Aplicación de técnicas de análisis de datos para encontrar patrones espaciales no obvios dentro del proyecto Alianza CERES Salto-Afro, Norte del Cauca

Duvan Andres Giraldo Quiñonez

Trabajo de grado para optar el título de Ingeniero en Sistemas

Director

Beatriz Eugenia Marin Ospina

Magister en sistemas geográficos

Institución Universitaria Antonio José Camacho

Facultad de ingenierías

Ingeniería en sistemas

2021

Dedicatoria

La dedicación y entrega durante los últimos 5 años es resultado de las personas que me rodean. El esfuerzo realizado en este trabajo va dedicado primordialmente a mis padres, por el apoyo incondicional a mis estudios y creer siempre en las decisiones que he tomado y todos mis familiares que aportaron de una u otra manera a mi desarrollo a través de esta increíble etapa.

Agradecimiento.

Agradezco a los investigadores Beatriz Eugenia Marin, Tania Isadora Mora, Claudia Patricia Valencia y Edwin Jair Nuñez, que forman parte del grupo GrinTic del semillero ITmedia al cual pertenezco, los cuales ayudaron a recompilar la información suministrada de los diferentes estudiantes, egresados y competencias de los diferentes programas académicos, fueron un gran apoyo en los diferentes eventos en los cuales se presentaron avances de mi proyecto, me ayudaron a crecer como profesional y en el correcto desarrollo de mis actividades.

Debo agradecer de manera especial y sincera nuevamente a la profesora Beatriz Eugenia Marin, por aceptarme como estudiante semillerista para realizar este proyecto de grado bajo su dirección, gracias a su guía, disponibilidad y paciencia ayudo a realizar el correcto manejo del proyecto, artículos y ponencias elaboradas en el transcurso del mismo, facilitarme el material adecuado para instruirme y formarme en esta área de conocimiento para ser aplicado de la manera correcta.

Doy las gracias a la Institución Universitaria Antonio José Camacho, que me permitió desarrollarme como profesional, dándome oportunidades que nunca imagine y hacerme vivir experiencias que llevare en mi corazón toda la vida.

Y, evidentemente, el agradecimiento más profundo va para mi familia. Sin la ayuda de ellos, colaboración y esfuerzo no habría sido posible llevar a cabo este camino, A mis padres Carlina y Wilson, que son un ejemplo de superación y sacrificio, mi abuela Maruja la cual es una madre para mí, acompañándome en todas las etapas de mi vida, mi novia Carolay que me lleva acompañando en todos los momentos difíciles de la carrera y mi tío y amigo Luis Miguel, por ser un ejemplo de superación sin sus concejos no hubiera seria la persona que soy hoy.

Contenido

Pág.

Introducción....3
1. Aplicación De Técnicas De Análisis De Datos Para Encontrar Patrones Espaciales No Obvios Dentro Del Proyecto Alianza CERES Salto-Afro, Norte del Cauca....4
1.1 Planteamiento del Problema....4
1.2 Justificación....5
1.3 Objetivos....5
1.3.1 Objetivo General....5
1.3.2 Objetivos Específicos....5
2. Marco Referencial....6
2.1 Antecedentes....6
2.2 Marco Contextual....6
2.3 Marco Teórico....7
2.3.4 Metodología KDD (Knowledge Discovery in Databases)....7
2.3.4.1 Etapa de selección....8
2.3.4.2 Etapa de preprocesamiento....8
2.3.4.3 Etapa de transformación:....8
2.3.4.4 Etapa de minería de datos:....8
2.3.4.4.1 Aprendizaje supervisado....9
2.3.4.4.2 Aprendizaje no supervisado....9
2.3.4.5 Etapa de interpretación....9
2.3.5 Técnicas de minería de datos....10
2.3.5.1 Reglas de Asociación....10
2.3.5.2 Clasificación....10
2.3.5.2.1 Árbol de decisiones....10
2.3.5.2.2 k vecinos más cercanos:....11

2.3.5.3 Clustering....12
2.3.5.3.1 K-Means....12
2.3.6 Weka....13
2.3.6.1 ZeroR....14
2.3.6.2 PART....15
2.3.6.2 J48....17
2.3.6.3 Apriori....18
2.3.6.4 Selección de atributos....19
2.4 Marco Conceptual....20
2.5 Marco Legal....20
3. Método....20
3.1 Integración de la información....20
3.2 Preprocesamiento de la información....23
3.3 Exploración de la información....24
4. Resultados....32
5. Discusión....39
6. Conclusiones....40
Referencias....41

Lista de Tablas

Pág.

Tabla 1 *Proceso de asociación del algoritmo A priori*.....18
Tabla 2. *Comparación de los algoritmos ZeroR, PART y J48, con los atributos seleccionados por los algoritmos Ranker e InfoGainAttributeEval como clase la situación académica actual*..33
Tabla 3. *Comparación de los algoritmos ZeroR, PART y J48, con los atributos seleccionados por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase la situación académica actual*.....34
Tabla 4. *Comparación de los algoritmos ZeroR, PART y J48, con los atributos seleccionados por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase el tiempo desempleado.*35
Tabla 5. *Comparación de resultados obtenidos del algoritmo simpleKmeans con los atributos seleccionados por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase el tiempo desempleado*.....37
Tabla 6. *Resultado del algoritmo simpleKmeans utilizando 2 conjuntos para encontrar agrupaciónes en los datos*.....38

Lista de Figuras

Pág.

Figura 1 *Etapas del proceso KDD*.....7
Figura 2. *Gráfica de un árbol de* decisiones.....11
Figura 3. *Grafica del algoritmo K vecinos más cercanos*.....12
Figura 4 *Algoritmo K-means*.....13
Figura 5. *Menú explorer del software Weka*.....14
Figura 6. *Ejecución del algoritmo ZeroR con el conjunto de datos Iris.*.....15
Figura 7. *Ejecución del algoritmo PART en el conjunto de datos Iris.*.....16
Figura 8. *Ejecución del algoritmo J48 con el conjunto de datos Iris.*.....17
Figura 9 *Diagrama de barras de competencias del programa de licenciatura en pedagogía infantil*.....22
Figura 10 *Diagrama de barras de competencias del programa Tecnología en producción*.....23
Figura 11 *Errores encontrados en las encuestas.*.....24
Figura 12 *Municipio de residencia y la relación con el género de los encuestados.*.....24
Figura 13 *Etnia y rango de edad*.....25
Figura 14 *Información del vínculo familiar y estado civil.*.....26
Figura 15 *Información laboral.*.....28
Figura 16 *Información académica.*.....30
Figura 17 *Evaluación de competencias en el programa académico que cursó*.....31
Figura 18. *Resultado arrojado por el algoritmo PART en el conjunto de datos seleccionado por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase la situación académica actual*.....34
Figura 19. *Resultado arrojado por el algoritmo J48 en el conjunto de datos seleccionado por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase el tiempo desempleado*.....36
Figura 20. *Resultado arrojado por el algoritmo simpleKmeans en el conjunto de datos seleccionado por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase el tiempo desempleado*.....37

Resumen

El proyecto Alianza CERES Salto-Afro, Norte del Cauca, fue una convocatoria en los municipios de Padilla, Santander de Quilichao, Puerto Tejada y Guachené, el cual buscaba llevar mayor cobertura de educación superior en el norte del Cauca a través de los programas de licenciatura en pedagogía infantil, tecnología en producción industrial, tecnología en logística y licenciatura en ciencias del deporte y la educación física, actualmente el proyecto no tiene información que permita conocer el estado de sus estudiantes y egresados, ni el efecto que logro tener la convocatoria a nivel personal y municipal

De acuerdo a lo planteado anteriormente, se ha propuesto “Implementar de técnicas de análisis de datos para encontrar patrones espaciales no obvios dentro del proyecto Alianza CERES Salto-Afro, Norte del Cauca”, esto se realizara como un aporte al proyecto para favorecer futuras convocatorias, planes regionales y fortalecimiento de los programas académicos aquí mencionados.

