: 2021.04.06 12:58:21 -05'00'

PhD. Ximena Quintana
Miembro del Tribunal

…….
Firma



Mag. Wayner Bustamante
Miembro del Tribunal

…….
Firma

DERECHOS DE AUTORÍA

La responsabilidad del contenido de este proyecto de graduación corresponde exclusivamente a: Rosa Virginia Tapia Estrada bajo la dirección de la PhD. Lida Barba, y al patrimonio intelectual de la Universidad Nacional de Chimborazo.

Autora

……..
Rosa Virginia Tapia Estrada
2300409394

Directora del Proyecto

…….
PhD. Lida Barba
0602582132

DEDICATORIA

Dedico este proyecto de investigación a toda mi familia quienes siempre estuvieron a mi lado apoyándome en los buenos y malos momentos de mi vida, a mis padres Abrahán Tapia y Sara Estrada que han sido mi apoyo incondicional para poder culminar mis estudios, a mi esposo e hija que siempre me brindaron su apoyo y consejos para poder culminar esta etapa importante de mi vida.

A mis hermanos Yolanda, Fabiola, Marco, Alex y Karelis; por brindarme su cariño, amor y ser ese soporte fundamental en esos momentos difíciles.

Rosa Virginia Tapia Estrada

"*El éxito es la suma de pequeños esfuerzos que se hacen día tras día****" Robert Collier.***

AGRADECIMIENTO

Agradezco infinitamente a Dios por protegerme y guiar mi camino en todos estos años de vida, por brindarme salud y darme las fuerzas necesarias para lograr culminar mi carrera universitaria.

Agradezco a mis padres, hermanos, suegros, esposo e hija por el apoyo incondicional, por sus consejos, por siempre estar presentes en cada paso de mi vida.

Mi gratitud entera para la Universidad Nacional de Chimborazo por abrirme las puertas y darme la oportunidad de ser una profesional, a la carrera de Ingeniería en Sistemas y Computación, profesores, compañeros y amigos de clases por compartir sus conocimientos y experiencias. En especial a la Ing. Lida Barba, Ph.D tutora de tesis quien me brindó su apoyo incondicional, de igual forma a mis tutores colaboradores, PhD Ximena Quintana y MsC. Wayner Bustamante.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

RESUMEN

La Universidad Nacional de chimborazo (UNACH) dispone de varios sistemas informáticas uno de ellos es el Sistema de Control Académico (SICOA) a través del cual fueron recopilados los datos de dos encuestas correspondientes a la accesibilidad a la educación virtual en el año 2019, modalidad adoptada por el confinamiento obligatorio efecto de la pandemia Covid-19.

El trabajo consistió en el análisis clúster enfocado a localizar los grupos homogéneos en su condición de accesibilidad a la educación virtual. Con la finalidad de evaluar la efectividad de la segmentación, se comparan los algoritmos de aprendizaje no supervisado K-means y K-medoids. El proceso de minería se llevó a cabo por medio de la metodología CRISP-DM, misma que permitió analiza, limpiar y construir los datos.

Los resultados presentan 3 grupos homogéneos de estudiantes y 2 grupos homogéneos de docentes. Las métricas distancia de clúster al centroide y Davies Bouldin determinaron que la mayor exactitud fue alcanzada por medio de K-means.

Palabras Clave: Clustering, K-Means, K- Medoids, Accesibilidad, Educación.

Abstract

The National University of Chimborazo (UNACH) works with several informatics systems. One of them is the Academic Control System (SICOA). This study used the SICOA to collect data from two surveys regarding the accessibility of virtual education in 2019 considering that schools adopted virtual education because of the obligated lockdown during the Covid -19 pandemic.

This work consisted of a cluster analysis focused on identifying homogenous groups in their conditions of accessibility to virtual education with the aim of assessing segmentation effectiveness. We compared not supervised K- means and K-medoids learning algorithms. The mining process was conducted through the CRISPP-DM methodology. This methodology allowed data analysis, cleaning, and building.

The results suggested three homogenous groups of students and two homogeneous groups of teachers. The metric distance between cluster to centroid and Davies Bouldin determined that the highest accuracy was reached through K-means.

Keywords: Clustering, K-Means, K- Medoids, Accessibility, Education.

Reviewed by: MsC. Adriana Cundar, Ph.D.

ENGLISH PROFESSOR

c.c. 1709268534

INTRODUCCIÓN

El incremento e importancia de lo que hoy conocemos como educación virtual o teleformación se ve reforzado con la creación casi generalizada de los campus virtuales y aulas virtuales de aprendizaje (González & Urbina, 2014). La pandemia COVID-19 ocasionó un cambio de vida y la educación no fue la excepción, en todos los niveles, los procesos de enseñanza – aprendizaje que se impartían en la modalidad presencial han pasado a la modalidad en línea, a sistemas de difusión por medios electrónicos o incluso se han suspendido por un tiempo. Ahora se maneja los sistemas en línea donde el alumno toma clases en vivo (síncronas) y pregrabadas (asíncronas) usualmente por medio de internet (Ruiz & Sánchez, 2020).

En este contexto la inteligencia artificial ofrece un gran potencial de aplicaciones en combinación con tecnologías emergentes o ya existentes, como es el caso del análisis de big data y la optimización de sistemas (Rodríguez & Castillo, 2017). Un resultado positivo de esto es que la IA facilita la vida en ciertas áreas en las que se necesita analizar datos (Adamssen, 2020).

Las técnicas de minería de datos se clasifican en dos categorías: algoritmos supervisados o predictivos y algoritmos no supervisados o de descubrimiento del conocimiento (Tuya, Ramos & Dolado, 2007). El aprendizaje supervisado necesita ser entrenado mediante pares de entradas y salidas (Pino, Gómez & Martínez, 2007), mientras que el aprendizaje no supervisado se ajusta a las observaciones.

La agrupación facilita la descripción concisa de un conglomerado de datos multidimensional complejo, lo cual se encuentra sustituyendo la descripción de todos los elementos de un conglomerado por la de un representante característico del mismo. En algunos contextos, como el de la minería de datos, la agrupación es un método de estudio no supervisado puesto que explora encontrar relaciones entre variables descriptivas pero no la que guardan con respecto a una variante objetivo.

El número de grupos o clústers no es conocido de antemano, por tanto los grupos se crean en función de la naturaleza de los datos, se trata de una técnica de clasificación post hoc (Pérez & Santín, 2008).

Con los antecedentes presentados en el trabajo de investigación se pretende identificar grupos homogéneos entre estudiantes y profesores de la UNACH con accesibilidad a la educación virtual.

Los conglomerados resultantes, deberán mostrar un alto grado de homogeneidad interna y un alto grado de heterogeneidad externa entre conglomerado (Pedroza, 2007).

Las instituciones educativas permanentemente requieren tomar las decisiones más acertadas para mejorar su gestión, una oportunidad efectiva se encuentra en los datos, más aún ante un cambio obligatorio de modalidad de estudio.

CAPÍTULO I

Planteamiento del problema y justificación

La actual crisis sanitaria global producto del Covid-19, ha ocasionado que las actividades de las organizaciones se desarrollen en condiciones y escenarios diferentes. En este contexto, el sistema de educación superior no es la excepción, las instituciones de gobierno, de manera específica el Ministerio del Trabajo y el Consejo de Educación Superior, CES, ha promulgado leyes y normativas que direccionan la aplicación del teletrabajo y la educación virtual, como mecanismo de aislamiento entre las personas para evitar posibles contagios y que al mismo tiempo busca no paralizar los servicios. La Universidad Nacional de Chimborazo, en el mes de junio/2020 adoptó la ejecución del periodo académico mayo-octubre 2020 en modalidad virtual, basados en la información nacional proporcionada por las instancias legales tales como el comité de operaciones especiales (COE) nacional y COE local, así como de la información que fue recabada por medio del Sistema de Control Académico SICOA, misma que buscaba identificar la accesibilidad a la educación virtual por parte de los profesores y los estudiantes. Los datos que fueron recabados han sido analizados con técnicas estadísticas convencionales, mismas que no permiten extraer conocimiento no obvio y significante, que conllevaría a mejorar la toma de decisiones. Mediante la técnica de minería de datos Clustering también conocida como Agrupamiento, se pretende segmentar los datos en grupos homogéneos, en este caso, que presenten determinadas variables o características que les permita acceder a la educación virtual, en escalas cualitativas y cuantitativas. La importancia en la identificación de grupos homogéneos radica en que la institución podrá realizar el seguimiento de los mismos, apoyando a aquellos cuyas características los haga más susceptibles a no lograr consolidar su proceso académico por medio de la permanencia y ejecución exitosa de la educación virtual. Aunque la educación virtual pudiese ser temporal, el presente estudio permitirá contar con un buen referente de conocimiento válido y aplicable en procesos futuros de educación virtual que tengan regularidad, como por ejemplo la ejecución de carreras o programas de estudios en esa modalidad.

1.2. OBJETIVOS

1.2.1. Objetivo General

- Identificar grupos homogéneos entre estudiantes y profesores de la UNACH con accesibilidad a la educación virtual, por medio de análisis clúster haciendo uso de herramientas de inteligencia artificial.

1.2.2. Objetivos Específicos

- Preparar los datos obtenidos de la encuesta de accesibilidad virtual de profesores y estudiantes que serán utilizados para el análisis clustering.
- Seleccionar las metodologías, técnicas y herramientas de inteligencia artificial en aprendizaje no supervisado para el análisis de agrupamiento.
- Diseñar, implementar y evaluar el modelo de clustering para comparar su rendimiento e identificar grupos homogéneos de estudiantes y profesores a partir de los datos analizados.

CAPÍTULO II

Fundamentación teórica

2.1. Minería de datos

Es el proceso de descubrir automáticamente información valiosa en grandes volúmenes de datos, con el fin de extraer reglas, patrones, asociaciones, relaciones 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).

Según 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

Abarca una gran cantidad de escenarios, entre ellos el de la educación (Rosado & Verjel, 2016). 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 (Peña, 2014).

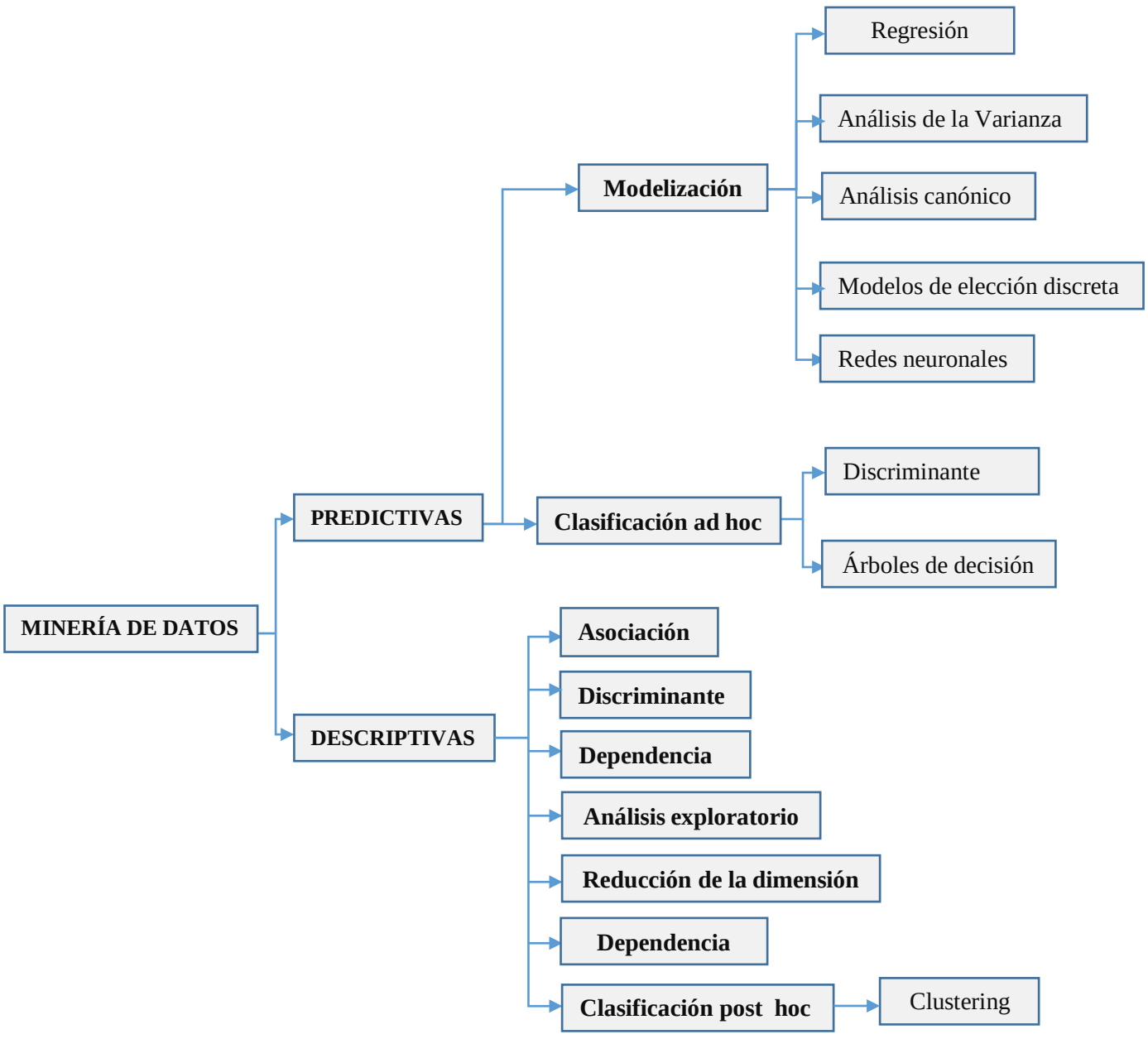
Algunas de las tareas sustanciales de la minería de datos son la identificación de aplicaciones para las técnicas existentes, y desarrollar nuevas técnicas para dominios tradicionales o de nueva aplicación, como el comercio electrónico y la bioinformática (Riquelme, Ruiz, & Gilbert, 2006).

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. Técnicas de la minería de datos

De acuerdo con Pérez & Santín, (2008), Las técnicas de minería de datos propiamente dichas pueden clasificarse según el esquema siguiente:

Figura 1: *clasificación de algoritmos*

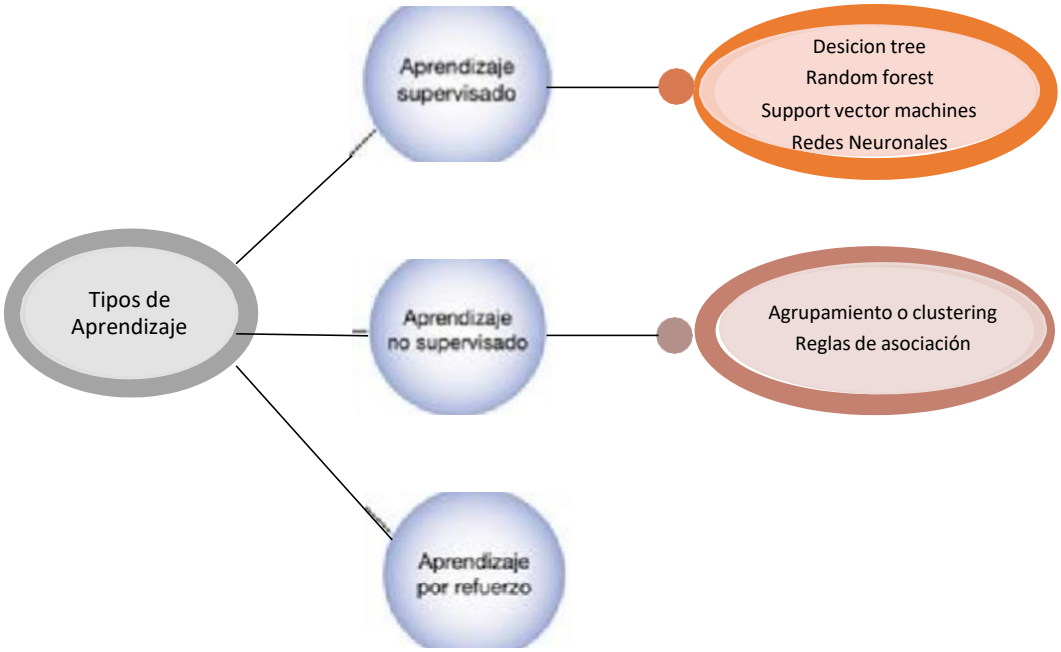


2.4. La inteligencia artificial (IA)

Es un área extraordinariamente vital, amplia y multidisciplinar. El comportamiento inteligente humano, como el que la inteligencia Artificial trata de emular y/o simular, presenta complejos *cognitivos, perceptivos, heurísticos, sociales, colaborativos*, *etc*. Entre todos ellos, se encuentra una amplia variedad de orientaciones tomando en cuenta aspectos del pensamiento humano. De acuerdo con Benítez, (2014), las aplicaciones más frecuentes de la IA incluyen campos como la robótica, el análisis de imágenes o el tratamiento automático de textos.

Los algoritmos de IA de aprendizaje supervisado y no supervisado son los siguientes:

Figura 2: *Tipos de aprendizaje de algoritmos*



Fuente: *Beunza, 2020*

Aprendizaje supervisado

Según Beunza, (2020), el aprendizaje supervisado se refiere a un tipo de modelos de machine learning que se entrenan con un conjunto de ejemplos en los que los resultados de salida (outcome, eventos o labels) son conocidos. Los modelos aprenden de esos resultados conocidos y realizan ajustes es sus parámetros interiores para adaptarse a los datos de entrada. Cuenta con los algoritmos de:

Regresión: tiene como resultado un número específico. Si las etiquetas suelen ser un valor numérico, mediante las variables de las características, se pueden obtener dígitos como dato resultante.

Aprendizaje no supervisado

El objetivo es la extracción de la información significativa, sin la referencia de variables de salida conocidas y mediante la exploración de la estructura de los datos. Estos parten de datos no etiquetados en los que no existe información de clasificación ni de evento dependiente continuo. Su labor es encontrar patrones de agrupamiento (clusters) para sujetos que por sus características en todas las variables introducidas en el modelo son parecidos. Una categoría de algoritmo no supervisado es el clustering o agrupamiento.

Aprendizaje por refuerzo

Funciona mediante intervención humana durante el proceso de aprendizaje. Básicamente consiste en premiar o penalizar las decisiones parciales que va tomado el algoritmo en cada paso del aprendizaje (Beunza, Puertas, & Moreno, 2020).

2.5. Algoritmos de clustering

Buscan la generación de nuevos conjuntos a partir de datos analizados (Suárez, 2014). A estos algoritmos también se le conoce como clasificación no supervisada (Garre, & Cuadrado, & Sicilia, 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).

Clustering 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).

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

Algoritmos jerárquicos: Estos algoritmos de clustering crean una descomposición jerárquica formando un árbol o dendograma que secciona la esencia de los datos recursivamente en conglomerados 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).

Los algoritmos de clustering buscan formar subconjuntos más pequeños de objetos hasta que cada uno de los objetos este en un clúster individual (Arroyo, & Borja, 2018).

2.5.1. Dominios de aplicación

De acuerdo con (Chapman & Hall, 2014), algunos dominios de aplicación comunes en los que surge el problema de agrupamiento son los siguientes:

- **Paso intermedio para otros problemas fundamentales de minería de datos:** Dado que la agrupación en clúster puede considerarse una forma de resumen de datos, a menudo sirve como un paso intermedio clave para muchos problemas fundamentales de minería de datos, como la clasificación o el análisis de valores atípicos.
- **Filtración colaborativa:** en los métodos de filtrado colaborativo, el agrupamiento proporciona un resumen de usuarios con ideas afines.
- **Segmentación de clientes:** esta aplicación es bastante similar al filtrado colaborativo, ya que crea grupos de clientes similares en los datos.
- **Resumen de datos:** Muchos métodos de agrupamiento están estrechamente relacionados con los métodos de reducción de dimensionalidad, el resumen de datos puede ser útil para crear representaciones de datos compactas, que son más fáciles de procesar e interpretar en una amplia variedad de aplicaciones.
- **Detección dinámica de tendencias:** Se pueden utilizar muchas formas de algoritmos dinámicos y de transmisión para realizar la detección de tendencias en una amplia variedad de aplicaciones de redes sociales.
- **Análisis de datos multimedia:** Una variedad de diferentes tipos de documentos, como imágenes, audio o video, entran en la categoría general de datos multimedia. En muchos casos. Los datos pueden ser multimodales y pueden contener diferentes tipos, en tales casos, el problema se vuelve aún más desafiante.
- **Análisis de datos biológicos:** Los datos biológicos se han vuelto omnipresentes en los últimos años, debido al éxito del esfuerzo del genoma humano y la creciente capacidad de recopilar diferentes tipos de datos de expresión génica. Los datos biológicos generalmente se estructuran

como secuencias o como redes. Los algoritmos de agrupamiento proporcionan buenas ideas de los elementos clave en los datos, así como secuencias inusuales.

- **Análisis de redes sociales:** En estas aplicaciones, la estructura de una red social se utiliza para determinar las comunidades importantes en la red subyacente.

2.5.2. Área de agrupación de datos

Según Chapman & Hall, (2014), el trabajo en el área de agrupación de datos generalmente se divide en varias categorías amplias:

- **Centrado en la técnica:** dado que la agrupación en clústeres es un problema bastante popular, no es sorprendente que numerosos métodos, como técnicas probabilísticas, técnicas basadas en la distancia, técnicas espectrales, técnicas basadas en la densidad y técnicas basadas en la reducción de la dimensionalidad, se utilicen para la agrupación.
- **Información adicional a partir de variaciones de agrupamiento:** también se han diseñado varios conocimientos para diferentes tipos de variaciones de agrupación. por ejemplo, el análisis visual, el análisis supervisado, el análisis de conjuntos o el análisis de múltiples vistas se pueden utilizar para obtener información adicional.
- **Centrado en el tipo de datos:** diferentes aplicaciones crean diferentes tipos de tipos de datos con diferentes propiedades. por ejemplo, una máquina de ECG producirá puntos de datos de series de tiempo que están altamente correlacionados entre sí.

2.6. Agrupamiento K-means

El algoritmo de conglomerado de K-medias se basa en el análisis de grupos, divide los datos encontrados en bloques, separados y agrupados por características semejantes (Gutiérrez & Molina, 2016).

El algoritmo K-Means reúne las observaciones en K grupos diferentes, donde el analista determina el número de clústers (K), halla los K mejores clústers, entendiendo como mejor clúster aquel cuya varianza interna (intra-cluster variation) sea lo más pequeña posible (Amat, 2017).

2.7. Exactitud en los algoritmos de clustering

Según 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 componente de cada clúster debe ser lo más próximo posible a los otros componentes del mismo clúster.
- **Separación:** Los clústers deben tener una división considerable entre ellos.

2.8. Funcionamiento de K_means

Según Amat, (2017), k-means es un algoritmo que consta de cuatro pasos:

- Fijar el número K de clústers que deseo crear.
- Escoger de forma aleatoria k observaciones del set de datos como centroides iniciales.
- Designar cada una de las observaciones al centroide más próximo.
- Para cada uno de los K clústers recalcular su centroide.

$$\text{cj} = \frac{1}{\lfloor Cj \rfloor} \sum_{\forall \times \in \boldsymbol{Cj}} \boldsymbol{z}$$

Donde $\boldsymbol{z}$ representa un elemento del conjunto de datos, que pertenece al cluster Cj, cj es un centroide y $\lfloor Cj \rfloor$ corresponde al número de elementos en el cluster Cj.

Continuar con los pasos 3 y 4 hasta que sus asignaciones no cambien o pueda alcanzar el número máximo de iteraciones estipulado.

Una desventaja de este algoritmo es que el resultado obtenido es dependiente de la selección inicial de los centroides de los clústers y puede converger a óptimos locales. Por lo tanto, la selección de los centroides iniciales afecta al proceso principal de k-means y a la participación resultante de este proceso.

De acuerdo a Amat (2017) k-means es un algoritmo que presenta ciertas ventajas y desventajas que sedescriben en la tabla 1:

Tabla 1: *Ventajas y desventajas del algoritmo K-means*

Ventajas	Desventajas
K-means es uno de los métodos de clustering más utilizados.	Necesita que el beneficiario defina el número de clústers
Fácil de entender, fácil de adaptar	Las agrupaciones resultantes pueden variar dependiendo de la asignación aleatoria inicial de los centroides
K-means es el más utilizado en la comunidad de aprendizaje no supervisado	No es conveniente para grandes conjuntos de datos.

Fuente: *Amat, 2017*

CAPÍTULO III

Metodología

La metodología utilizada para el desarrollo del presente trabajo fue de tipo cualitativa y cuantitativa, porque en la investigación se utilizaron herramientas informáticas, cálculos matemáticos y cálculos estadísticos, como también ver el comportamiento de cada clúster para poder analizar los datos y obtener resultados.

3.1. Tipo de investigación

Con la revisión de la literatura se realizó una investigación de tipo bibliográfica, que está basada en técnicas y estrategias que fueron empleadas para identificar, acceder y verificar aquellos documentos como artículos científicos, libros, revistas, entre otros con respecto al tema de estudio, como material de apoyo para el respaldo del trabajo de investigación y que se encuentran con las debidas citas según las normas APA.

3.2. Método de investigación

Para esta investigación se utilizó el método analítico, que contribuye con la descomposición de un fenómeno en sus elementos constitutivos, además ayudo a conocer los objetivos del estudio, sus características con las cuales se pudieron explicar, comprender mejor su comportamiento y establecer nuevas teorías (Echaverría, Gómez, Aristizábal, & Vanegas, 2010).

3.3. Unidad de análisis

Para esta investigación la unidad de análisis está basada en las encuestas realizadas por medio del Sistema de Control Académico (SICOA) específicamente de la Universidad Nacional de Chimborazo, destinadas a recolección de la información de los estudiantes y profesores.

3.4. Población de estudio

La población corresponde a la totalidad de profesores y estudiantes encuestados, matriculados en las cuatro unidades académicas y en la coordinación de competencias lingüísticas. Facultad de Ingenieria, Facultad de Ciencias de la Salud, Facultad de Ciencias de la Educación Humanas y Tecnologías y Facultad de Ciencias Políticas y Administrativas en el periodo mayo – octubre 2020 de las 31 carreras de la Universidad Nacional de Chimborazo, siendo una población total de 9.978 estudiantes y de 644 profesores, que respondieron la encuesta.

3.5. Metodología CRISP-DM

CRISP-DM corresponde a un proceso estándar de la industria para la minería de datos, (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). De acuerdo con KDnuggets (2014), a partir de 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%, razón por la que se utiliza en la presente investigación.

Tabla 2: *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. En el nivel más alto de la metodología constan seis fases para el proceso de minería de datos, las que se describen a continuación:

Comprensión del negocio o problema.

Comprensión de los datos.

Preparación de los datos.

Modelado.

Evaluación.

Implementación.

3.5.1. Comprensión del negocio o problema: Esta es una de las etapas más importante, debido a que si no se tiene una correcta comprensión del problema, o negocio, de nada servirán las etapas siguientes. Las actividades principales de esta etapa son:

- **Identificación del problema*:*** Consiste en entender y delimitar la problemática, así como identificar los requisitos, supuestos, restricciones y beneficios del proyecto.
- **Determinación de objetivos:** Puntualiza las metas a lograr al proponer una solución basada en un modelo de minería de datos Espinosa, (2020).
- **Evaluación de la situación actual:** Especifica el estado actual antes de implementar la solución de minería de datos propuesta, a fin de tener un punto de comparación que permita medir el grado de éxito del proyecto Espinosa, (2020).

3.5.2. Comprensión de datos: Las actividades principales de esta etapa son:

- **Recolección de datos:** Consiste en obtener los datos a utilizar en el proyecto identificando las fuentes, las técnicas empleadas en su recolección, los problemas encontrados en su obtención y la forma como se resolvieron los mismos Espinosa,(2020).
- **Descripción de datos:** Identifica el tipo, formato, volumetría y significado de cada dato Espinosa,(2020).
- **Exploración de datos:** Radica en aplicar pruebas estadísticas básicas que permitan conocer las propiedades de los datos a fin de entenderlos lo mejor posible Espinosa,(2020)

3.5.3. Preparación de datos: Generalmente esta es la etapa que consume más tiempo en el proyecto, y es donde se seleccionan los datos que se transforman de acuerdo con los resultados de la etapa anterior a fin de utilizarlos en la etapa de modelado. Las actividades principales de esta etapa son:

- **Limpieza de datos:** Aplicación de diferentes técnicas, por ejemplo, normalización de datos, discretización de campos numéricos, tratamiento de valores ausentes, tratamiento de duplicados e imputación de datos.
- **Creación de indicadores:** Genera indicadores que potencien la capacidad predictiva de los datos a partir de los datos existentes y ayuden a detectar comportamientos interesantes para modelar.

- **Transformación de datos:** Cambia el formato o estructura de ciertos datos sin modificar su significado, a fin de aplicarles alguna técnica particular en la etapa de modelado.

3.5.4. Modelado: En esta etapa se obtiene propiamente el modelo de minería de datos. Sus actividades principales son:

- **Selección de técnica de modelado*:*** Consiste en la **s**elección de la técnica más apropiada de acuerdo con el tipo de inconveniente a resolver, los datos libres, las herramientas de minería de datos disponibles, así como el dominio de la método elegido.
- **Selección de datos de prueba:** En algunos tipos de modelos se requiere dividir la muestra en datos de entrenamiento y de validación.
- **Obtención del modelo:** Genera el excelente modelo por medio de un proceso iterativo para la modificación de parámetros.

3.5.5. Evaluación del modelo: En esta etapa se determina la propiedad del modelo con soporte en el análisis de las métricas adheridas, analizando los resultados con resultados previos, o también, analizando los resultados con asistencia de especialistas en el dominio del problema. De acuerdo con los resultados de esta etapa se determina seguir con la última fase de la metodología, regresar a alguna de las etapas anteriores o incluso partir de cero con un nuevo proyecto.

3.5.6. Implementación del modelo: Esta etapa explota, mediante acciones específicas, el conocimiento adquirido por medio del modelo. Aquí también es importante documentar los resultados de manera clara para el usuario final y asegurarse de que todas las etapas de la metodología se documenten debidamente para hacer una revisión del proyecto a fin de obtener lecciones aprendidas durante el proceso. Así mismo, monitorear las acciones para detectar áreas de oportunidad o incluso nuevos problemas.

3.6. Técnicas de análisis e interpretación de la información

3.6.1. Desarrollo de la minería de datos

Esta investigación se encuentra realizada bajo la metodología para minería de datos CRISP-DM, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de minería de datos. La metodología se realiza en seis fases como son la comprensión del negocio, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación e implantación. La serie de fases, no es obligatoriamente

rígida. Cada fase es descompuesta en varias tareas regulares de segundo nivel (Moine, 2011) .