Palabras clave: Análisis de datos, patrones, programas académicos, municipios.

Abstract

The Alianza CERES Salto-Afro project, Norte del Cauca, was a call in the municipalities of Padilla, Santander de Quilichao, Puerto Tejada and Guachené, which sought to bring greater coverage of higher education in the north of Cauca through the programs of degree in child pedagogy, technology in industrial production, technology in logistics and a degree in sports science and physical education, currently the project does not have information that allows knowing the status of its students and graduates, nor the effect that the call to personal and municipal level

According to the above, it has been proposed to "Implement data analysis techniques to find non-obvious spatial patterns within the CERES Salto-Afro Alliance project, Norte del Cauca", this will be carried out as a contribution to the project to favor future calls. , regional plans and strengthening of the academic programs mentioned here.

Keywords: Data analysis, patterns, academic programs, municipalities.

Introducción

En el año 2019, el proyecto Alianza CERES Salto-Afro, Norte del Cauca, finalizo su convocatoria, de la cual hasta el momento no se tienen datos claros que permitan validar la repercusión del programa en los diferentes municipios que participaron en el desarrollo de las actividades, ni de la situación actual de los egresados y estudiantes que ingresaron en los programas académicos licenciatura en pedagogía infantil, tecnología en producción industrial, tecnología en logística y licenciatura en ciencias del deporte y la educación física, .

A través de la información recompilada del proyecto por medio de encuestas, competencias de los programas académicos y demás información recolectada por el grupo de investigación se busca encontrar patrones y asociaciones utilizando técnicas de análisis de datos como lo son agrupación o clustering y clasificación, permitiendo aportar al proyecto interpretación de los datos resultantes en cada uno de los ciclos de vida de la metodología KDD la cual nos permite realizar el proceso de análisis de datos de manera iterativa y mejorar en cada una de sus etapas, relacionando en cada una de ellas como atributo principal el municipio de residencia de los estudiantes y egresados, lo cual asignara un contexto geográfico a todos los atributos e interpretaciones en el desarrollo de la implementación.

Los resultados obtenidos servirán como insumo para favorecer nuevas ampliaciones de cobertura, a su vez ayudar a fortalecer los programas académicos ofrecidos en la alianza generando mejores posibilidades para los estudiantes, egresados y los planes de desarrollo que tengan los diferentes municipios del norte del Cauca.

1. Aplicación De Técnicas De Análisis De Datos Para Encontrar Patrones Espaciales No Obvios Dentro Del Proyecto Alianza CERES Salto-Afro, Norte del Cauca.

1.1 Planteamiento del Problema

Existe gran cantidad de datos referentes al proyecto Alianza CERES Salto-Afro, Norte del Cauca que finalizó en el año 2019, donde intervinieron entidades públicas, instituciones educativas y organizaciones privadas, con el fin de brindar a la población del norte del Cauca facilidades de acceso a la educación superior en los cuales participaron los municipios de Guachené, Padilla, Puerto Tejada y Villa Rica. Los programas creados fueron licenciatura en pedagogía infantil, tecnología en producción industrial, tecnología en logística y licenciatura en ciencias del deporte y la educación física, se esperaba con esta oferta impactar en la economía de los municipios. Hasta el momento se encuentran egresados y estudiantes que participaron en el proyecto Alianza CERES Salto-Afro, Norte del Cauca, de los cuales no se tienen datos del impacto del proyecto en los municipios y en la vida de cada uno ellos.

Por estas razones en este proyecto se propone crear un grupo de investigación conformado por las docentes Claudia Patricia Valencia y Tania Isadora Mora, que exploran los contextos geográficos y competencias de los programas de licenciatura en pedagogía infantil y Tecnología en producción en los municipios de Guachené, Padilla, Santander de quilichao y Puerto tejada, buscando evaluar el impacto del proyecto Alianza CERES Salto-Afro, Norte del Cauca.

En el presente proyecto de grado se reúne gran parte de la información recolectada por los investigadores para explorar a través de técnicas de análisis de datos documentando posibles patrones o asociaciones en los datos que permitan vincular los intereses de los municipios, los programas académicos y los estudiantes para aportar como parte del macro proyecto Alianza CERES Salto-Afro, Norte del Cauca.

1.2 Justificación

Con este proyecto se busca encontrar información a través de la aplicación de técnicas de análisis de datos, que permita relacionar a los estudiantes y egresados del proyecto Alianza CERES Salto-Afro, Norte del Cauca, que favorezca a la realización de nuevas ampliaciones de coberturas y tomar acciones efectivas con la ayuda de un flujo de trabajo sistematizado que permita su réplica en los diferentes municipios.

De igual forma favorecer al proyecto Alianza CERES Salto-Afro, Norte del Cauca, entregando una retroalimentación para su mejora en futuros proyectos que se realicen y un fortalecimiento a los programas académicos ofrecidos por la Institución Universitaria Antonio José Camacho.

1.3 Objetivos

1.3.1 Objetivo General

Aplicar técnicas de análisis de datos en la información recolectada sobre el proyecto Alianza CERES Salto-Afro, Norte del Cauca para encontrar patrones espaciales interesantes.

1.3.2 Objetivos Específicos

- Seleccionar las variables de valor para el proyecto
- Procesar la información a partir de las técnicas de minería de datos que mejor se ajusten a las variables.
- Interpretar la información obtenida después de aplicar las técnicas de análisis de datos.
- Evaluar la precisión de los modelos resultantes.

2. Marco Referencial

2.1 Antecedentes

En el documento que presenta Dueñas (2009), da a conocer la aplicación de minería de datos en la inteligencia de negocios, aprovechando la gran cantidad de información que almacenan las empresas enfatizando en los datos espaciales los cuales en su mayoría de los casos no se utilizan efectivamente, ya que en estos se pueden encontrar soluciones a problemas específicos en empresas que contienen grandes cantidades de datos como lo es el marketing, la caracterización de usuarios, detección de fraudes, control de tráfico, logrando encontrar asociaciones entre datos espaciales y no espaciales logrando un mayor grado de competitividad en el mercado.

Gutiérrez & Molina (2015), presenta el uso de la minería de datos y su importancia para resolver problemas empresariales en específico para las micro y pequeñas empresas, explicando las técnicas de minería de datos más comunes y metodologías de extracción de conocimiento que se pueden utilizar para realizar planeaciones económicas, análisis de mercado, análisis de perfiles de clientes y tomas de decisiones para una mejor gestión y avance de la empresa para alcanzar sus objetivos dando una visibilidad global en un entorno local.

En el artículo de Menacho (2017), realizaron un proyecto para lograr determinar el desempeño de los alumnos matriculados en los cursos al inicio de un semestre, utilizando técnicas de minería de datos para predecir el rendimiento de los estudiantes, para lograr determinar los diferentes factores que influyen en el correcto aprendizaje de la materia, además, poder tomar acciones en la forma de enseñanza y mejora en los currículos académicos de cada una de las asignaturas, aumentando la aprobación de los estudiantes y ofreciendo educación superior con mayor calidad.

2.2 Marco Contextual

El proyecto Alianza CERES Salto-Afro, Norte del Cauca nace en el año 2012 a través de la convocatoria del Ministerio de Educación Nacional, contando con la participación de la Escuela Nacional del Deporte (END), las alcaldías de Guachené, Padilla, Puerto Tejada, Villa

Rica, la Gobernación del Cauca, la Fundación Propal, la Corporación Río Cauca Palenke, el Consejo Comunitario de las Comunidades Negras Pílamo Palenque y la institución Universitaria Antonio José Camacho liderando el proyecto, teniendo como objetivo brindar mayor cobertura a la educación superior, creándose los programas de Licenciatura en Pedagogía Infantil, Tecnología en Producción Industrial, Tecnología en Deporte, Técnico Profesional en Logística Empresarial y Técnico Profesional en Procesos Empresariales (López, 2018).

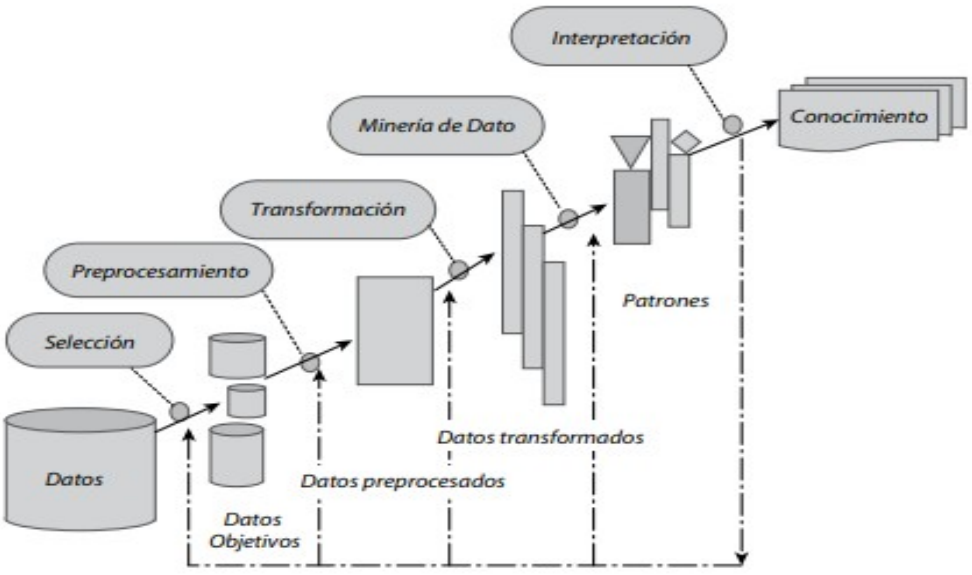
El convenio finalizó en el año 2018, continuando con el proceso en el que se encuentran hasta el momento egresados y estudiantes que participaron en la alianza, de los cuales no se tienen datos del impacto que generó el proyecto en los municipios y en la vida de cada uno ellos.

2.3 Marco Teórico

2.3.4 Metodología KDD (Knowledge Discovery in Databases)

Para lograr extraer conocimiento útil de los datos se lleva a cabo un ciclo de vida, uno de los más utilizados es el KDD (descubrimiento de conocimiento en bases de datos, por sus siglas en inglés), el cual consta de las siguientes etapas:

Figura 1 *Etapas del proceso KDD.*



Tomado de Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (Timarán Pereira et al., 2016),

2.3.4.1 Etapa de selección.

En esta etapa se definen los objetivos a alcanzar con el proyecto, al tener claridad en tu objetivo facilitará los siguientes pasos del proceso KDD. De igual manera en esta etapa se recopilan los datos y se digitalizan los que se encuentren de forma física para ser seleccionados.

2.3.4.2 Etapa de preprocesamiento.

La etapa de preprocesamiento se encarga de limpiar y ajustar los datos que puedan presentar errores o valores no deseados en las siguientes etapas, algunas de las irregularidades que se revisan en esta etapa pueden ser el remplazo valores desconocidos como lo son los valores nulos y valores que se encuentran fuera del rango esperado, este proceso se puede realizar aplicado técnicas estadísticas como lo pueden ser la media, moda, mínimo, máximo y en algunas ocasiones se podría optar por valores aleatorios o eliminando completamente la muestra.

2.3.4.3 Etapa de transformación:

En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos (Timarán Pereira et al., 2016).

Para lograr obtener los resultados esperados es necesario que los datos se encuentren en su estado más óptimo en cada una de sus características, ya que cada una de estas debe aportar al modelo una cercanía a nuestro dato objetivo, sería perjudicial para el modelo tener características que proporcionen un sesgo y aumenten la maldición de la dimensionalidad, por estas razones en la etapa de transformación se deben seleccionar y buscar las características que mejor representen nuestro objetivo.

2.3.4.4 Etapa de minería de datos:

Con la necesidad del ser humano de lograr interpretar grandes volúmenes de datos nacen algoritmos y modelos estadísticos, que facilitan la extracción de información útil, de la cual no se tenía anteriormente un conocimiento práctico y utilizarlo como beneficio de manera más

sencilla. A estas técnicas se le conoce como minería de datos, “el cual es un proceso asistido por computadora que excava y analiza enormes conjuntos de datos”(Raval, 2012), este conjunto de datos el cual es procesado por la minería de datos se le denomina Big data.

2.3.4.4.1 Aprendizaje supervisado

El aprendizaje por supervisión es uno de los dos conjuntos que envuelven la mayor parte de los algoritmos de minería de datos y de aprendizaje de máquina, esté consta en generalmente en entregar un modelo de datos ya establecido como entrenamiento al algoritmo que deseamos utilizar, este modelo ayudará a encontrar un punto de partida al algoritmo para lograr predecir o estimar algún resultado que nosotros deseamos. La mayor parte de las técnicas de minería de datos como lo son regresiones lineales y clasificación trabajan sobre esta lógica de entrenamiento de los algoritmos, como ejemplo podríamos describir un clasificador de frutas, al cual se le entrena con imágenes de diferentes tipos de frutas y su respectivo nombre, este algoritmo irá empezando a comprender por las imágenes que corresponde a la fruta asignada, así cuando se desee identificar, podría determinar a qué clase corresponde el objeto que le hemos dado en observación.

2.3.4.4.2 Aprendizaje no supervisado

Los algoritmos de aprendizaje no supervisado no contienen un modelo de entrenamiento inicial, estos tratan de encontrar agrupaciones entre las similitudes en los datos, este algoritmos es ideal para analizar conjuntos de datos que no se logren etiquetar o su labor de etiquetado sea muy ardua, como pueden ser la segmentación de usuarios en plataformas, entre las técnicas más comunes se puede encontrar la agrupación en clústeres, el cual se encarga en agrupar los elementos más similares en diferentes conjuntos.

2.3.4.5 Etapa de interpretación.

En esta etapa se analizan los patrones reflejados por las técnicas de minería de datos y si es necesario se retoma nuevamente a cualquiera de las etapas del ciclo de vida, se consideran cambios y mejoras a realizar al sistema, como la eliminación o uso de nuevas herramientas para la mejora de los resultados, de igual manera esta etapa significa la culminación del proceso de

minería de datos, obteniendo los resultados esperados y traducir los datos a una manera fácil de comprender por la parte interesada.

2.3.5 Técnicas de minería de datos.

2.3.5.1 Reglas de Asociación.

Las reglas de asociación es una técnica de aprendizaje supervisado que permite encontrar situaciones en común que comparten las variables en un conjunto de datos para extraer patrones frecuentes, asociaciones y correlaciones interesantes, un ejemplo común es la relación de los productos comprados en un supermercado, como podría ser {leche, harina}$\Rightarrow$ {huevos}, esto podría identificar que si un usuario compra leche y harina, existe la posibilidad que compre huevos. Esto ayudaría a las tiendas a realizar marketing y estrategias de ventas mucho más efectivas, este tipo de técnicas también son utilizadas frecuentemente en tiendas virtuales, para lograr calcular productos relacionados y así realizar recomendaciones de compras antes de finalizar la pasarela de pago, un algoritmo de reglas de asociación común son el a priori.

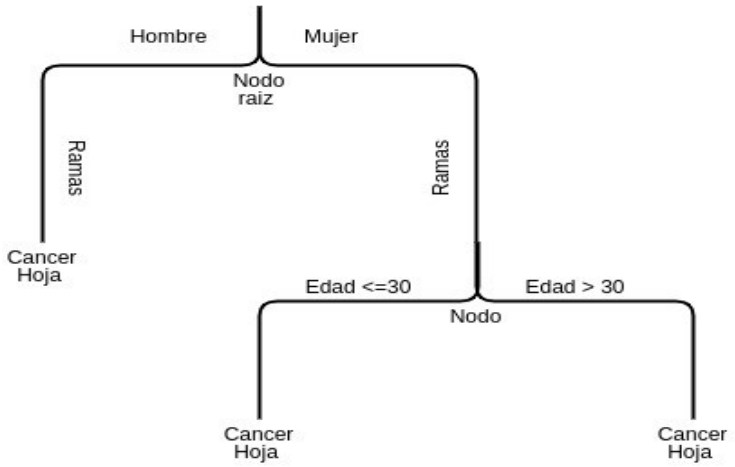
2.3.5.2 Clasificación.