3.7. Fase de comprensión del negocio

La misión de la Universidad Nacional de Chimborazo es: crear, desarrollar, transferir y difundir el conocimiento, los saberes y la cultura a través de la aplicación de procesos de formación académica, investigación y vinculación; bajo principios de pertinencia, integralidad, interculturalidad, equidad, preservación del ambiente, fortaleciendo el talento humano, para la construcción de una mejor sociedad.

Su visión es ser la institución de educación superior líder de la Zona 3 del Ecuador, con reconocimiento nacional y proyección Internacional.

Específicamente la institución al realizar la minería de datos busca mejorar los servicios académicos que provee a la comunidad, durante la pandemia de COVID-19, analizando la accesibilidad virtual de sus estudiantes.

3.8. Evaluación de la situación

La Universidad Nacional de Chimborazo, en el mes de mayo de 2020, adoptó la ejecución del periodo académico mayo-octubre 2020 en modalidad virtual, debido a la crisis de COVID-19 y con la finalidad de mitigar el impacto de la pandemia, ha sido recabada información que busca identificar el acceso a la educación virtual por parte de los profesores y los estudiantes a los medios electrónicos. Esta información se encuentra almacenada en el sistema de control académico SICOA y ha sido exportada a una hoja de cálculo, la cual requiere una preparación previa para la comprensión y utilización en el modelado.

3.9. Determinación de los objetivos de la minería de datos

El objetivo de esta investigación es identificar grupos similares entre estudiantes y profesores de la UNACH con accesibilidad a la educación virtual, por medio de análisis clúster, a través de la técnica de minería de datos K-means.

A continuación, se describen los objetivos específicos para alcanzar la meta propuesta:

- Comprender los datos proporcionados a través de la descripción, exploración y verificación de los datos recolectados.
- Preparar la información a través de la selección, limpieza, integración y formateo de los datos.
- Construir el modelo de minería de datos aplicando la técnica de análisis clúster K-means en constante con la técnica K-medoids con la finalidad de identificar su efectividad.

- Analizar los resultados obtenidos de la aplicación del modelo en torno a la identificación de grupos homogéneos entre estudiantes y profesores que presentan características de accesibilidad a la educación virtual.

3.10. Fase de comprensión de los datos

3.10.1. Recolección de datos iniciales

La Universidad Nacional de Chimborazo, ha provisto una base de datos almacenada en el SICOA y exportada a una hoja de cálculo la que contiene la información de accesibilidad virtual de 9.978 estudiantes y 644 profesores. Una vez respondidas 29 preguntas se suman un total de 327.882 registros, con 11 campos, los cuales son los siguientes: Periodo, Facultad, Carrera, Nivel, Estudiante, Documento de Identidad, Sexo, Teléfono Celular, Condiciones de conectividad, Dispositivos disponibles, Tipos de dispositivos.

3.11. Descripción de los datos

Dos bases de datos han sido provistas las cuales poseen: 327.882 registros, con 11 campos para estudiantes y 20.165 con 9 campos para profesores. El detalle de las preguntas formuladas se encuentra en el anexo 1.

A continuación, se presenta la descripción de los datos:

Tabla 3: *Base de datos de estudiantes.*

Nº	Campo	Tipo de dato	Escala	Descripción
1	Periodo	String	Nominal	Periodo académico de los estudiantes
2	Facultad	String	Nominal	Facultad en la que se encuentran los estudiantes
3	Carrera	String	Nominal	Carrera en la que se encuentran los estudiantes
4	Nivel	String	Ordinal	Nivel académico en el cual se encuentran los estudiantes
5	Estudiante	String		Apellidos y nombres del estudiante
6	DocumentoIdentidad	Número		Documento de identificación del estudiante en su mayoría cédula
7	Sexo	String	Nominal	Género del estudiante
8	TelefonoCelular	String		Teléfono celular del estudiante
9	Condiciones de conectividad	String	Nominal	Pregunta relacionada a las condiciones de accesibilidad virtual de los estudiantes
10	Dispositivos disponibles	String	Nominal	Respuesta relacionada a los dispositivos disponibles para la accesibilidad virtual de

				los estudiantes
11	Tipos de dispositivos	String	Nominal	Descripción adicional de los tipos de dispositivos a la accesibilidad virtual.

Fuente: Virginia Tapia (Autor).

Tabla 4: *Base de datos de profesores.*

Nº	**Campo**	**Tipo de dato**	**Escala**	**Descripción**
1	Periodo	String	Nominal	Periodo académico de los profesores
2	Facultad	String	Nominal	Facultad en la que se encuentran los profesores
3	Profesor	String		Apellidos y nombres del profesor
4	DocumentoIdentidad	Número		Documento de identificación del profesor en su mayoría cédula
5	Sexo	String	Nominal	Género del profesor
6	TelefonoCelular	String		Teléfono celular del profesor
7	Condiciones de conectividad	String	Nominal	Pregunta relacionada a las condiciones de accesibilidad virtual de los profesores
8	Dispositivos disponibles	String	Nominal	Respuesta relacionada a los dispositivos disponibles para la accesibilidad virtual de los profesores
9	Tipos de dispositivos	String	Nominal	Descripción adicional de los tipos de dispositivos a la accesibilidad virtual.

Fuente: Virginia Tapia (Autor).

3.12. Exploración de datos

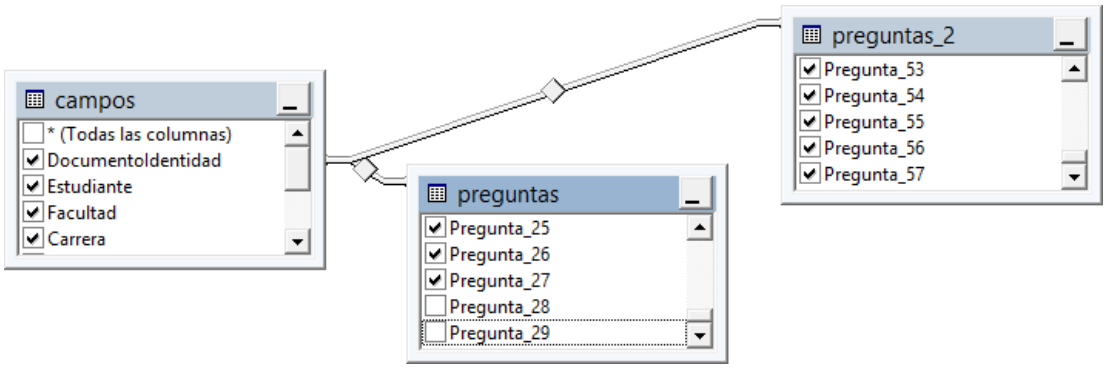
A continuación, se procede a su exploración, cuyo final es hallar una estructura global para los datos, se concatenaron las respuestas 8 y 9 debido a que corresponden a la misma naturaleza, adicionalmente se realizó la conversión de filas a columnas por cada pregunta, para ello se usó de la herramienta QlikView en su versión 12.5. A partir de la utilización de QlikView, se cambiaron el nombre de las preguntas originales a Pregunta 1, Pregunta 2, Pregunta 3, etc (**Ver Anexo 1**), con la finalidad de que se visualicen de forma corta las mismas y una vez que se migre la información a la base de datos y a la herramienta de minería, estas no se reduzcan, dificultando la visualización posterior. Adicionalmente, se utiliza la herramienta de importación de datos de SQL Server, de las tablas generadas con la herramienta QlikView con la finalidad de generar dos vistas una para estudiantes y otra para profesores.

Posteriormente, con la herramienta SQL Server y Talend Opend Studio en su versión 7.3.1, se realiza un primer análisis en donde se puede identificar que las preguntas 8.1, 8.2, 8.3, 8.4 y 12 para el caso

de los estudiantes y 8.1, 8.2., 8.3, 8.4., 10., 11., para profesores se encuentran con un problema debido a que la mismas poseen múltiples respuestas, dificultando poseer un registro por cada documento de identidad, problemas que posteriormente afectarán el resultado del modelo, por lo que se realiza una separación de estas preguntas con la herramienta QlikView según su respuesta, de esta forma se crean 28 preguntas adicionales para estudiantes y 37 para profesores. (**Ver Anexo 1**).

A través de la utilización de SQL server se importa la tabla preguntas_2 y se añade a la vista.

Figura 3: *importación de la tabla preguntas_2.*



Fuente: Virginia Tapia (Autor).

Se utiliza la herramienta Talend Data Quality en su versión 7.3.1 en la cual se hace uso del "Análisis de tabla" presentando los siguientes valores un total de 9.978 registros únicos.

Figura 4: *Análisis de la Tabla.*

Analysis Results

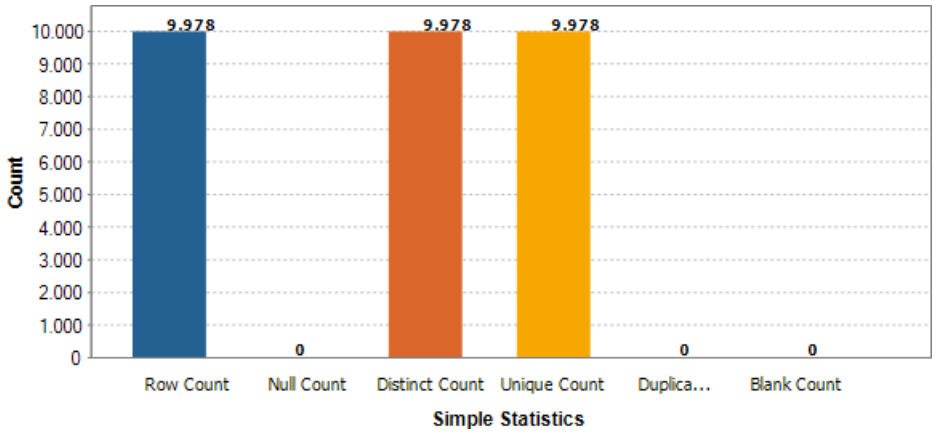
View:View_base_completa1

Where Rule Indicator

Label	Count	%
Row Count	9978.00	100.00%

Fuente: Virginia Tapia (Autor).

Figura 5: *Conteo de datos iniciales*

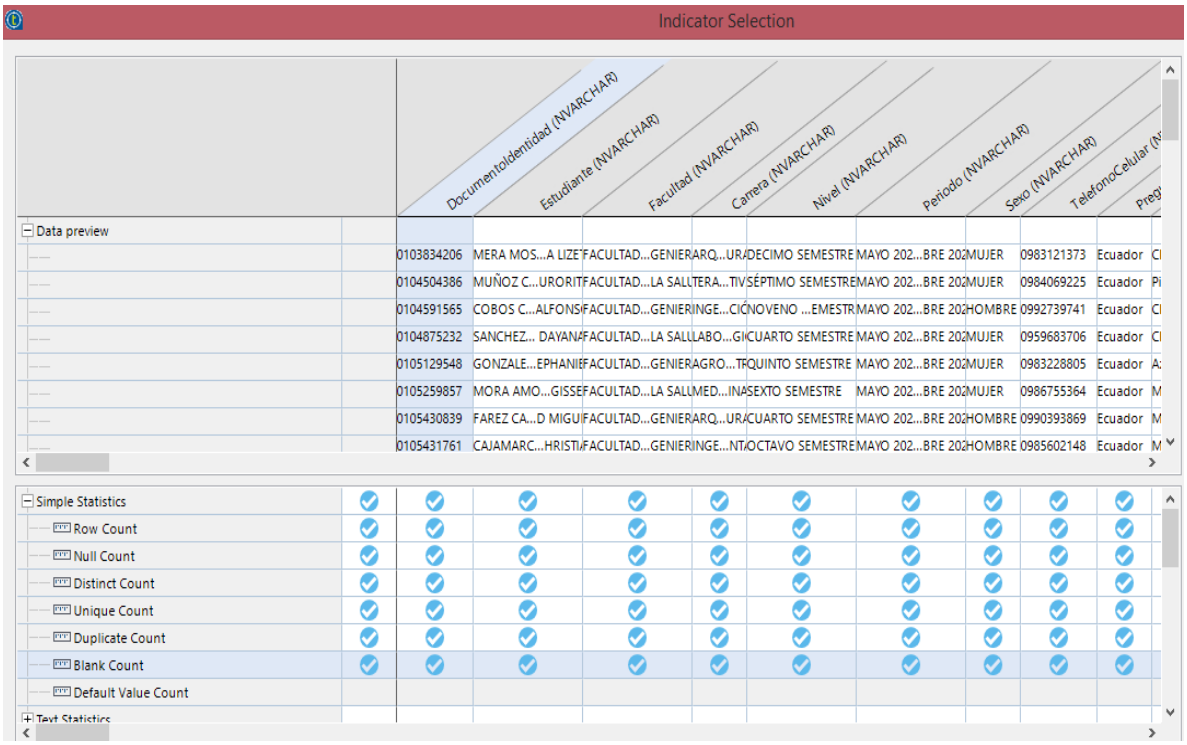


Fuente: Virginia Tapia (Autor).

3.13. Verificación de la calidad de los datos

Utilizando la herramienta Talend Data Quality se realiza un *Análisis de Campos* en el cual se utiliza las secciones de *Estadísticas simples*, *Estadísticas de texto*, *Frecuencia de valores* y *Frecuencia de patrones* para observar la calidad de la base de datos de accesibilidad virtual.

Figura 6: *Vista de interfaz de Talend Data Quality para escoger el análisis*



Fuente: Virginia Tapia (Autor).

a. Base de datos estudiantes

Los campos que no presentan ningún problema de calidad de datos son los siguientes: DocumentoIdentidad, Estudiante, Facultad, Carrera, Sexo, Pregunta_1, Pregunta_7, Pregunta_14, Pregunta_15, Pregunta_16, Pregunta_18, Pregunta_32, Pregunta_34, Pregunta_35, Pregunta_36, Pregunta_40, Pregunta_41, Pregunta_42, Pregunta_43, Pregunta_44, Pregunta_45, Pregunta_46, Pregunta_47, Pregunta_48, Pregunta_51, Pregunta_52, Pregunta_53, Pregunta_56, Pregunta_57.

b. Base de datos profesores

Los campos sin inconvenientes son los siguientes: DocumentoIdentidad, Facultad, Profesor, Sexo, Pregunta_Doc_1, Pregunta_Doc_2, Pregunta_Doc_6, Pregunta_Doc_7, Pregunta_Doc_8, Pregunta_Doc_9, Pregunta_Doc_10, Pregunta_Doc_10, Pregunta_Doc_11, Pregunta_Doc_12, Pregunta_Doc_17, Pregunta_Doc_20, Pregunta_Doc_21, Pregunta_Doc_22, Pregunta_Doc_23, Pregunta_Doc_24, Pregunta_Doc_26, Pregunta_Doc_27, Pregunta_Doc_28, Pregunta_Doc_29, Pregunta_Doc_30, Pregunta_Doc_31, Pregunta_Doc_32, Pregunta_Doc_33, Pregunta_Doc_34, Pregunta_Doc_35, Pregunta_Doc_36, Pregunta_Doc_37, Pregunta_Doc_38, Pregunta_Doc_39, Pregunta_Doc_40, Pregunta_Doc_41, Pregunta_Doc_42, Pregunta_Doc_43, Pregunta_Doc_44, Pregunta_Doc_45, Pregunta_Doc_46, Pregunta_Doc_47, Pregunta_Doc_48, Pregunta_Doc_49, Pregunta_Doc_50, Pregunta_Doc_51, Pregunta_Doc_52, Pregunta_Doc_53, Pregunta_Doc_54, Pregunta_Doc_55, Pregunta_Doc_56, Pregunta_Doc_57, Pregunta_Doc_58, Pregunta_Doc_59, Pregunta_Doc_60, Pregunta_Doc_61, Pregunta_Doc_62.