Esta técnica es una de las más conocidas en machine learning y minería de datos de aprendizaje supervisado, se basa en enseñar a los algoritmos a predecir la categoría de un objeto o un atributo faltante, esta técnica utiliza dos pasos para determinar la clasificación. El primer paso se basa en el entrenamiento del algoritmo, en el cual se crea un conjunto de datos previamente clasificados, que se utilizará para asignar parámetros iniciales, este paso también es conocido como aprendizaje supervisado. El segundo paso sirve para predicción de los datos sin clasificar o datos futuros, algunas de las técnicas de clasificación más comunes son:

2.3.5.2.1 Árbol de decisiones.

Esta técnica de aprendizaje supervisado se basa en dividir las características que aparecen en los datos para lograr predecir el resultado, estas divisiones generan un mapa en forma de árbol o diagrama de flujo, lo cual facilita su interpretación como pueden ver en la Figura 2. En esta estructura se inicia en un nodo raíz del cual saldrán las ramificaciones que representa una regla de decisión, llegando a un nuevo nodo interno el cual tendrá una nueva característica o algún nodo hoja el cual será una de las clases que deseamos predecir.

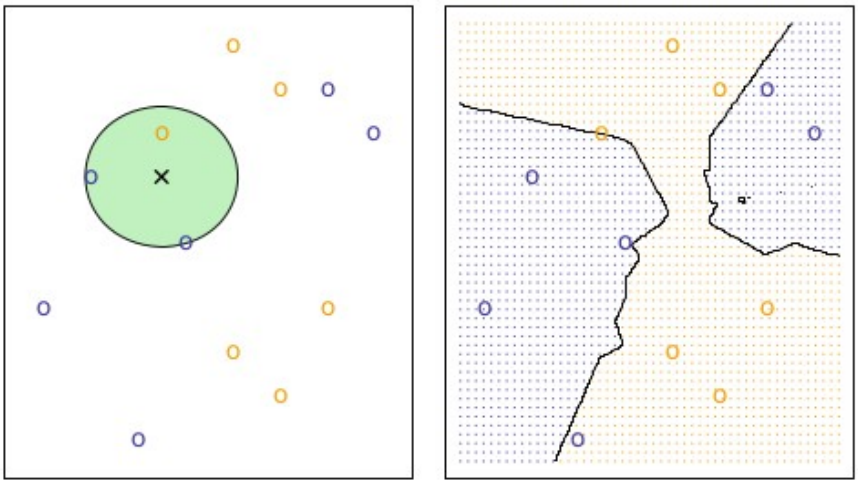
Figura 2. *Gráfica de un árbol de* decisiones



2.3.5.2.2 k vecinos más cercanos:

Este algoritmo de aprendizaje supervisado trabaja con un conjunto de datos de entrenamiento los cuales ya se encuentran clasificados, los cuales servirán para predecir los nuevos elementos utilizando los k vecinos más cercanos a este, utilizando una regla muy sencilla de encontrar la clase más frecuente entre los K elementos que rodean al elemento y asignando la clase que representa, para evitar algún empate en el cálculo se utiliza un valor impar de K, utilizando esta técnica podemos encontrar que “la tasa de error es pequeña en la práctica. En teoría, se sabe que la tasa de error asintótica a medida que el número de muestras de prototipos se vuelve muy grande se acerca a la tasa de error óptima de Bayes y en realidad tiende a ella cuando k aumenta”(Laaksonen & Oja, 1996), en la figura 3 podemos ver de manera gráfica el funcionamiento del algoritmo.

Figura 3. *Grafica del algoritmo K vecinos más cercanos*



Tomado de An Introduction to Statistical Learning (James et al., 2013)

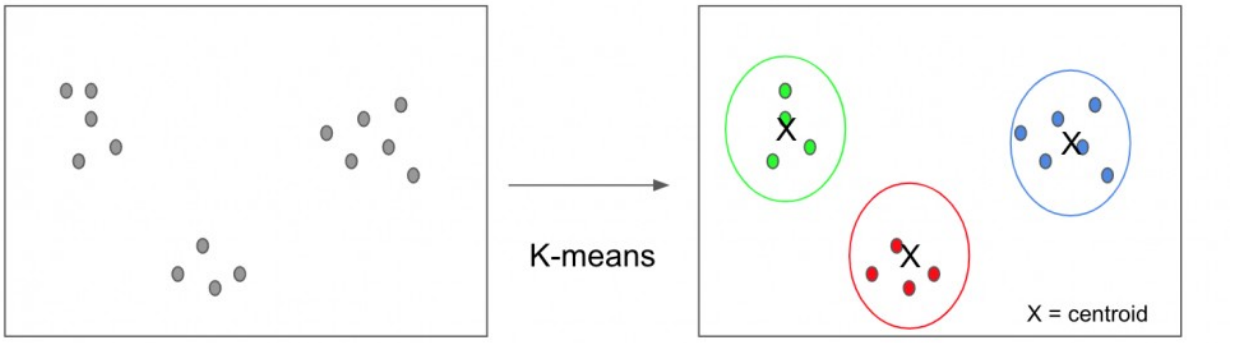
2.3.5.3 Clustering.

La agrupación en clústeres se refiere a la división de datos en grupos de objetos similares, esta técnica se basa en aprendizaje no supervisado en el cual cada grupo consta de objetos que son similares entre sí y diferentes a los objetos de otros grupos (Berkhin, 2006), la técnica de clustering son algo similares a la técnica clasificación, pero esta no necesita datos de entrenamiento, por lo cual la hace no supervisada, esto hace que no existan unas clases definidas con anterioridad, únicamente se agrupara por la similitud entre los objetos que se encuentran en los datos, se podría ver aplicada esta técnica en la segmentación del mercado, para lograr crear diferentes grupos de clientes y así ofrecer diferentes servicios como podría ser préstamos, tasas de intereses preferenciales o descuentos, uno de sus algoritmos más famosos es el K-means.

2.3.5.3.1 K-Means

Es uno de los algoritmos de clustering más utilizados para la agrupación en clúster, el cual se basa en recibir como parámetro de entrada los elementos que se desean agrupar en las dimensiones que se encuentren y el número de K agrupaciones que desean realizar, este algoritmo intentara minimizar la suma de los cuadrados entre cada elemento en el conjunto de datos, es decir trata de encontrar los elementos menos dispersos (Hartigan & Wong, 1979), para ser agrupados en uno de los K conjuntos y obteniendo cada uno de los cancroides los cuales son las coordenadas de cada uno de los grupos.

Figura 4 *Algoritmo K-means*

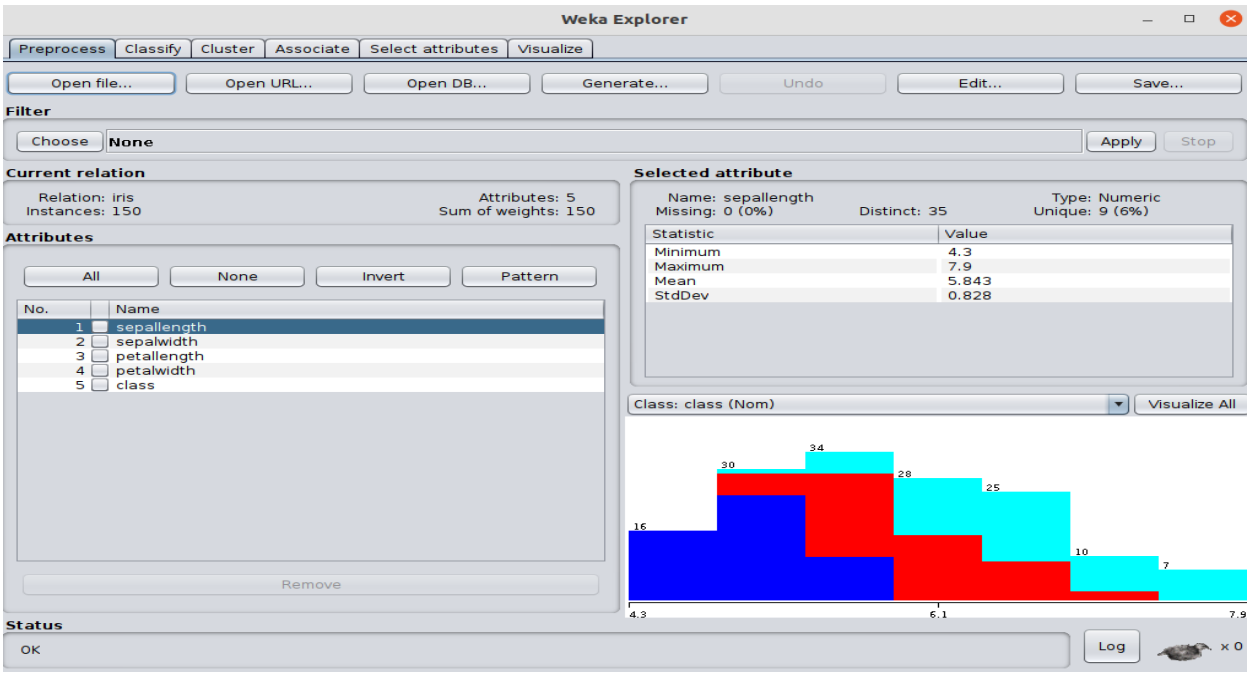


Tomado de K-means Clustering en Machine Learning (K-medias). (2020, mayo 2). DATA SCIENCE. https://datascience.eu/es/aprendizaje-automatico/k-means-clustering-en-machine-learning/

2.3.6 Weka

Es una plataforma para la aplicación de minería de datos el cual contiene múltiples herramientas para el preprocesamiento de los datos, permitiendo la subida de múltiples formatos como archivos, bases de datos o la posibilidad de generar tus propios datos de prueba, de igual forma la posibilidad de visualizar los datos, atributos y algunos datos estadísticos de este como lo es el valor máximo y mínimo, media y desviación estándar como ven en la figura 5, Weka cuenta con una gran variedad de algoritmos incorporados, para resolver los principales problemas de minería de datos como lo son: regresión, clasificación, agrupación, reglas de asociación y selección de atributos, contando también con una gran comunidad que mantienen el repositorio del cual se pueden descargar miles de algoritmos e instalarlo en tu plataforma, algunos de los algoritmos que aparecen pre instalados en Weka son:

Figura 5. *Menú explorer del software Weka.*



Tomado de Data Mining: Practical Machine Learning Tools and Techniques, (Frank et al., 2016)

2.3.6.1 ZeroR

Es uno de los algoritmos de clasificación más simples, este únicamente predice la clase o categoría principal, si en los datos no se encuentra una clase predominante clasifica en la primera que encuentre, principalmente se basa en realizar reglas de asociación, aunque este algoritmo no presente un poder predictivo, puede servir como valor de partida para las siguientes ejecuciones de los diferentes modelos generados, este tipo de algoritmos permite datos de tipo binarios, fechas, nominales y numéricos.

En la figura 6 se visualiza la ejecución del algoritmo ZeroR con el conjunto de datos iris, este contiene 150 instancias de las cuales se obtienen 5 atributos los cuales pueden visualizar en la salida de Weka como el largo del pétalo (petal length), ancho del pétalo (petal width), largo del sépalo (sepal length), ancho del sépalo (sepal width) y las clases que cuenta con 3 de tipos, iris-setosa, iris-versicolor e iris-virginica, en los resultados se visualizan las instancias que clasificó correctamente que fueron el 33.33% de las 150 existentes, pero si se observa la matriz de confusión que genera weka, se puede observar que todas las instancias fueron agrupadas en

una sola clase, la cual fue Iris-setosa esto ocurre al no encontrar una clase predominante en los datos.

Figura 6. *Ejecución del algoritmo ZeroR con el conjunto de datos Iris.*

```
=== Classifier model (full training set) ===

ZeroR predicts class value: Iris-setosa

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          50               33.3333 %
Incorrectly Classified Instances       100               66.6667 %
Kappa statistic                          0
Mean absolute error                      0.4444
Root mean squared error                  0.4714
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    1,000    0,333      1,000    0,500      ?        0,500     0,333     Iris-setosa
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,333     Iris-versicolor
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,333     Iris-virginica
Weighted Avg.    0,333    0,333    ?          0,333    ?          ?        0,500     0,333

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
 50  0  0 |  b = Iris-versicolor
 50  0  0 |  c = Iris-virginica
```

Tomado de Data Mining: Practical Machine Learning Tools and Techniques, (Frank et al., 2016)

2.3.6.2 PART

Es la evolución del algoritmo C4.5 el cual crea árboles de decisiones parciales para obtener las reglas de asociación, este algoritmo se diferencia del C4.5 por su mayor rapidez de ejecución en los conjuntos de datos basándose en el procedimiento de dividir y vencer, este tipo de algoritmos permite datos de tipo binario y nominal para la clase que se desea predecir como atributos permite además de los anteriormente mencionados, tipos numéricos.

En la Figura 7, se encuentra la salida generada por la aplicación Weka del algoritmo PART con el conjunto de datos iris, encontramos que clasificó correctamente el 94% de las

instancias con respecto al resultado mostrado por el algoritmo ZeroR, se encuentra una mejora significativa a los 33.33% obtenidos, algo a resaltar es la lista de decisiones que elaboró PART para lograr clasificar las instancias del conjunto de datos, en la matriz de confusión se logra observar como clasifico cada una de las clases, una interpretación que se puede obtener es que existe una confusión entre la clase Iris-versicolor e Iris-virginica las cuales comparten características similares entre sus atributos.

Figura 7. *Ejecución del algoritmo PART en el conjunto de datos Iris.*

```
=== Classifier model (full training set) ===

PART decision list
------------------

petalwidth <= 0.6: Iris-setosa (50.0)

petalwidth <= 1.7 AND
petallength <= 4.9: Iris-versicolor (48.0/1.0)

: Iris-virginica (52.0/3.0)

Number of Rules  :      3


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         141               94      %
Incorrectly Classified Instances         9                6      %
Kappa statistic                          0.91
Mean absolute error                      0.0482
Root mean squared error                  0.1794
Relative absolute error                 10.8379 %
Root relative squared error             38.0567 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,980    0,000    1,000      0,980    0,990      0,985    0,990     0,987     Iris-setosa
                 0,940    0,060    0,887      0,940    0,913      0,868    0,954     0,878     Iris-versicolor
                 0,900    0,030    0,938      0,900    0,918      0,879    0,959     0,914     Iris-virginica
Weighted Avg.    0,940    0,030    0,941      0,940    0,940      0,911    0,968     0,926

=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  5 45 |  c = Iris-virginica
```

Tomado de Data Mining: Practical Machine Learning Tools and Techniques, (Frank et al., 2016)

2.3.6.2 J48

Este algoritmo es una adaptación del C4.5 realizada por Weka, que permite predecir los resultados, gracias a un entrenamiento previo con los datos el algoritmo logra construir un árbol de decisiones dividiendo cada característica en los nodos, este tipo de algoritmos permite datos de tipo binario y nominal para la clase que se desea predecir.

En la Figura 8, se encuentra la ejecución del algoritmo J48 con el conjunto de datos Iris, se encuentra una mejora con respecto a los algoritmos anteriores con un 96% de las instancias clasificadas correctamente, de igual forma nos enseña el árbol que genera para clasificar las entidades el cual consta de 5 hojas.