El análisis de los campos que poseen problemas de calidad de datos tanto de la base de datos de estudiantes como de profesores se puede observar en el **Anexo 2**.

Se exportan los valores a una hoja de cálculo con la finalidad de corregir problemas de escritura errónea más rápidamente.

En el campo **Nivel** se corrigen y mantienen las siguientes clases: DÉCIMO SEMESTRE, SÉPTIMO SEMESTRE, NOVENO SEMESTRE, CUARTO SEMESTRE, QUINTO SEMESTRE, SEXTO SEMESTRE, OCTAVO SEMESTRE, DECIMO, SEGUNDO SEMESTRE, PRIMER SEMESTRE, TERCER SEMESTRE, NOVENO, SEGUNDO, PRIMER, SEPTIMO, OCTAVO, TERCER, CUARTO, CUARTO CURSO, SEXTO CURSO.

En el campo **Periodo** se elimina los 12 registros de “MAYO 2020 - ABRIL 2021”.

En el campo **Pregunta_3** (3. Ciudad actual de residencia durante la emergencia sanitaria), se

corrigen múltiples valores para que corresponda a una escritura estándar. Algunos valores de provincia se eliminan y valores en forma de caracteres que no tienen respuesta válida.

En los siguientes campos se retiran las tildes y los caracteres especiales, se transforma a mayúsculas con la finalidad de mejorar en cierta parte la calidad de datos al tratarse de preguntas abiertas, en la cual existen muchas respuestas, pero que sin embargo puede hallarse algún patrón importante.

Base de datos profesores

Periodo

Se eliminan 3 filas correspondientes al periodo MAYO 2020 - ABRIL 2021, por no encontrarse en el periodo de análisis, se unifica en una sola clase el periodo MAYO 2020 - OCTUBRE 2020.

En el campo **Pregunta_Doc_3** (3. Ciudad actual de residencia durante la emergencia sanitaria), se corrigen múltiples valores para que corresponda a una escritura estándar.

Adicionalmente en la sección de formateo de los datos se hará referencia a varios cambios que se realizan a los atributos para que puedan ser utilizados en el análisis de Clustering.

3.14. Fase de preparación de los datos

3.14.1 Selección de datos

Los siguientes campos no serán utilizados para el análisis y por ende para el proceso de preparación de los datos:

- **DocumentoIdentidad**: Por ser un campo único, este campo no se utilizará para el análisis, sin embargo se mantendrá en la selección de campos.
- **TelefonoCelular**: Por ser un campo único.
- **Pregunta_5** (5. Dirección de su lugar de residencia durante la emergencia sanitaria): Por representar 2221 clases diferentes, que no contribuye al análisis de clústeres.
- **Pregunta_12, Pregunta_13, Pregunta_21, Pregunta_28, Pregunta_29:** Debido a que fueron separadas en las preguntas desde la 30 a la 57.

Limpieza de los datos

Se utiliza las siguientes sentencias SQL, con la finalidad que el valor "-", sea reconocido por la herramienta de rapidminer como errores se utiliza la siguiente sentencia SQL en SQL Server:

```
update [dbo].[preguntas] Set [Pregunta_1]=NULL WHERE [Pregunta_1]='-';
update [dbo].[preguntas] Set [Pregunta_2]=NULL WHERE [Pregunta_2]='-';
update [dbo].[preguntas] Set [Pregunta_3]=NULL WHERE [Pregunta_3]='-';
```

```
update [dbo].[preguntas] Set [Pregunta_4]=NULL WHERE [Pregunta_4]='-';
update [dbo].[preguntas] Set [Pregunta_5]=NULL WHERE [Pregunta_5]='-';
...
```

Y con la finalidad que en las preguntas 30 a 57, se pueda reconocer el valor de 2 como el si se reemplaza por 1 debido a que representaba un error por los niveles del estudiante.

```
update [dbo].[preguntas_2] Set [Pregunta_30]=1 WHERE [Pregunta_30]=2;
update [dbo].[preguntas_2] Set [Pregunta_31]=1 WHERE [Pregunta_31]=2;
update [dbo].[preguntas_2] Set [Pregunta_32]=1 WHERE [Pregunta_32]=2;
update [dbo].[preguntas_2] Set [Pregunta_33]=1 WHERE [Pregunta_33]=2;
update [dbo].[preguntas_2] Set [Pregunta_34]=1 WHERE [Pregunta_34]=2;
```

Se exportan los valores a una hoja de cálculo con la finalidad de corregir problemas de mala escritura más rápidamente.

En el campo **Nivel** se corrigen y mantienen las siguientes clases: DÉCIMO SEMESTRE, SÉPTIMO SEMESTRE, NOVENO SEMESTRE, CUARTO SEMESTRE, QUINTO SEMESTRE, SEXTO SEMESTRE, OCTAVO SEMESTRE, DECIMO, SEGUNDO SEMESTRE, PRIMER SEMESTRE, TERCER SEMESTRE, NOVENO, SEGUNDO, PRIMER, SEPTIMO, OCTAVO, TERCER, CUARTO, CUARTO CURSO, SEXTO CURSO.

En el campo **Periodo** se elimina los 12 registros de “MAYO 2020 - ABRIL 2021”.

En el campo **Pregunta_3** (3. Ciudad actual de residencia durante la emergencia sanitaria), se corrigen múltiples valores para que corresponda a una escritura estándar. Algunos valores de provincia se eliminan y valores en forma de caracteres que no tienen respuesta válida.

En los siguientes campos se retira las tildes, los caracteres especiales, se transforma a mayúsculas con la finalidad de mejorar en cierta parte la calidad de datos al tratarse de preguntas abiertas, en la cual existen muchas respuestas, pero que sin embargo puede hallarse algún patrón importante para este análisis.

En el campo Pregunta_4 (4. Parroquia donde se encuentra su lugar de residencia durante la emergencia sanitaria)

En el campo Pregunta_19 (11. Explique aquí alguna circunstancia o condición importante referente a su situación actual, que pueda influir en su función de estudiante dentro del proceso de enseñanza - aprendizaje virtual)

Integración de los datos

La base de datos en SQL Server, incluye 3 tablas de SQL Server denominadas campos, preguntas, preguntas_2; las mismas que son integradas en la vista, "View_base_completa2", sin incluir los campos antes mencionados en "Selección de datos" de este documento.

Formateo de los datos

Para el formateo de los datos se utiliza la herramienta RapidMiner en su versión 9.8, se procede a la preparación de los datos para lo cual se utiliza el siguiente modelo:

- Para corregir los valores con errores, es decir los que presentan valores nulos o blancos, por la cantidad menor de valores nulos en la mayoría de preguntas se va a utilizar el operador "Replace Missing Values".
- Por seguridad se utiliza el operador "Remove Duplicates", con la finalidad de que si se presenta alguna fila duplicada la elimina.
- Se normalizan los datos con el operador "Normalize", debido a que K-means es un algoritmo basado en distancias, y los registros que se presentan en la base de datos no están en iguales condiciones con respecto a sus valores, para que estas compitan en iguales términos, se utiliza la transformación Z incluida en este operador, que resta la media y divide sobre la desviación estándar.

3.15. Herramientas informáticas

3.15.1. Talend data quality

Los datos de alta calidad mejoran la analítica empresarial para aumentar los ingresos y pueden proporcionar eficiencias operativas masivas para reducir los costos.

Talend Data Quality limpia datos inexactos e inconsistentes, identifica y resuelve registros duplicados y brinda la capacidad de aumentar y mejorar sus datos.

Amplía la elaboración de perfiles con paneles de control en tiempo real para obtener información sobre la calidad y también proporciona el conducto no solo para identificar problemas, sino también para crear procesos automáticamente para resolver y limpiar datos.

3.15.2. Herramienta rapidminer

Permite trabajar los procesos que participan en un proyecto: sirve para crear modelos, la adquisición de datos, la selección de datos, la transformación de los datos, la selección de atributos, la transformación de los atributos, el aprendizaje/modelización y la validación. Además, permite el crecimiento de procesos de análisis de datos mediante el encadenamiento de operadores a través

de un medio descriptivo. Lo que hace factible incrementar la productividad a través de modelos que solucionan los problemas de clasificación, predicción y segmentación de la información (Jaramillo & Arias, 2015).

Es una herramienta de Business Inteligence (Inteligencia de Negocios) de código abierto desarrollada y mantenida por RapidMiner. El software es conocido anteriormente como YALE (Yet Another Learning Environment). Entorno para aprendizaje automático y para procesos de minería de datos (Bermúdez & Acevedo, 2010).

CAPÍTULO IV:

Resultados y discusión

4. Fase de preparación de los datos

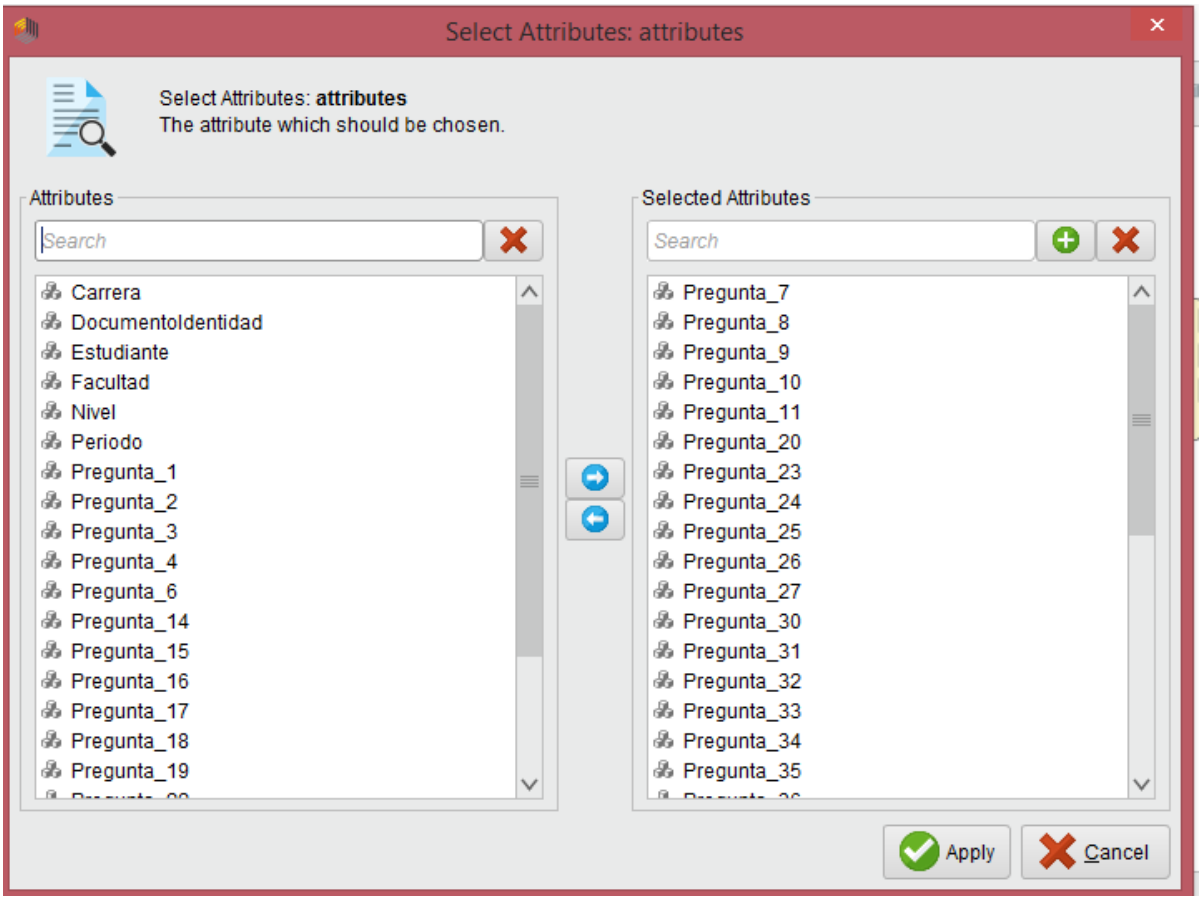
4.1. Selección de datos

En este trabajo de investigación se aplican las técnicas de Clustering basadas en distancias K-means y K-medoids, con fines comparativos para determinar la mejor agrupación, las cuales utilizan datos cuantitativos. Se escogerá de la base de datos limpia los atributos que representen un valor cuantitativo, o que permitan construir un campo de este tipo, como se describe a continuación:

Base de datos estudiantes:

Pregunta_7 a Pregunta_11, Pregunta_20, Pregunta_23 a Pregunta_27, Pregunta_30 a Pregunta_53.

Figura 7: *Vista de la selección inicial de atributos para estudiantes en RapidMiner*

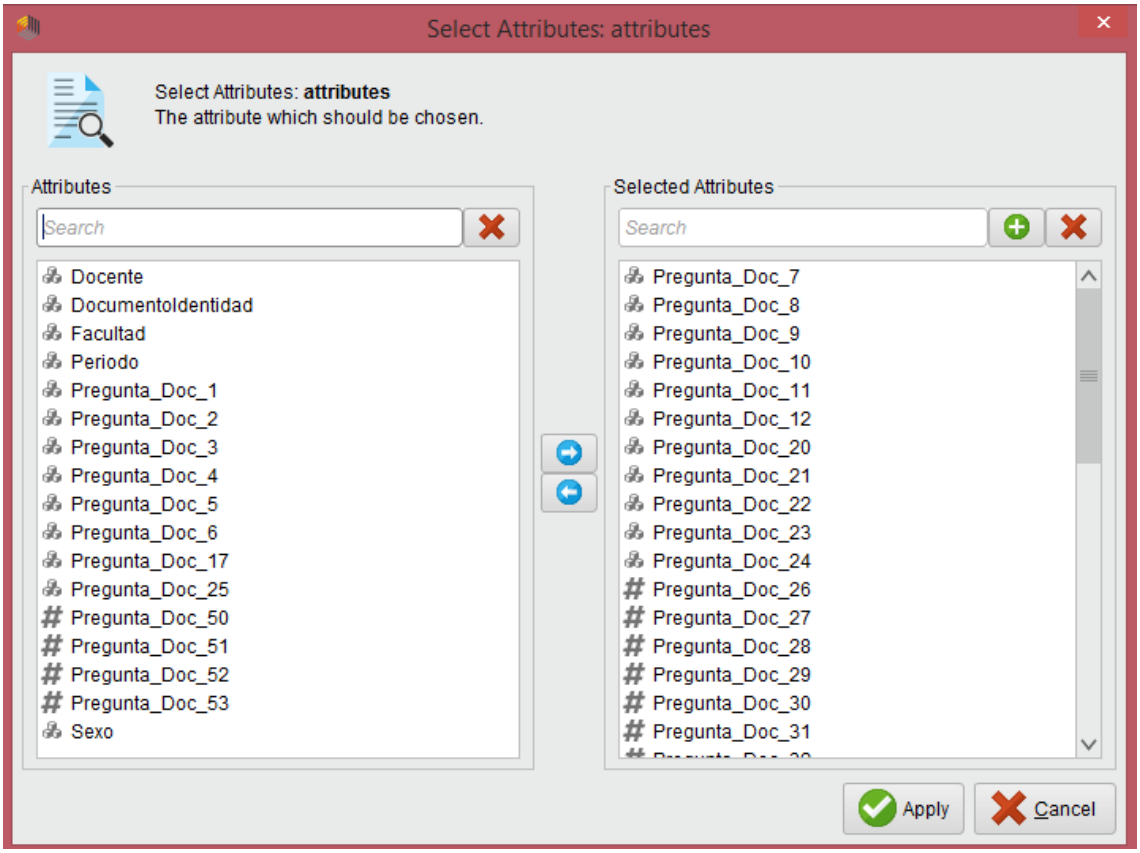


Fuente: Virginia Tapia (Autor).