Figura 8. *Ejecución del algoritmo J48 con el conjunto de datos Iris.*

```
J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,980    0,000    1,000      0,980    0,990      0,985    0,990     0,987     Iris-setosa
                 0,940    0,030    0,940      0,940    0,940      0,910    0,952     0,880     Iris-versicolor
                 0,960    0,030    0,941      0,960    0,950      0,925    0,961     0,905     Iris-virginica
Weighted Avg.    0,960    0,020    0,960      0,960    0,960      0,940    0,968     0,924

=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

2.3.6.3 Apriori

Utiliza múltiples interacciones a la base de datos para encontrar las asociaciones, este algoritmo utiliza dos procesos primero se genera un conjunto de datos denominado candidatos los cuales son seleccionados tras el conteo de cada uno de los elementos de la base de datos para luego ser filtrados teniendo en cuenta la frecuencia en la cual aparece como se muestra en la Tabla 1, obteniendo así los elementos más frecuentes este filtro se basa en la propiedad apriori que dice que “todos los conjuntos de elementos secundarios (k-1) de los conjuntos de elementos k frecuentes deben ser frecuentes”(Zhao & Bhowmick, 2003), es decir que si uno de los

conjuntos de elementos no es frecuentes sus siguientes súper conjuntos tampoco lo serán, esta frecuencia se mide con el cálculo del soporte, el cual es el conteo del elemento sobre la cantidad de lista de elementos, esto nos ayudará a definir un soporte mínimo para considerar qué elementos no son frecuentes en la lista. Luego se encuentran nuevos candidatos uniendo los elementos más frecuentes obtenidos y repitiendo el filtro con un elemento nuevo como vemos en la Tabla 1, este tipo de algoritmo no admite valores de tipo numéricos, únicamente acepta datos de tipo binarios y nominales.

Tabla 1 *Proceso de asociación del algoritmo A priori*

Lista de elementos
I_1, I_2, I_5
I_2, I_4
I_1, I_2
I_2, I_3
I_1, I_3
I_1, I_2, I_3, I_5
I_2, I_3
I_1, I_2, I_3
I_1, I_2, I_5, I_6

1 elemento	Conteo
I_1	6
I_2	8
I_3	5
I_4	1
I_5	3
I_6	1

Elementos seleccionados
I_1
I_2
I_3
I_5

2 elementos	Conteo
I_1, I_2	6
I_1, I_3	3
I_1, I_5	3
I_2, I_3	4
I_2, I_5	3
I_3, I_5	1

Elementos seleccionados
I_1, I_2
$I_{1,} I_3$
$I_{1,} I_5$
I_2, I_3
I_2, I_5

3 Elementos	Conteo
I_1, I_2, I_5	3
I_1, I_2, I_3	2

2.3.6.4 Selección de atributos

En Weka se encuentra una sección en el menú dedicada a este apartado denominada selección de atributos, esta opción permitirá seleccionar las características adecuadas para nuestro modelo de entrenamiento y no llenar de datos innecesarios o redundantes. Weka implementa dos métodos para realizar la selección de atributos, el primero es el evaluador de atributos el cual se encarga de valorar cada uno de los atributos, realizando diferentes

combinaciones en el conjunto de datos, el segundo es el método de búsqueda, este ayudara al algoritmo a encontrar de forma más eficiente cada uno de los conjuntos que pueda generarse en la información, algunos evaluadores de atributos requieren un método de búsqueda específico, por este motivo se generan combinaciones entre las más comunes se encuentran:

Método de búsqueda=Ranker
Método de evaluación= InfoGainAttributeEval

Ranker según Weka es un método el cual clasifica los atributos por sus evaluaciones individuales ordenándolos en una lista de mayor a menor importancia, este algoritmo se utiliza con evaluadores de atributos como lo es infoGainAttributeEval, este evaluador asigna un valor a cada característica encontrada en los datos dependiendo del beneficio que aporte a la clase seleccionada.

Método de búsqueda= Greedy Stepwise
Método de evaluación= ClassifierSubsetEval

El método Greedy Stepwise realiza una búsqueda voraz entre los atributos según Weka, puede comenzar sin/todos los atributos o desde un punto arbitrario en el espacio. Se detiene cuando la adición/eliminación de cualquier atributo restante da como resultado una disminución en la evaluación, este método de busca se utiliza método de evaluación wrapper, los cuales utilizan un modelo de clasificación para puntuar los subconjuntos de cada una de las características.

2.4 Marco Conceptual

Patrones: Se refiere a la forma en la cual la minería de datos logra descubrir conocimiento interesante a partir de grandes cantidades de datos (Han et al., 2011).

Contenido curricular: Es el plan o la planificación, por la cual se organizan los procesos escolares de enseñanza/aprendizaje (Angulo & Blanco, 1994).