Base de datos profesores:

Pregunta_Doc_7 a Pregunta_Doc_12, Pregunta_Doc_20 a Pregunta_Doc_24, Pregunta_Doc_26 a Pregunta_Doc_62.

Figura 8: *Vista de la selección inicial de atributos para profesores en RapidMiner*



Fuente: Virginia Tapia (Autor).

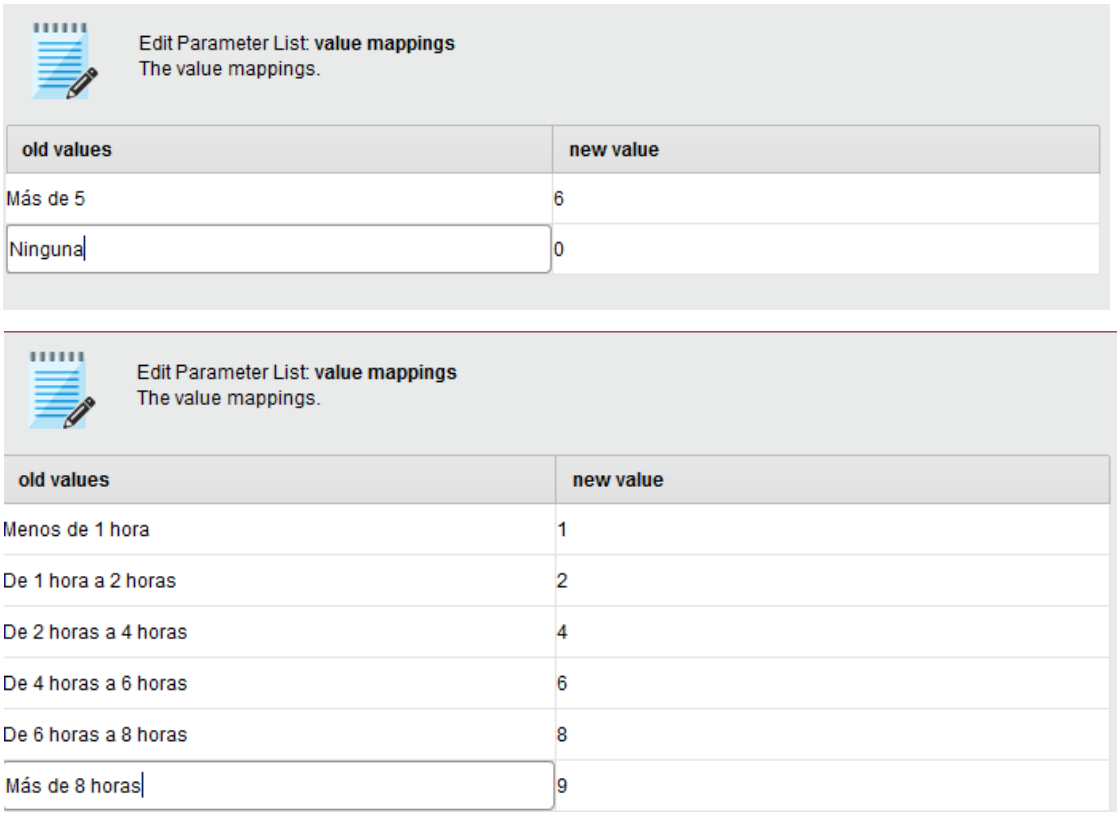
4.2. Limpieza de los datos

En la sección de calidad de datos se adelantó ciertos procedimientos de limpieza debido a que al no tener claro los valores en la transposición de datos realizada podían presentarse valores cuantitativos los cuales podrían servir para este análisis. Una vez seleccionados los datos haciendo uso de la herramienta RapidMiner en su versión 9.8 se usó del operador Replace Missing Values, debido a que son muy pocos los valores en blanco encontrados en esta base de datos y Remove Duplicates con la finalidad de asegurar que todas las filas sean únicas.

Se aplica el operador Map para modificar los valores de varias preguntas tanto de la base de estudiantes como de profesores que poseen los valores de “Más de 5” y “Ninguna”, tales valores

no permiten que estos atributos puedan utilizarse como valores numéricos en el análisis. De igual manera para la pregunta relacionada al acceso a un dispositivo electrónico la cual se presenta en rangos de hora se cambia a valores numéricos entendiéndose hasta el número de horas indicado (Pregunta_27 para estudiantes, Pregunta_Doc_24 para Profesores).

Figura 9: *Utilización de los operadores Map en RapidMiner*



Fuente: Virginia Tapia (Autor).

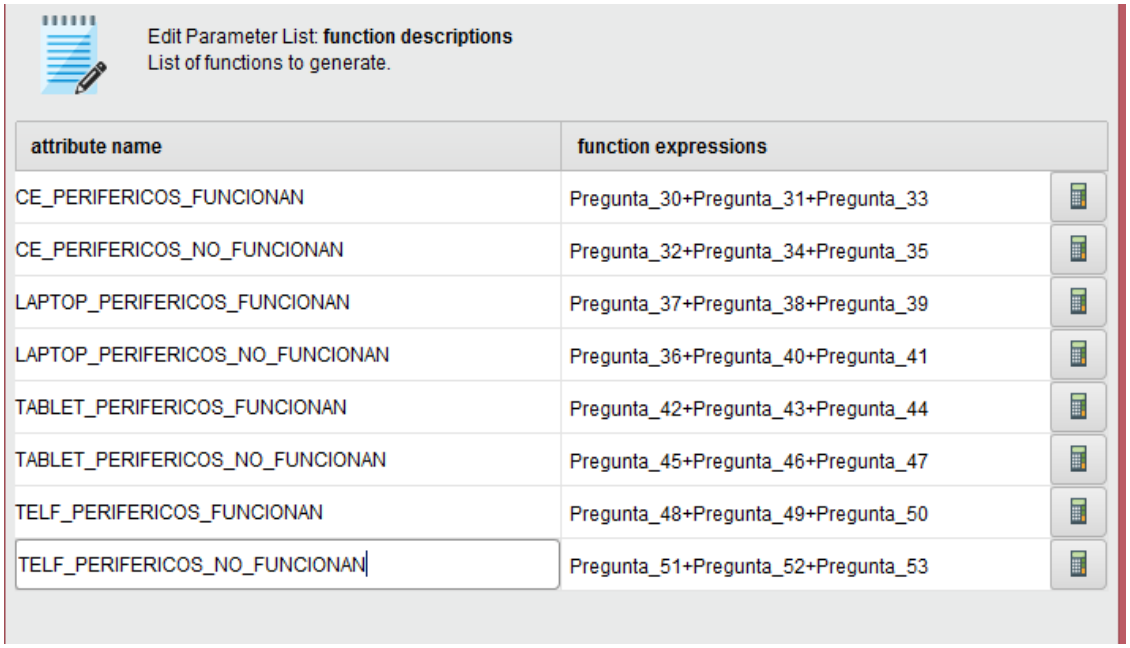
El operador Replace ha sido también utilizado con la finalidad de eliminar el valor de “Mbps” para la Pregunta_23 en la base de datos de estudiantes y para la Pregunta_20 en la de profesores, con la finalidad que se convierta en un valor numérico que pueda ingresar al modelo.

Al poseer un dataset con valores numéricos, sin embargo, aún reconocidos como nominales, se utiliza el Operador Parse Number con la finalidad que estos se transformen a valores numéricos que puedan utilizarse en el análisis.

4.3. Integración y formateo de los datos

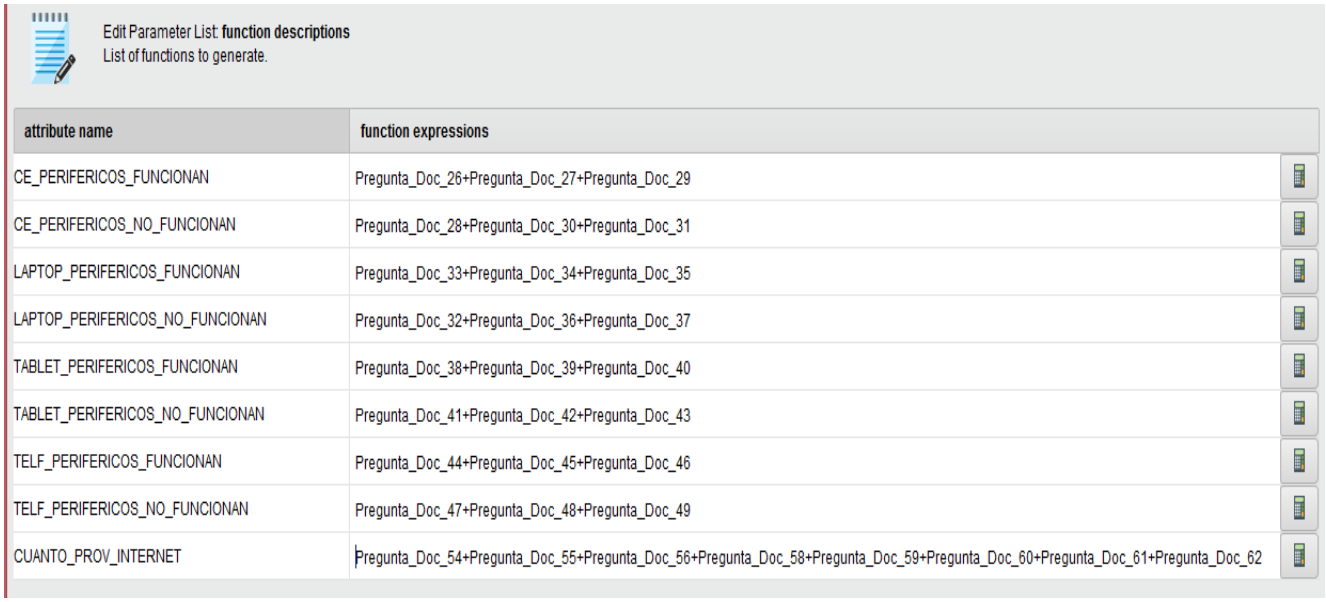
A través de la utilización del operador Generate Atributes se crean atributos cuantitativos para el modelo, estos se pueden observar en las figura 10 y 11. Como se presentan a continuación:

Figura 10: *Generación de nuevos atributos para estudiantes*



Fuente: Virginia Tapia (Autor).

Figura 11: *Generación de nuevos atributos para estudiantes*



Fuente: Virginia Tapia (Autor).

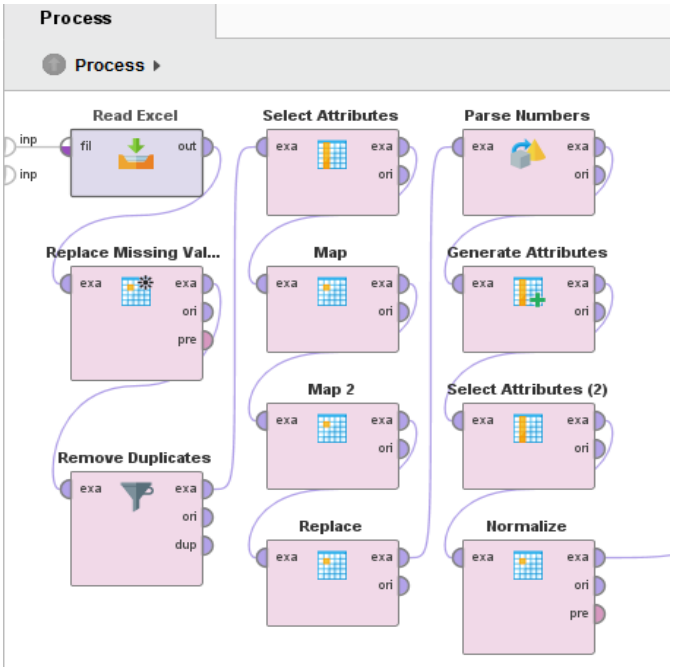
Tabla 5: *Descripción de nuevos atributos creados*

Nombre de nuevo atributo	Descripción
CE_PERIFERICOS_FUNCIONAN	Número de periféricos que funcionan en un computador personal.
CE_PERIFERICOS_NO_FUNCIONAN	Número de periféricos que no funcionan en un computador personal.
LAPTOP_PERIFERICOS_FUNCIONAN	Número de periféricos que funcionan en laptops.
LAPTOP_PERIFERICOS_NO_FUNCIONAN	Número de periféricos que no funcionan en laptops,
TABLET_PERIFERICOS_FUNCIONAN	Número de periféricos que funcionan en tablets.
TABLET_PERIFERICOS_NO_FUNCIONAN	Número de periféricos que no funcionan en tablets,
TELF_PERIFERICOS_FUNCIONAN	Número de periféricos que funcionan en teléfonos inteligentes.
TELF_PERIFERICOS_NO_FUNCIONAN	Número de periféricos que no funcionan en teléfonos inteligentes,
CUANTO_PROV_INTERNET	Número de proveedores de internet que posee en su hogar.

Fuente: Virginia Tapia (Autor).

Se seleccionan nuevamente los atributos útiles y finalmente se normalizan los datos con el operador "Normalize", debido a que K-means es un algoritmo basado en distancias, y los registros que se presentan en la base de datos no están en iguales condiciones con respecto a sus valores, para que estas compitan en iguales términos, se utiliza la transformación Z incluida en este operador, que resta la media y divide sobre la desviación estándar.

Figura 12: *Fase de preparación de los datos en RapidMiner*



Fuente: Virginia Tapia (Autor).

4.4. Fase de modelado

4.4.1. Selección de la técnica de modelado

Las técnicas de modelado que se utilizaron en este trabajo de investigación son K-means y K-medoids, algoritmos que están basados en distancias.

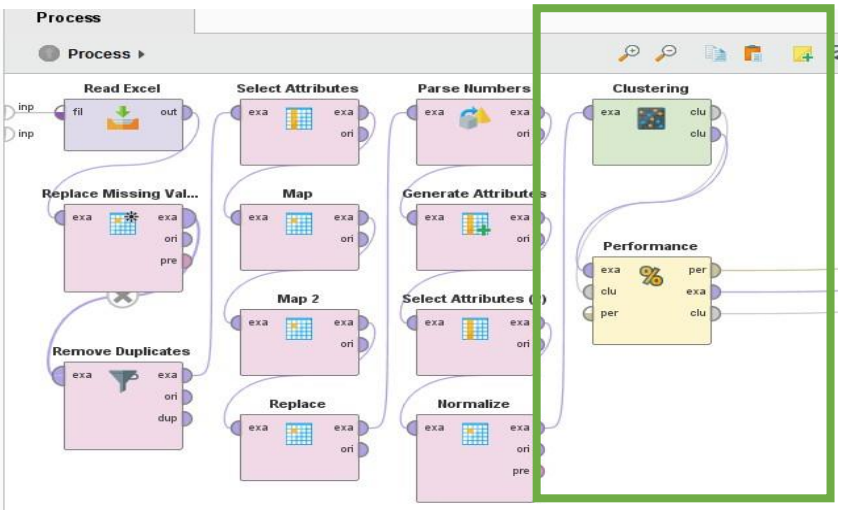
4.4.2. Generación del plan de prueba

Para las pruebas se realizó una comparativa del número de clústers utilizando el operador *Cluster Distance Performance* este operador se utiliza para la evaluación del rendimiento de los métodos de agrupación basados en centroides y entrega una lista de valores de criterios de rendimiento basados en centroides de clúster y la métrica DAVIES-BOULDIN: los algoritmos que producen clústeres con bajas distancias intra-clúster (alta similitud intra-clúster) y altas distancias entre clústeres (baja similitud inter-clúster) tendrán un índice de DAVIES-BOULDIN bajo, el algoritmo de agrupación que produce una colección de los conglomerados con el índice de Davies-Bouldin más pequeño se consideran el mejor algoritmo según este criterio. Para el caso de estudiantes el valor ideal es de 3 para estudiantes y de 2 para profesores. Se han realizado pruebas con un campo de 2 a 7 clústeres.

4.4.3. Construcción del modelo

Se ha construido en la herramienta RapidMiner los siguientes flujos. En los cuales se puede observar el modelo de clustering y el performance, se muestra de forma similar tanto para el algoritmo K-means como para el K-medoids.

Figura 13: *Modelo para Clustering K-means y K-medoids*



Fuente: Virginia Tapia (Autor)

4.4.4. Evaluación del modelo

En las figuras 14 y 15 se pueden observar el promedio de distancia del grupo con respecto al centroide y la métrica de Davis Bouldin de la base de datos de estudiantes. El operador Cluster Distance Performance de Rapid Miner toma este modelo de conglomerado de centroides y el conjunto de conglomerados como entrada y evalúa el rendimiento del modelo en función de los centroides del conglomerado.

Para estudiantes un mejor promedio de la distancia del grupo al centroide está determinado por la técnica de minería de datos K-means, sin embargo, la métrica de Davies Bouldin es más pequeña con K-medoids.

La medida de rendimiento basada en el promedio de la distancia en el clúster, se calcula a través del promedio de la distancia que existe entre el centroide y todos los ejemplos de un grupo, esta distancia se denomina inercia intra clases, y la mejor es la más pequeña, debido a que indica la separación de los individuos con respecto al centro de gravedad. Así también, la métrica de Davies Bouldin, una métrica para evaluar el agrupamiento, los algoritmos que producen conglomerados con distancias bajas entre los mismos (alta similitud entre los conglomerados) tendrán un índice de Davies-Bouldin bajo.

Davies-Bouldin index (DB)

Éste índice está definido 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 14: *Rendimiento Clustering estudiantes en Rapidminer*

K-means

```
PerformanceVector:
Avg. within centroid distance: -16.030
Avg. within centroid distance_cluster_0: -12.802
Avg. within centroid distance_cluster_1: -19.562
Avg. within centroid distance_cluster_2: -41.895
Davies Bouldin: -2.293
```

K-medoids

```
PerformanceVector:
Avg. within centroid distance: -23.047
Avg. within centroid distance_cluster_0: -83.654
Avg. within centroid distance_cluster_1: -26.888
Avg. within centroid distance_cluster_2: -21.128
Davies Bouldin: -2.054
```

Fuente: Virginia Tapia (Autor).

Tabla 6. *Promedio de métricas de rendimiento estudiantes*

Medida	K-Means	K-Medoids
Clúster 0	12.802	83.654
Clúster 1	19.562	26.888
Clúster 2	41.595	21.128
Davies Bouldin	2.293	2.054

Para la base de profesores tanto el promedio de la distancia del grupo de clústeres al centroide y la métrica de Davies Bouldin es mejor con el algoritmo de K-means.

Figura 15: *Rendimiento Clustering profesores en Rapidminer*

K-means

```
PerformanceVector:
Avg. within centroid distance: -18.294
Avg. within centroid distance_cluster_0: -17.637
Avg. within centroid distance_cluster_1: -42.511
Davies Bouldin: -1.292
```

K-medoids

```
PerformanceVector:
Avg. within centroid distance: -25.291
Avg. within centroid distance_cluster_0: -22.012
Avg. within centroid distance_cluster_1: -29.064
Davies Bouldin: -2.125
```

Fuente: Virginia Tapia (Autor).

Tabla 7. *Promedio de métricas de rendimiento profesores*

Medida	K-Means	K-Medoids
Clúster 0	17.637	22.012
Clúster 1	42.511	29.064
Davies Bouldin	1.192	2.125

Fuente: Virginia Tapia (Autor).

Davies-Bouldin index (DB)

Éste índice está definido como:

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

Donde k es el número de clústeres, σi es la distancia promedio entre cada punto en el clúster i y el centroide del clúster, σj es la distancia promedio entre cada punto del clúster j y el centroide del clúster, y d (ci, cj) es la distancia entre los centroides de los 2 clústeres.

En RapidMiner el signo negativo no es representativo en los valores resultantes de Davis Bouldin, debido a que internamente en el cálculo puede multiplicarse por -1 para que pueda ejecutar un minimizado en él, por lo tanto deberá utilizarse como una medida de distancia absoluta.

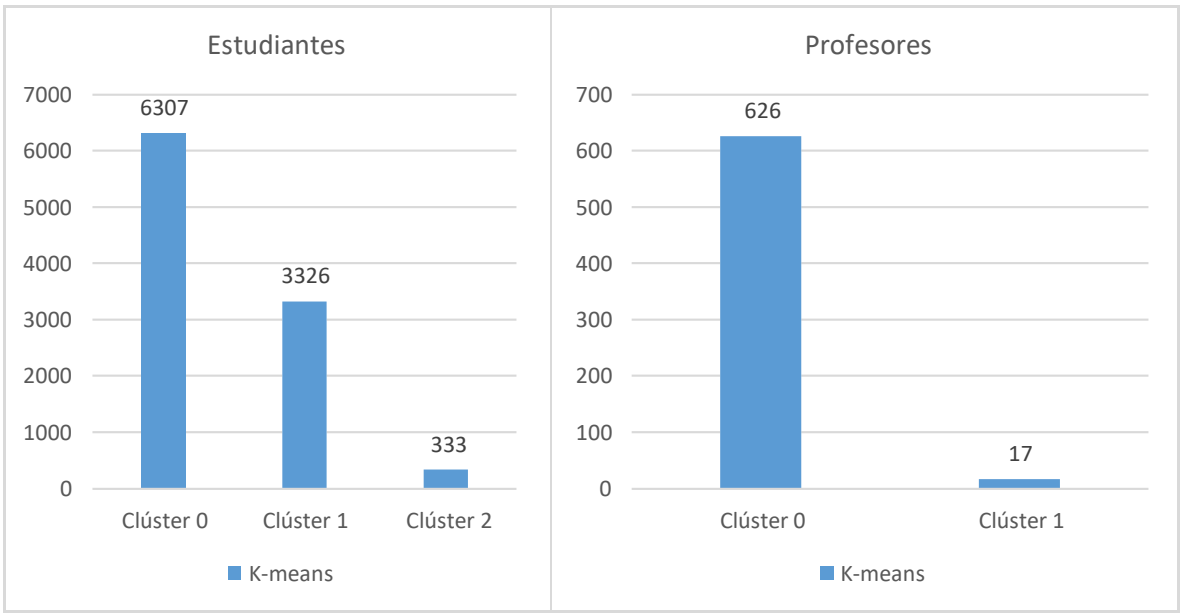
Por tal razón debido a los resultados presentados anteriormente, se decide utilizar la técnica de K-means para realizar el análisis de clustering con todos los datos en la fase de evaluación presentada a continuación.

4.5.Fase de evaluación

4.5.1. Evaluación de los resultados

Se puede observar la figura 16 se presenta el número de individuos tanto de estudiantes como profesores que se agrupan en los diferentes clústeres, utilizando la técnica K-means.

Figura 16: *Resultado de agrupamiento de estudiantes y profesores*



Fuente: Virginia Tapia (Autor).

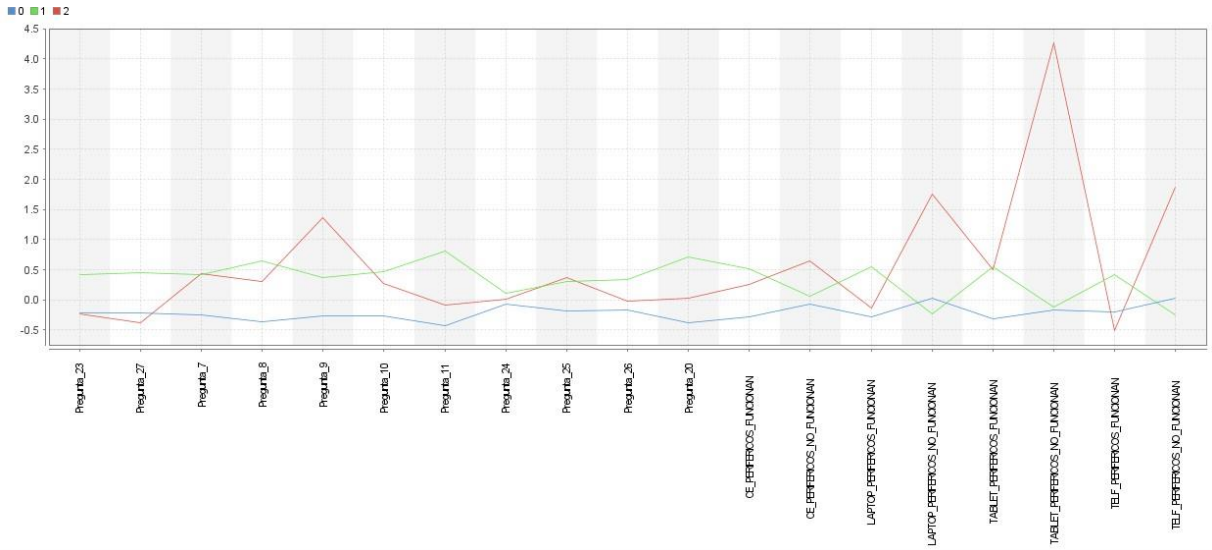
Figura 17: *Porcentaje de agrupamiento*

Fuente: Virginia Tapia (Autor).

En términos porcentuales se identifica que el 63% de estudiantes pertenece al clúster 0 denominado estudiantes con accesibilidad virtual estable, el 34% al clúster 1 denominado estudiantes con alta accesibilidad virtual y únicamente el 3% pertenece al clúster 2 denominado estudiantes con baja accesibilidad virtual. Así también el 97% de profesores pertenecen al clúster 0 denominado docentes con accesibilidad virtual alta y solo el 3% pertenece al clúster 1 denominado docentes con accesibilidad virtual baja.

En la figura 18 que corresponde a la aplicación de K-means en estudiantes, se puede observar que el clúster 0 el más grande en cuanto a su población, se encuentra estable, no se presentan picos altos entre sus variables y todas se encuentra bajo 0, es decir este grupo mayoritario posee una accesibilidad baja a dispositivos virtuales, y sus variaciones están dadas por periféricos que no se encuentran funcionando correctamente en los dispositivos computacionales que poseen. El clúster 1 por su parte representa a un grupo alto de estudiantes, pero mucho menor al clúster 0 aproximadamente la mitad del mismo, que tienen un alto número de dispositivos que no presentan problemas de funcionamiento en cuanto a sus periféricos, a este grupo se puede determinar que tiene una alta accesibilidad virtual. Finalmente, el clúster 2, un grupo minoritario de estudiantes, el cual merece atención no tiene accesibilidad de un internet rápido, además poseen bajo acceso a la disponibilidad de un dispositivo. En su mayoría este grupo accede a clases virtuales a través de una tableta, poseen un bajo número de dispositivos inteligentes y en cuanto a sus periféricos no funcionan en su mayoría, presentándose altas variaciones en el funcionamiento correcto de los mismos, este grupo puede catalogarse como un grupo que no tiene buena accesibilidad virtual.

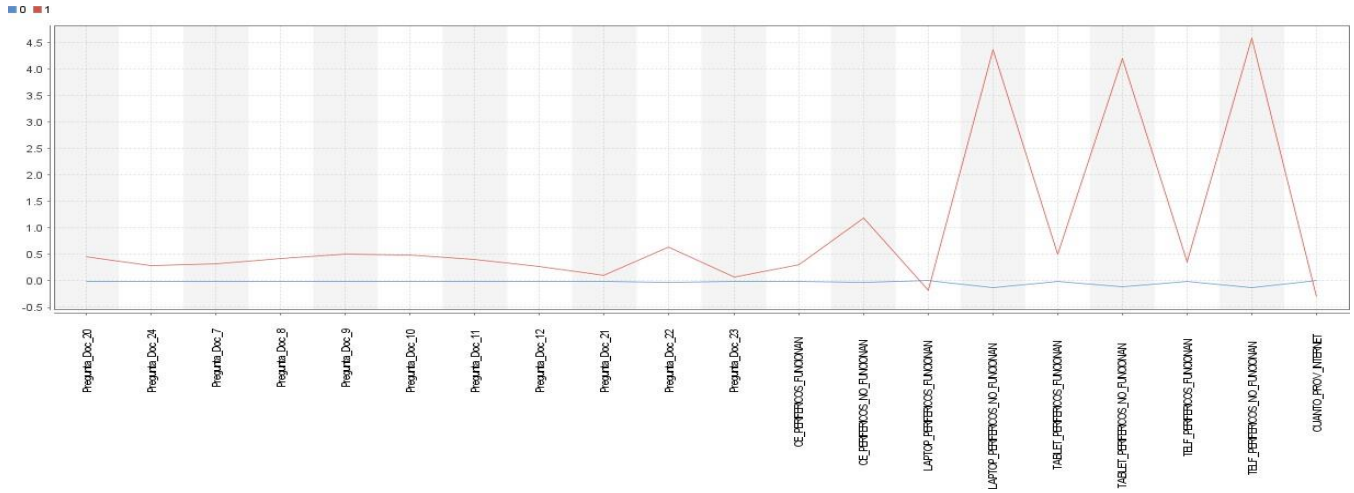
Figura 18: *K-means Estudiantes*



Fuente: Virginia Tapia (Autor)

En la figura 19 que corresponde a la aplicación de K-means en profesores, se puede observar que el clúster 0 el más grande en cuanto a su población, se encuentra estable, no se presentan picos altos entre sus variables y todas se encuentran cercanas a cero, además sus variaciones son bajas y están dadas por un funcionamiento adecuado de sus periféricos computacionales, se puede pensar que este grupo no presenta problemas en cuanto accesibilidad virtual se refiere. Por el contrario, el clúster 1, que es un grupo sumamente más pequeño, este presenta muchas dificultades en cuanto al funcionamiento de sus dispositivos periféricos.

Figura 19: *K-means Profesores*



Fuente: Virginia Tapia (Autor)

4.5.2. Proceso de revisión y determinación de futuras fases

El análisis desarrollado en minería de datos ha permitido evidenciar información valiosa para la toma de decisiones de las autoridades de la Universidad Nacional de Chimborazo, tanto a nivel de estudiantes y profesores con el propósito de mejorar la accesibilidad a la educación virtual.

Al cumplir los objetivos de la minería de datos, este informe representa el mayor aporte en cuanto a esta investigación y la fase de implementación que conlleva su plan, monitorización, está dada por la entrega de los modelos en los cuales se podrá en el futuro utilizar nuevos datos y realizar comparaciones de la situación de conectividad y accesibilidad por profesores y estudiantes.