Competencias: La competencia supone la capacidad de afrontar demandas complejas, en un contexto determinado, poniendo en relación y movilizando prerrequisitos psicosociales que incluyen aspectos tanto cognitivos como no cognitivos (Rychen & Salganik, 2006).

2.5 Marco Legal

Por motivos de el correcto uso, manejo y tratamiento de datos recolectado en las encuestas realizadas a los estudiantes, egresados del proyecto Alianza CERES Salto-Afro, Norte del Cauca, cuyo manejo se encuentra regulado por la Ley 1581 de 2012 y el Decreto 1377 de 2013, se ha realizado un acuerdo de confidencialidad manejo y protección de datos el cual hace constar que, se hace responsable de mantener y conservar de manera confidencial todas las informaciones personales, comerciales, contables, técnicas, académicas o de cualquier otro tipo suministradas en la ejecución y ejercicio de sus funciones y/o proyecto o proceso académico, administrativo o investigativo específico, proteger la información ya sea de manera verbal, escrita visual, tangible, intangible digital o cualquier otro medio, de igual forma responder disciplinaria y penalmente por el mal uso o destinación inadecuada que le dé a la información confidencial y/o bases de datos suministradas.

3. Método

3.1 Integración de la información.

El proyecto Alianza CERES Salto-Afro, Norte del Cauca se encuentran datos de estudiantes, egresados, programas académicos e industria, los cuales deben ser compilados para tener una vista unificada de los datos dispersos, para lograr encontrar inconsistencias, redundancia en los datos y lograr extraer la información de relevancia para el proyecto, en el grupo de investigación se ha realizado la recolección de la información a través de encuestas las cuales han sido integradas y consolidadas en un mismo documento para obtener un informe claro de la cantidad de datos ingresados.

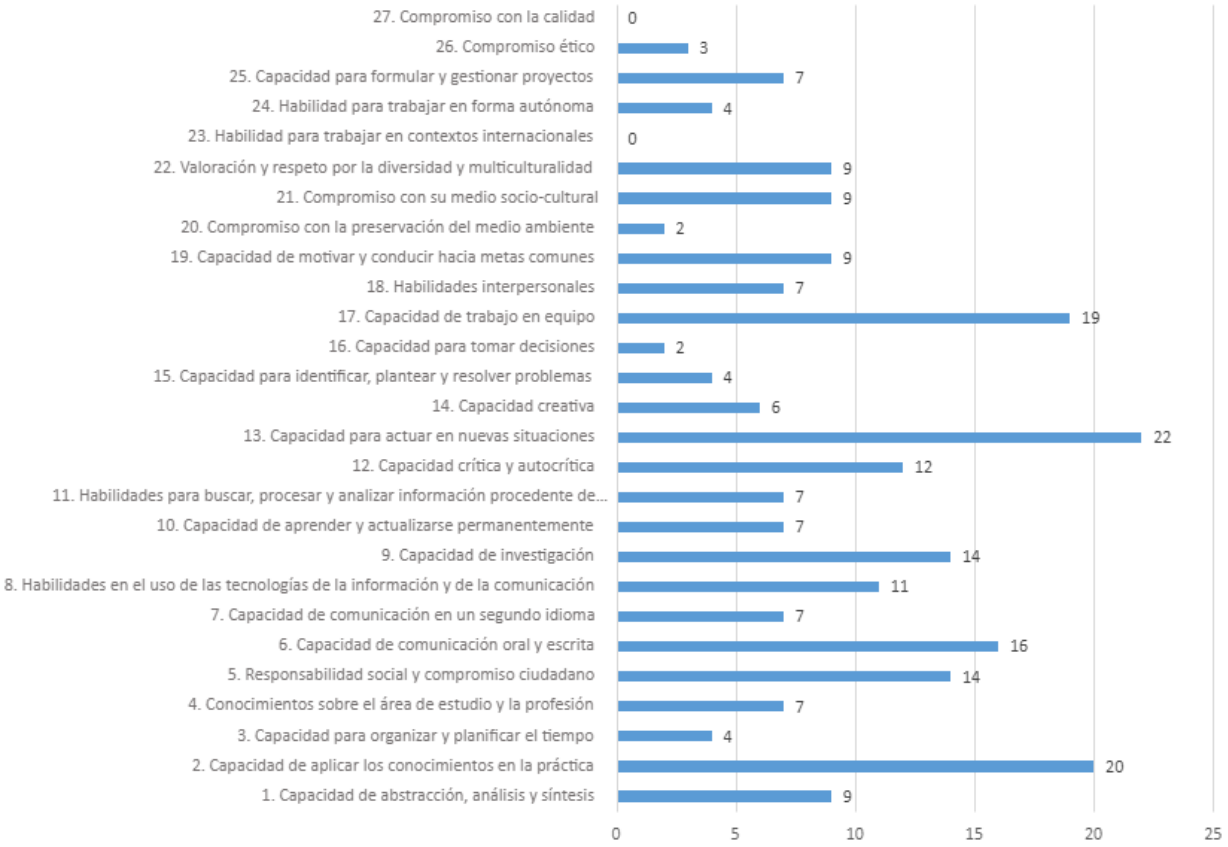
Anteriormente esta información iba ser recolectada realizando una investigación de campo, asistiendo a cada uno de estos municipios y con la ayuda de la alcaldía realizar la recolección de la información de estudiantes y egresados que participaron en la alianza, pero con la actual pandemia provocada por el Covid-19, hubo dificultad en la movilidad por las constantes cuarentenas que el gobierno implementó en los diferentes territorios del país.

Para obtener la perspectiva académica de cada uno de los estudiantes se les pidió evaluar del 1 al 5 siendo 1 insuficiente, 2, deficiente, 3 aceptable, 4 bueno y 5 excelente, las competencias del programa las cuales fueron. Trabajo en equipo, liderazgo, emprendimiento y manejo de conflictos relaciones interpersonales y comunicación, para lograr establecer un estándar entre los datos obtenidos de las competencias del programa y las competencias del mercado, se utilizaron las competencias Tuning Latinoamérica, de las cuales se fijó una o más relaciones a las evaluadas en la encuesta y se asignó una calificación aprobatoria haciendo referencia en el caso de los estudiantes a la correcta obtención de la competencia siendo entre 1 y 3 competencias no desarrolladas y adquiridas en el programa, entre 4 y 5 competencias desarrolladas y adquiridas en el programa.

En la encuesta se evaluó el trabajo en equipo el cual tiene una relación directa con la competencia Tuning Latinoamerica #17 capacidad de trabajo en equipo, liderazgo se asignaron 3 competencias Tuning Latinoamerica que se consideraron a fines las cuales son, #13 capacidad para actuar en nuevas situaciones, #16 capacidad para tomar decisiones y #15 capacidad para formular y gestionar proyectos, emprendimiento se relacionó con las competencias #14 capacidad creativa, #16 capacidad para tomar decisiones y #19 capacidad de motivar y conducir hacia metas comunes, en manejo de conflictos y relaciones interpersonales y comunicación se identificaron las mismas 3 competencias para cada una de ellas las cuáles fueron, #6 capacidad de comunicación oral y escrita, #17 capacidad de trabajo en equipo y #18 habilidades interpersonales, teniendo en cuenta estas competencias se lograra encontrar información de los datos de la perspectiva que tienen los estudiantes con respecto al programa académico y sus habilidades adquiridas.

Por otra parte los investigadores identificaron y evaluaron las competencias que el programa deseaba que los estudiantes adquirieran, tomando como base el micro currículo de los programas de tecnología en producción y licenciatura en pedagogía infantil, realizaron diagramas en los cuales se logran identificar que las competencias del programa de licenciatura en pedagogía infantil con más énfasis son la capacidad para actuar en nuevas situaciones, capacidad de aplicar los conocimientos en la práctica, capacidad de trabajo en equipo, capacidad de investigación y responsabilidad social y compromiso ciudadano como se muestra en la Figura 9.