CONCLUSIONES

- Concerniente a la preparación de los datos, se ha utilizado la herramienta QlikView para la transposición, es decir cambiar los datos mostrados en filas a columnas, de la información de la encuesta de accesibilidad virtual. Además, se ha hecho uso de la herramienta Talend Data Quality para realizar la exploración de la calidad de datos y una vez corregida la base de datos, esta ha sido de gran utilidad para la aplicación de las técnicas de Clustering.
- Se ha utilizado la revisión bibliográfica para elegir la metodología de CRISP-DM para la realización de la minería de datos. Además, se ha seleccionado dos técnicas de agrupamiento en clústeres, las cuales son: K-means y K-medoids, las mismas que están centradas en algoritmos de distancia y utilizan datos cuantitativos para su ejecución. Concerniente a la herramienta de inteligencia artificial, una vez seleccionados y formateados los datos, se ha diseñado e implementado los modelos de Clustering para estudiantes y profesores, utilizando las dos técnicas de agrupamiento seleccionadas a través de la herramienta RapidMiner.
- Una vez diseñado e implementado el modelo de Clustering, se ha comparado su rendimiento mediante la métrica de Davies Bouldin y como resultado de esto la técnica de K-means para estudiantes y profesores proporciona una mejor calificación. Los resultados de la aplicación de las técnicas de Clustering indican que para estudiantes el número óptimo de clústeres son tres y para profesores dos. Para estudiantes el clúster 0 es el mayoritario y presenta una población estable, le sigue el clúster 1 con un alto número de estudiantes que no presentan mayor dificultad de accesibilidad virtual y el clúster 2, un grupo minoritario el cual no tiene una adecuada accesibilidad virtual. Para profesores el clúster 0 el grupo es mayoritario en gran medida y no presenta problemas en este ámbito, por el contrario, el clúster 1 con un número de individuos sumamente bajo si presenta graves problemas de accesibilidad virtual. Al encontrar este grupo 1 que es más pequeño pero que presenta muchas dificultades, eso podría ser negativo para el trabajo en una Institución de Educación Superior donde se utilizan tecnologías para el desarrollo de actividades, más aún ahora que se encuentran desarrollando teletrabajo. Representando un problema complejo al momento de que los docentes necesiten interactuar o hacer uso de las herramientas tecnológicas, indispensables para el teletrabajo y la educación virtual.

RECOMENDACIONES

- Se recomienda que si es utilizada una encuesta como en esta ocasión se prevea la aplicación de respuestas cerradas y utilizar cada pregunta como un campo por separado, para evitar la transposición de datos. Para esta investigación la herramienta QlikView fue de gran utilidad para transponer los datos en lugar de consultas SQL o procedimientos almacenados debido a que reduce una gran cantidad de tiempo.
- Además, se recomienda realizar una exploración previa de la calidad de datos a través del software Talend Data Quality u otra herramienta con este fin, antes de realizar la preparación de los datos.
- Si el tiempo para realización es corto y se requiere aplicar únicamente estas dos técnicas K-means y K-medoids, se recomienda seleccionar de antemano los datos cuantitativos y realizar todo el proceso de preparación únicamente con las columnas pertinentes.
- Se recomienda analizar esta información con otras técnicas de minería de datos para contrastar los resultados obtenidos en esta investigación. Y finalmente para las autoridades se recomienda dar atención a los clústeres en los cuales se han identificado problemas en la accesibilidad virtual.

REFERENCIAS BIBLIOGRÁFICAS

Adamssen, J. (2020). *Inteligencia Artificial, Aprender sobre chatbots, robótica y otras aplicaciones comerciales.*

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

Bermúdez, J. A., & Acevedo, R. Á. (2010). *Análisis para predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies. Doctoral dissertation. Universidad Tecnológica de Pereira. Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la Computación. Ingeniería de Sistemas y Computación*, Pereira.

Benítez, R., Escudero, B. G., Kanaan, S. I., & Robó, D. M. (2014). *Inteligencia Artificial Avanzada.Gran via de les corts Catalanes*, Barcelona.

Beunza, J. N., Puertas, S. E., & Condés, M. E. (2014). *Inteligencia Artificial.* Barcelona, España.

Beunza, J. N., Puertas, S. E., & Condés, M. E. (2020). *Inteligencia Artificia en entornos sanitariosl.* Barcelona, España.

Echaverría, J. D., Gómez, C. A., Aristizábal, M. U., & Vanegas, J. O. (2010). *El método analítico como método natural*. Universidad de Antioquia, Colombia.

Escolano, R. M., Cazorla, Q, Galipienso, A. M., & Lozano O. A. (2003). *Inteligencia Artificial Modelos, Técnicas y Áreas de aplicación.* Madrid.

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

Gonzalez, G. C., & Urbina, C. S. (2014). Presentación del número monográfico "Experiencias y retos actuales en los campus virtuales universitarios", en Revista de Educación a Distancia Número 40. Disponible en: www.um.es/ead/red/40

Gutiérrez, J. A., & Molina, B. T. (2016). Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. *Revista de investigación de la facultad de ingeniería.* Disponible en: https://journal.universidadean.edu.co/index.php/Revistao/article/view/1440

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

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

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

Hernández, P. R., Tomás, M. V., Felipe, R. A., & Núñez, C. F. (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.

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

Chapman, & Hall, (2014) *Data clustering Algorithms and applications.* London, New York: Data Mining and Knowledge Discovery Series.

Jain, A. K., Murty, M. N., & 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.

Jaramillo, A., & Arias, H. P. (2015). Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje. *Revista Tecnológica-ESPOL, 28(1).*

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

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

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

Mamani, R. 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.

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

Pedroza, P. H. (2007). *Sistemas de análisis estadístico con SPSS.* Instituto Nicaragüense de Tecnología Agropecuaria. Managua, Nicaragua.

Pérez, L. C, & Santín, G. D. (2008). *Minería de Datos Técnicas y herramientas* (1ra ed.). Madrid.

Pino, R. D., Gómez, G. A., & Martínez, N. A. (2020). *Introducción a la Inteligencia Artificial*. Sistemas expertos, redes neuronales artificiales y computación evolutiva, Universidad de Oviedo.

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

Rodríguez, C. P., & Castillo, H. A. (2017). *Sociedad Digital en España.* Editorial Ariel, S.A. Barcelona, España.

Rosado, G. A., & Verjel, A. I. (2016). Aplicación de la minería de datos en la educación en línea. *Revista colombiana de Tecnologías de avanzada, 7.*

Ruíz, V. E., Sánchez, J. F., & López, V. S. (2020). *Movilidad virtual de experiencias educativas.* Sociedad Mexicana de computación en la educación A.C. México.

Suárez, F. 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.

Tuya, J. I., Ramos, R. I., & Dolado, C. J. (2020). *Técnicas cuantitativas para la gestión en la ingeniería del software.* Netbiblo, S.L. España.

ANEXOS

Plan Del Proyecto

Tabla 7: *Plan de proyecto de minería de datos.*

Actividades	Duración (días)	Tareas a desarrollar	Técnicas a emplear
Comprensión del negocio o problema	2	Determinar los objetivos del negocio. Evaluación de la situación. Determinar los objetivos de la minería de datos. Realizar el plan del proyecto.	N/A
Comprensión de los datos	3	Recolección de datos iniciales. Descripción de los datos. Exploración de los datos. Verificación de la calidad de los datos.	Ejecución de consultas. Gráficos de frecuencia Resumen de errores en los datos.
Fase de preparación de los datos	10	Selección de datos. Estructuración de los datos. Integración de los datos. Formateo de los datos.	Ejecución de consultas Creación de vistas
Fase de modelado	5	Selección de la técnica de modelado. Generación del plan de prueba. Construcción del modelo. Evaluación del modelo.	Análisis de clúster K-means, K-medoids
Fase de evaluación	5	Evaluación de los resultados Proceso de revisión. Determinación de futuras fases.	Generación de gráficos y tablas estadísticas.
Fase de implantación	5	Informe final. Revisión del proyecto.	N/A

Fuente: Virginia Tapia (Autor).

ANEXO 1: Actualización del nombre de las preguntas

Base de datos de estudiantes

1. ¿En qué país se encuentra actualmente durante la Emergencia Sanitaria?	Pregunta_1
2. Provincia actual de residencia durante la Emergencia Sanitaria	Pregunta_2
3. Ciudad actual de residencia durante la emergencia sanitaria	Pregunta_3
4. Parroquia donde se encuentra su lugar de residencia durante la emergencia sanitaria	Pregunta_4
5. Dirección de su lugar de residencia durante la emergencia sanitaria	Pregunta_5
6. ¿Su hogar tiene acceso a equipos tecnológicos? Entendiéndose por equipos tecnológicos computadoras, tablets, celulares, etc.	Pregunta_6
7.1 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos: Computadora de escritorio	Pregunta_7
7.2 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos: Laptop	Pregunta_8
7.3 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos: Tablet	Pregunta_9
7.4 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos: Radio	Pregunta_10
7.5. Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos: Teléfonos inteligentes (con acceso a internet)	Pregunta_11
8.2 Indicar el estado de sus dispositivos tecnológicos: Laptop	Pregunta_12
8.4 Indicar el estado de sus dispositivos tecnológicos: Teléfono inteligente	Pregunta_13
9. ¿Su hogar dispone de internet?	Pregunta_14
10.1 Si su hogar no dispone de acceso a equipos tecnológicos o acceso a internet. Usted ha considerado: Comprar los equipos tecnológicos necesarios	Pregunta_15

10.2 Si su hogar no dispone de acceso a equipos tecnológicos o acceso a internet. Usted ha considerado: Contratar servicio de internet o plan de datos	Pregunta_16
10.3 Si su hogar no dispone de acceso a equipos tecnológicos o acceso a internet. Usted ha considerado: Suspender sus estudios hasta que se establezcan nuevamente las clases presenciales	Pregunta_17
10.4 Si su hogar no dispone de acceso a equipos tecnológicos o acceso a internet. Usted ha considerado: Solicitar una beca o ayuda económica para adquirir las herramientas tecnológicas necesarias para realizar sus estudios	Pregunta_18
11. Explique aquí alguna circunstancia o condición importante referente a su situación actual, que pueda influir en su función de estudiante dentro del proceso de enseñanza - aprendizaje virtual.	Pregunta_19
7.5 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos. Televisión	Pregunta_20
12. El acceso a internet desde su hogar es a través de:	Pregunta_21
13. ¿Con qué proveedor de internet tiene contratado su plan de datos o internet fijo/inalámbrico?	Pregunta_22
14. Escoja la velocidad de su servicio de internet	Pregunta_23
15. ¿Cuántas personas en su hogar se encuentran actualmente estudiando de manera online?	Pregunta_24
16. ¿Cuántas personas en su hogar se encuentran actualmente tele trabajando?	Pregunta_25
17. ¿Cuántas personas hacen uso del internet en su hogar simultáneamente, incluyéndose usted?	Pregunta_26
18. ¿Cuántas horas al día podría usted tener acceso a un dispositivo electrónico de su hogar conectado a internet? (computadora, laptop, tablet, celular)	Pregunta_27
8.1 Indicar el estado de sus dispositivos tecnológicos: Computador de escritorio	Pregunta_28
8.3 Indicar el estado de sus dispositivos tecnológicos: Tablet	Pregunta_29
Creadas a partir de la transposición y separación de preguntas con varias respuestas	
8.1.1. Micrófono funciona adecuadamente	Pregunta_30
8.1.2. Parlantes o audífonos funcionan adecuadamente	Pregunta_31
8.1.3. Cámara no funciona	Pregunta_32
8.1.4. Cámara funciona adecuadamente	Pregunta_33
8.1.5. Micrófono no funciona	Pregunta_34
8.1.6. Parlantes o audífonos no funcionan	Pregunta_35
8.2.1. Micrófono no funciona	Pregunta_36
8.2.2. Micrófono funciona adecuadamente	Pregunta_37
8.2.3. Parlantes o audífonos funcionan adecuadamente	Pregunta_38
8.2.4. Cámara funciona adecuadamente	Pregunta_39
8.2.5. Parlantes o audífonos no funcionan	Pregunta_40
8.2.6. Cámara no funciona	Pregunta_41
8.3.1. Micrófono funciona adecuadamente	Pregunta_42
8.3.2. Cámara funciona adecuadamente	Pregunta_43
8.3.3. Parlantes o audífonos funcionan adecuadamente	Pregunta_44
8.3.4. Micrófono no funciona	Pregunta_45
8.3.5. Cámara no funciona	Pregunta_46
8.3.6. Parlantes o audífonos no funcionan	Pregunta_47
8.4.1. Micrófono funciona adecuadamente	Pregunta_48
8.4.2. Cámara funciona adecuadamente	Pregunta_49

8.4.3. Parlantes o audífonos funcionan adecuadamente	Pregunta_50
8.4.4. Micrófono no funciona	Pregunta_51
8.4.5. Cámara no funciona	Pregunta_52
8.4.6. Parlantes o audífonos no funcionan	Pregunta_53

12.1. Servicio de internet o datos compartidos de algún vecino o familiar que no vive en su hogar.	Pregunta_54
12.2. Servicio contratado de internet fijo o inalámbrico con algún proveedor como Cnt, Claro, Netlife, u otro.	Pregunta_55
12.3. Internet desde dispositivos móviles, plan de datos (para celulares o tablets)	Pregunta_56
12.4. Ninguno	Pregunta_57

Base de datos de profesores

1. ¿En qué país se encuentra actualmente durante la Emergencia Sanitaria?	Pregunta_Doc_1
2. Provincia actual de residencia durante la Emergencia Sanitaria	Pregunta_Doc_2
3. Ciudad actual de residencia durante la emergencia sanitaria	Pregunta_Doc_3
4. Parroquia donde se encuentra su lugar de residencia durante la emergencia sanitaria	Pregunta_Doc_4
5. Dirección de su lugar de residencia durante la emergencia sanitaria	Pregunta_Doc_5
6. ¿Su hogar tiene acceso a equipos tecnológicos? Entendiéndose por equipos tecnológicos computadoras, tablets, celulares, etc.	Pregunta_Doc_6
7.1 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos. Computador de escritorio	Pregunta_Doc_7
7.2 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos. Laptop	Pregunta_Doc_8
7.3 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos. Tablet	Pregunta_Doc_9
7.4 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos. Radio	Pregunta_Doc_10
7.5 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos. Televisión	Pregunta_Doc_11
7.6 Escoger los dispositivos tecnológicos que dispone su hogar e indicar el número de los mismos. Teléfonos inteligentes (con acceso a internet)	Pregunta_Doc_12
8.1 Indicar el estado de sus dispositivos tecnológicos: Computador de escritorio	Pregunta_Doc_13
8.2 Indicar el estado de sus dispositivos tecnológicos: Laptop	Pregunta_Doc_14
8.3 Indicar el estado de sus dispositivos tecnológicos: Tablet	Pregunta_Doc_15
8.4 Indicar el estado de sus dispositivos tecnológicos: Teléfono inteligente	Pregunta_Doc_16
9. ¿Su hogar dispone de internet?	Pregunta_Doc_17
10. El acceso a internet desde su hogar es a través de:	Pregunta_Doc_18
11. ¿Con qué proveedor de internet tiene contratado su plan de datos o internet fijo/inalámbrico? *	Pregunta_Doc_19
12. Escoja la velocidad de su servicio de internet	Pregunta_Doc_20
13. ¿Cuántas personas en su hogar se encuentran actualmente estudiando online?	Pregunta_Doc_21
14. ¿Cuántas personas en su hogar se encuentran actualmente tele trabajando?	Pregunta_Doc_22
15. ¿Cuántas personas hacen uso del internet en su hogar simultáneamente, incluyéndose usted?	Pregunta_Doc_23
16. ¿Cuántas horas al día podría usted tener acceso a un dispositivo electrónico de su hogar conectado a internet? (computadora, laptop, tablet, celular)	Pregunta_Doc_24
17. Explique aquí alguna circunstancia o condición importante referente a su situación actual, que pueda influir en su función de profesor dentro del proceso de enseñanza - aprendizaje virtual	Pregunta_Doc_25
Creadas a partir de la trasposición y separación de preguntas con varias respuestas	
8.1.1. Micrófono funciona adecuadamente	Pregunta_Doc_26

8.1.2. Parlantes o audífonos funcionan adecuadamente	Pregunta_Doc_27
8.1.3. Cámara no funciona	Pregunta_Doc_28
8.1.4. Cámara funciona adecuadamente	Pregunta_Doc_29
8.1.5. Micrófono no funciona	Pregunta_Doc_30
8.1.6. Parlantes o audífonos no funcionan	Pregunta_Doc_31
8.2.1. Micrófono no funciona	Pregunta_Doc_32

8.2.2. Micrófono funciona adecuadamente	Pregunta_Doc_33
8.2.3. Parlantes o audífonos funcionan adecuadamente	Pregunta_Doc_34
8.2.4. Cámara funciona adecuadamente	Pregunta_Doc_35
8.2.5. Parlantes o audífonos no funcionan	Pregunta_Doc_36
8.2.6. Cámara no funciona	Pregunta_Doc_37
8.3.1. Micrófono funciona adecuadamente	Pregunta_Doc_38
8.3.2. Cámara funciona adecuadamente	Pregunta_Doc_39
8.3.3. Parlantes o audífonos funcionan adecuadamente	Pregunta_Doc_40
8.3.4. Micrófono no funciona	Pregunta_Doc_41
8.3.5. Cámara no funciona	Pregunta_Doc_42
8.3.6. Parlantes o audífonos no funcionan	Pregunta_Doc_43
8.4.1. Micrófono funciona adecuadamente	Pregunta_Doc_44
8.4.2. Cámara funciona adecuadamente	Pregunta_Doc_45
8.4.3. Parlantes o audífonos funcionan adecuadamente	Pregunta_Doc_46
8.4.4. Micrófono no funciona	Pregunta_Doc_47
8.4.5. Cámara no funciona	Pregunta_Doc_48
8.4.6. Parlantes o audífonos no funcionan	Pregunta_Doc_49
10.1. Servicio contratado de internet fijo o inalámbrico con algún proveedor como Cnt, Claro, Netlife, u otro.	Pregunta_Doc_50
10.2. Internet desde dispositivos móviles, plan de datos (para celulares o tablets	Pregunta_Doc_51
10.3. Servicio de internet o datos compartidos de algún vecino o familiar que no vive en su hogar	Pregunta_Doc_52
10.4. Ninguno	Pregunta_Doc_53
11.1. Cnt	Pregunta_Doc_54
11.2. Netlife	Pregunta_Doc_55
11.3. Otro:	Pregunta_Doc_56
11.5. Ninguno	Pregunta_Doc_57
11.4. Movistar	Pregunta_Doc_58
11.5. Claro	Pregunta_Doc_59
11.6. Puntonet	Pregunta_Doc_60
11.7. Tv Cable	Pregunta_Doc_61
11.8. Telconet	Pregunta_Doc_62

ANEXO 2: Calidad de datos

Base de datos de estudiantes

A continuación, se presentan los campos los cuales necesitan un formateo de datos y presentan probl

Nivel

Valor	Count	%
PRIMER SEMESTRE	1546	15.49%
SEGUNDO SEMESTRE	1423	14.26%
TERCER SEMESTRE	1256	12.59%
CUARTO SEMESTRE	1032	10.34%
SEXTO SEMESTRE	928	9.30%
QUINTO SEMESTRE	891	8.93%
OCTAVO SEMESTRE	693	6.95%
SEPTIMO SEMESTRE	626	6.27%
NOVENO SEMESTRE	469	4.70%
DECIMO SEMESTRE	417	4.18%

▾ **Value Low Frequency**

Valor	Count	%
CUARTO CURSO	5	0.05%
SEXTO CURSO	7	0.07%
DÉCIMO SEMESTRE	16	0.16%
PRIMER	27	0.27%
TERCER	35	0.35%
CUARTO	36	0.36%
SEGUNDO	44	0.44%
TERCERO SEMESTRE	63	0.63%
OCTAVO	77	0.77%
DECIMO	84	0.84%

Se presenta un problema en las clases debido a que existen clases que no se encuentran representadas sea por año o por semestre respectivamente.

Periodo

▼ **Value Frequency**

Valor	Count	%
MAYO 2020 - OCTUBRE 2020	9966	99.88%
MAYO 2020 - ABRIL 2021	12	0.12%

Existen únicamente 12 datos del periodo “MAYO 2020 – ABRIL 2021”, los cuales no son datos de análisis para esta investigación.

TelefonoCelular

▼ **Simple Statistics**

Label	Count	%
Row Count	9978	100.00%
Null Count	0	0.00%
Distinct Count	2	0.02%
Unique Count	0	0.00%
Duplicate Count	2	0.02%
Blank Count	0	0.00%

Presenta 2 teléfonos duplicados, el cual no representa mayor inconveniente debido a que este campo será retirado para el análisis.

Pregunta_2

▼ **Value Low Frequency**

Valor	Count	%
-	1	0.01%
Santa Elena	9	0.09%
Galápagos	16	0.16%
Manabí	29	0.29%
Azuay	30	0.30%
Los Ríos	32	0.32%
Zamora Chinchipe	50	0.50%
Guayas	58	0.58%
Esmeraldas	67	0.67%
El Oro	78	0.78%

Presenta únicamente un valor en blanco.

Pregunta_3

Valor	Count	%
Riobamba	4787	47.98%
Ambato	590	5.91%
Quito	384	3.85%
Guano	367	3.68%
Guaranda	226	2.26%
LATACUNGA	213	2.13%
PUYO	177	1.77%
Santo Domingo	125	1.25%
TENA	107	1.07%
Macas	105	1.05%

Value Low Frequency

Valor	Count	%
Riobsmba	1	0.01%
Pepinales	1	0.01%
Catón Cumandá	1	0.01%
YAGUACHI.	1	0.01%
Quito- Tumbaco	1	0.01%
Parroquia Tanicuchi barrio El Calv...	1	0.01%
canton Daule, guayas	1	0.01%
Riobambasi	1	0.01%
Guranda	1	0.01%
Montevideo	1	0.01%

Pregunta_4

Value Frequency

Valor	Count	%
Lizarzaburu	1045	10.47%
Maldonado	788	7.90%
Veloz	787	7.89%
VELASCO	594	5.95%
La Matriz	338	3.39%
Lizarzaburo	272	2.73%
YARUQUIES	151	1.51%
lican	147	1.47%
San Luis	121	1.21%
Matriz	116	1.16%

Value Low Frequency

Valor	Count	%
La Matriz Guano	1	0.01%
Guano,La Matriz	1	0.01%
Gabriel veintimilla	1	0.01%
Huachi Chico	1	0.01%
Sector Urbano	1	0.01%
El Playón de San Francisco	1	0.01%
Abraham Calazacón	1	0.01%
Parroquia santa ana	1	0.01%
Patricia Pilar	1	0.01%

Presentan problemas de calidad de datos en mala escritura que puede ser corregida.

Pregunta_5

Value Frequency

Valor	Count	%
Lizarzaburu	1045	10.47%
Maldonado	788	7.90%
Veloz	787	7.89%
VELASCO	594	5.95%
La Matriz	338	3.39%
Lizarzaburo	272	2.73%
YARUQUIES	151	1.51%
lican	147	1.47%
San Luis	121	1.21%
Matriz	116	1.16%

Value Low Frequency

Valor	Count	%
La Matriz Guano	1	0.01%
Guano,La Matriz	1	0.01%
Gabriel veintimilla	1	0.01%
Huachi Chico	1	0.01%
Sector Urbano	1	0.01%
El Playón de San Francisco	1	0.01%
Abraham Calazacón	1	0.01%
Parroquia santa ana	1	0.01%
Patricia Pilar	1	0.01%

Posee problemas de escritura y valores que no corresponden a una respuesta.

Pregunta_6

Value Low Frequency

Valor	Count	%
-	1	0.01%
NO	409	4.10%
SI	9568	95.89%

Presenta un valor en blanco

Pregunta_8

Value Frequency

Valor	Count	%
1	6210	62.24%
0	1857	18.61%
2	1489	14.92%
3	303	3.04%
4	78	0.78%
5	26	0.26%
Más de 5	13	0.13%
-	2	0.02%

Presenta dos valores nulos

Pregunta_9

Value Frequency

Valor	Count	%
0	8633	86.52%
1	1046	10.48%
2	202	2.02%
3	58	0.58%
4	23	0.23%
Más de 5	11	0.11%
5	4	0.04%
-	1	0.01%

Presenta un valor nulo

Pregunta_10

▾ **Value Frequency**

Valor	Count	%
1	5992	60.05%
0	2625	26.31%
2	1064	10.66%
3	212	2.12%
4	53	0.53%
Más de 5	22	0.22%
5	9	0.09%
-	1	0.01%

Presenta un valor nulo

Pregunta_11

▾ **Value Frequency**

Valor	Count	%
1	3156	31.63%
2	2628	26.34%
3	2006	20.10%
4	1092	10.94%
5	458	4.59%
0	431	4.32%
Más de 5	205	2.05%
-	2	0.02%

Presenta dos valores nulos

Pregunta_17

▾ **Value Low Frequency**

Valor	Count	%
-	1	0.01%
SI	2322	23.27%
NO	7655	76.72%

Un valor en blanco

Pregunta_19

▼ **Value Low Frequency**

Valor	Count	%
??	1	0.01%
A cubierta	1	0.01%
2 Personas de distintas carreras uti...	1	0.01%
A pesar de los cambios en el cam...	1	0.01%
A VECES EL INTERNET ES ALGO IN...	1	0.01%
2 computadoras para 4 personas	1	0.01%
A pesar de la emergencia algunos ...	1	0.01%
a pesar que antes no contaba con ...	1	0.01%
a veces el internet es lento	1	0.01%
A pesar de poseer herramientas te...	1	0.01%

Problemas de escritura, no se puede encontrar clases representativas, por tratarse de una pregunta abierta.

Pregunta_20

▼ **Value Low Frequency**

Valor	Count	%
-	1	0.01%
5	170	1.70%
4	314	3.15%
0	629	6.30%
3	1033	10.35%
2	2966	29.73%
1	4865	48.76%

Un valor en blanco

Pregunta_22

▼ **Value Low Frequency**

Valor	Count	%
-	1	0.01%
TELCONET	55	0.55%
TV CABLE	160	1.60%
MOVISTAR	161	1.61%
PUNTONET	223	2.23%
CLARO	380	3.81%
NINGUNO	812	8.14%
NETLIFE	1031	10.33%
OTRO	1764	17.68%
CNT	5391	54.03%

Presenta un valor en blanco

Pregunta_23

▼ **Value Low Frequency**

Valor	Count	%
-	1	0.01%
30 Mbps	77	0.77%
40 Mbps	96	0.96%
100 Mbps	139	1.39%
60 Mbps	307	3.08%
50 Mbps	535	5.36%
25 Mbps	562	5.63%
Ninguna	784	7.86%
15 Mbps	864	8.66%
20 Mbps	866	8.68%

Un valor en blanco

Pregunta_24

▼ **Value Frequency**

Valor	Count	%
2	3609	36.17%
3	2594	26.00%
1	1920	19.24%
4	956	9.58%
0	338	3.39%
5	323	3.24%
Más de 5	237	2.38%
-	1	0.01%

Un valor en blanco

Pregunta_25

▼ **Value Frequency**

Valor	Count	%
0	5075	50.86%
1	2826	28.32%
2	1501	15.04%
3	330	3.31%
4	138	1.38%
Más de 5	55	0.55%
5	51	0.51%
-	2	0.02%

Un valor en blanco

Pregunta_26

▼ **Value Frequency**

Valor	Count	%
4	2448	24.53%
3	2326	23.31%
5	1808	18.12%
Más de 5	1703	17.07%
2	1234	12.37%
1	458	4.59%
-	1	0.01%

Un valor en blanco

Pregunta_27

▼ **Value Frequency**

Valor	Count	%
Más de 8 horas	2649	26.55%
De 2 horas a 4 horas	2250	22.55%
De 1 hora a 2 horas	1754	17.58%
De 4 horas a 6 horas	1662	16.66%
De 6 horas a 8 horas	1303	13.06%
Menos de 1 hora	358	3.59%
-	2	0.02%

Dos valores en blanco

Pregunta_30

▼ **Value Frequency**

Valor	Count	%
0	7510	75.27%
1	2465	24.70%
2	3	0.03%

El valor de 2 representa una duplicidad de la respuesta, debido a que se encuentra en dos niveles o dos carreras al mismo tiempo, por lo que representa un problema de calidad de datos.

Lo mismo sucede en las preguntas: Pregunta_31, Pregunta_33, Pregunta_37, Pregunta_38, Pregunta_39, Pregunta_49, Pregunta_50, Pregunta_54 y Pregunta_55.

Base de datos profesores

Presentan inconvenientes de calidad de datos los siguientes campos:

Periodo

▼ **Value Frequency**

Valor	Count	%
MAYO 2020 - OCTUBRE 2020	590	91.33%
CI MAYO 2020 - OCTUBRE 2020	46	7.12%
CEF MAYO 2020 - OCTUBRE 2020	7	1.08%
MAYO 2020 - ABRIL 2021	3	0.46%

Los periodos de “mayo 2020 - abril 2021” pertenece a otro periodo de análisis, los otros periodos deben unificarse en una sola clase.

Pregunta_doc_3

▼ **Value Frequency**

Valor	Count	%
RIOBAMBA	565	87.46%
Guano	20	3.10%
AMBATO	9	1.39%
Quito	8	1.24%
Cuenca	5	0.77%
Guayaquil	3	0.46%
Loja	3	0.46%
BAÑOS DE AGUA SANTA	2	0.31%
Chambo	2	0.31%
Chimborazo	2	0.31%

Presenta datos con incorrecta escritura

Pregunta_doc_4

Valor	Count	%
Lizarzaburu	138	21.36%
Velasco	116	17.96%
Maldonado	94	14.55%
Veloz	93	14.40%
Lizarzaburo	35	5.42%
Juan de Velasco	12	1.86%
Velazco	10	1.55%
La Matriz	9	1.39%
RIOBAMBA	9	1.39%
Yaruquies	8	1.24%

Existen clases duplicadas por escritura incorrecta

Pregunta_doc_5

▼ **Value Frequency**

Valor	Count	%
RIOBAMBA	5	0.77%
La Primavera	4	0.62%
Las Acacias	3	0.46%
Colón 2938 y Venezuela	2	0.31%
Junín 34-14 y Francia	2	0.31%
Ciudadela Fausto Molina	2	0.31%
Cordova y valencia	2	0.31%
Conjunto San Antonio	2	0.31%
Espejo 15 - 06 y Boyacá	2	0.31%
GARCÍA MORENO Y NUEVA YORK	2	0.31%

Presenta datos con incorrecta escritura

Pregunta_doc_25

Valor	Count	%
ninguna	155	23.99%
Ninguna.	10	1.55%
Ninguno	9	1.39%
N/A	4	0.62%
LA VELOCIDAD DEL INTERNET	4	0.62%
NN	2	0.31%
internet lento	2	0.31%
Actividades del hogar	2	0.31%
Ninguno.	2	0.31%
.	2	0.31%