Figura 9 *Diagrama de barras de competencias del programa de licenciatura en pedagogía infantil*



Nota. Diagrama realizado por las investigadoras Tania Isadora Mora y Claudia Patricia Valencia.

Para el programa de tecnología en producción se obtuvo que las competencias más notorias son la capacidad de aplicar los conocimientos en la práctica, capacidad de abstracción, análisis y síntesis, conocimiento sobre el área de estudio y la profesión, capacidad para identificar, plantear y resolver problemas, capacidad para identificar, plantear y resolver problemas y capacidad de trabajo en equipo como se muestra en la figura 10.

Figura 10 *Diagrama de barras de competencias del programa Tecnología en producción*

Nota. Diagrama realizado por la investigadora Beatriz Eugenia Marin.

3.2 Preprocesamiento de la información.

En esta etapa se encarga de modificar o eliminar los datos que puedan generar problemas en la futura ejecución del algoritmo, algunas características tuvieron que ser normalizadas como en el caso de los municipios, los cuales en ocasiones fueron escritos con errores de sintaxis, ortografía o el cambio de cotejamiento que genera problemas con las tildes y ñ que puedan encontrarse en los datos como se muestra en la figura 11, como Guachené, que se encontraba de cuatro diferentes maneras Guachene, GuachenÃ©, GUACHENE y guachene, además la información de los estudios como lo es la asignatura que ha servido o le servirá para trabajar, se realizó el cambio de las escalas numéricas que se empleaban en las encuestas a nominales que correspondieran al peso de cada uno de los valores asignados, del mismo modo se eliminaron las personas que realizaron la encuesta y no eran participante del proyecto, esto para mantener la integridad de los datos la cual se encuentra enfocada a este mismo.

En términos generales se adecuo los datos que puedan tener errores para evitar futuras inconsistencias, normalizando la base de datos, corrigiendo campos vacíos o faltantes, esta etapa es de relevancia en todos los procesos de minería de datos ya que ayuda a realizar un correcto descubrimiento de conocimiento y obtener una respuesta confiable de cada uno de los algoritmos.

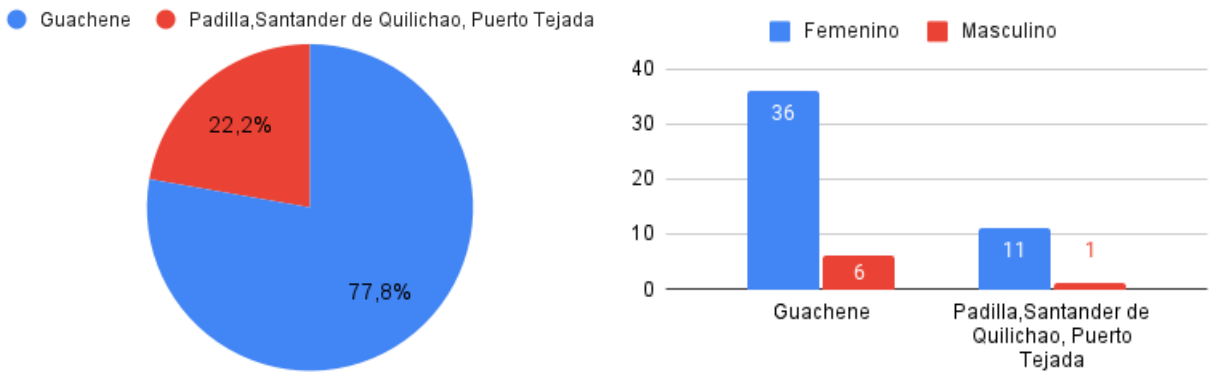
Figura 11 *Errores encontrados en las encuestas.*

1. Municipio de residencia	27. Â¿CuÃ¡l fuÃ© la asignatura que usted cree la ha servido o le servirÃ¡ para trabajar?	28. Trabajo en equipo:
Guachene	Principios administrativos	5
Guachene	Desarrollo de proyectos	4
Guachene	Muchas	4
GUACHENE	Ã‰tica del educador	4
guachene	todas	5
Guachene	PrÃ¡ctica pedagÃ³gica	5
GUACHENE	legislaciÃ³n	5
GuachenÃ©	Practica pedagogica	5
guachene	todas	5
guachenÃ¨ cauca	procesos productivos	4

3.3 Exploración de la información.

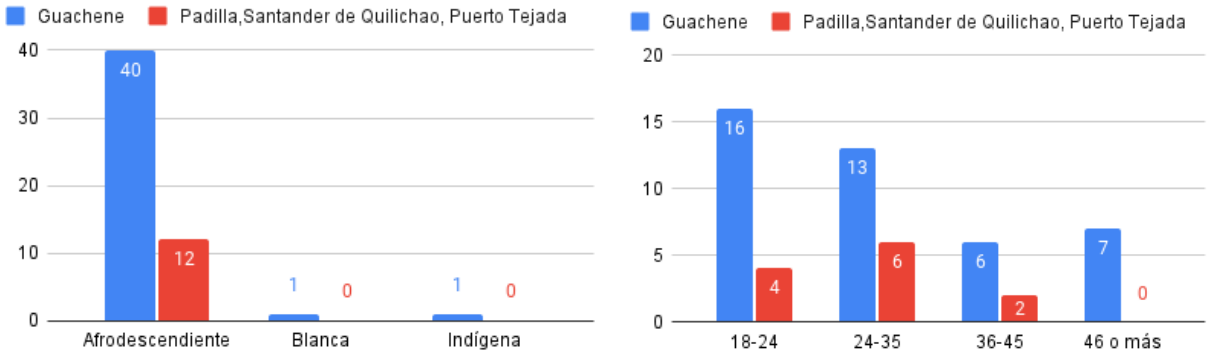
Los datos obtenidos de los 54 encuestas realizadas se evidencia que el 77,8% se encuentran en el municipio de Guachené, continuando Puerto tejada con un 16,7%, padilla con un 3,7% y de ultimo Santander de Quilichao con un 1,9%, por el bajo porcentaje de encuestados los municipios de Padilla, Santander de Quilichao y Puerto tejada se decidieron agruparlos en una sola categoría, como se muestra en la figura 12, en la cual se encuentra la relación de género en cada uno de los municipios, siendo el género femenino el más predominante con 87% aproximadamente y el masculino una minoría con un 13% aproximandamente.

Figura 12 *Municipio de residencia y la relación con el género de los encuestados.*



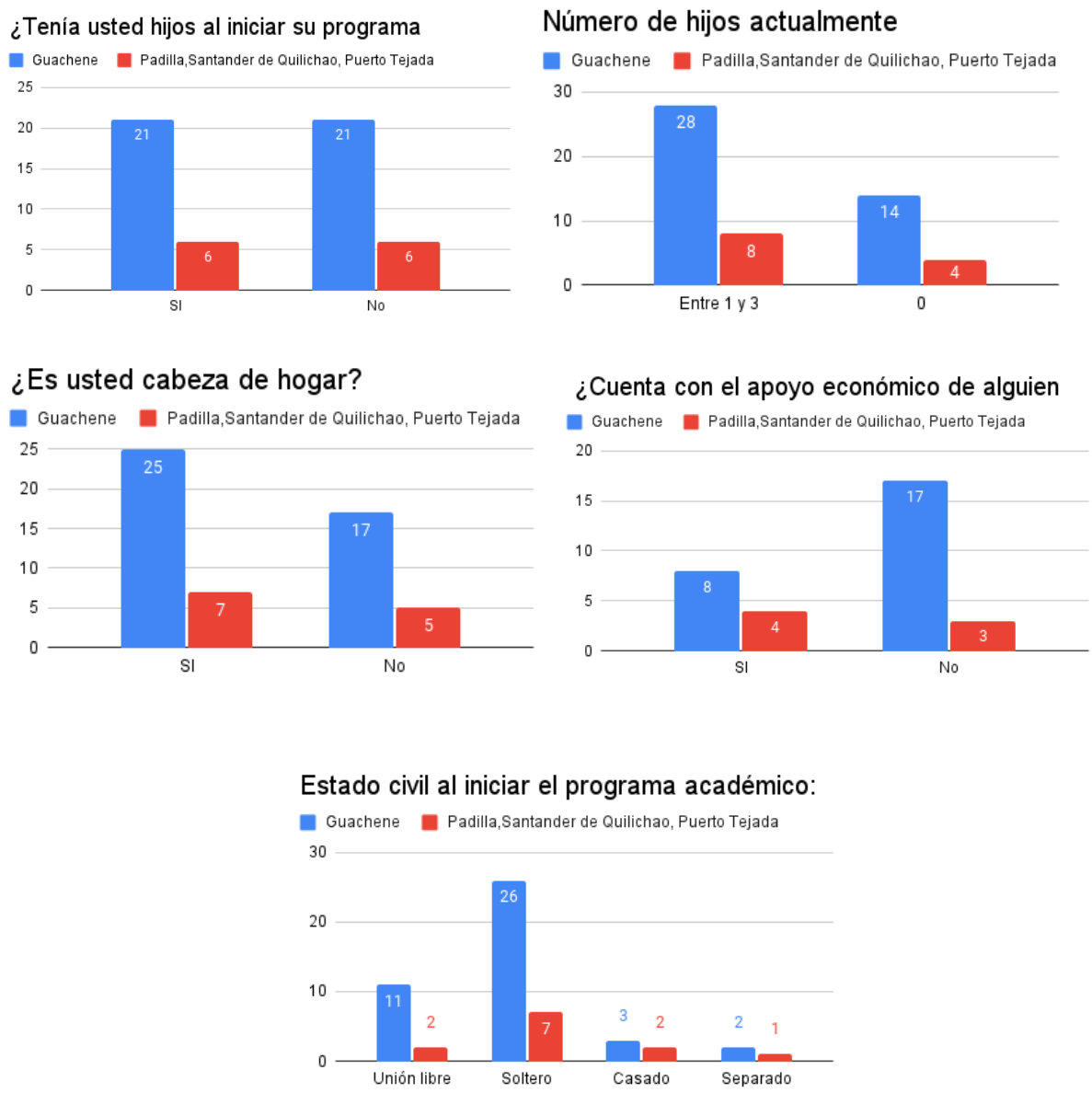
En las personas encuestadas se tomaron datos étnicos de los cuales se encontraron que la mayoría son afrodescendientes con 52 personas de las 54 encuestadas equivalentes a un 96,2%. Según el rango de edades los más representativos son 18-24 y 24-35 con 37% y 35,2% respectivamente, en la Figura 13 se puede ver la distribución con respectó a los municipios encuestados.

Figura 13 *Etnia y rango de edad.*



Para el proyecto es importante conocer el estado que se encontraba el estudiante antes y durante de su desarrollo como profesional o academico respecto a su situación familiar y estado civil, esta información se obtuvo realizando las preguntas que se encuentran en la Figura 14 obteniendo como resultado que el 59,3% de los encuestados son cabezas de hogar y de estos el 62,5% no cuenta con el apoyo economico de alguien más,.

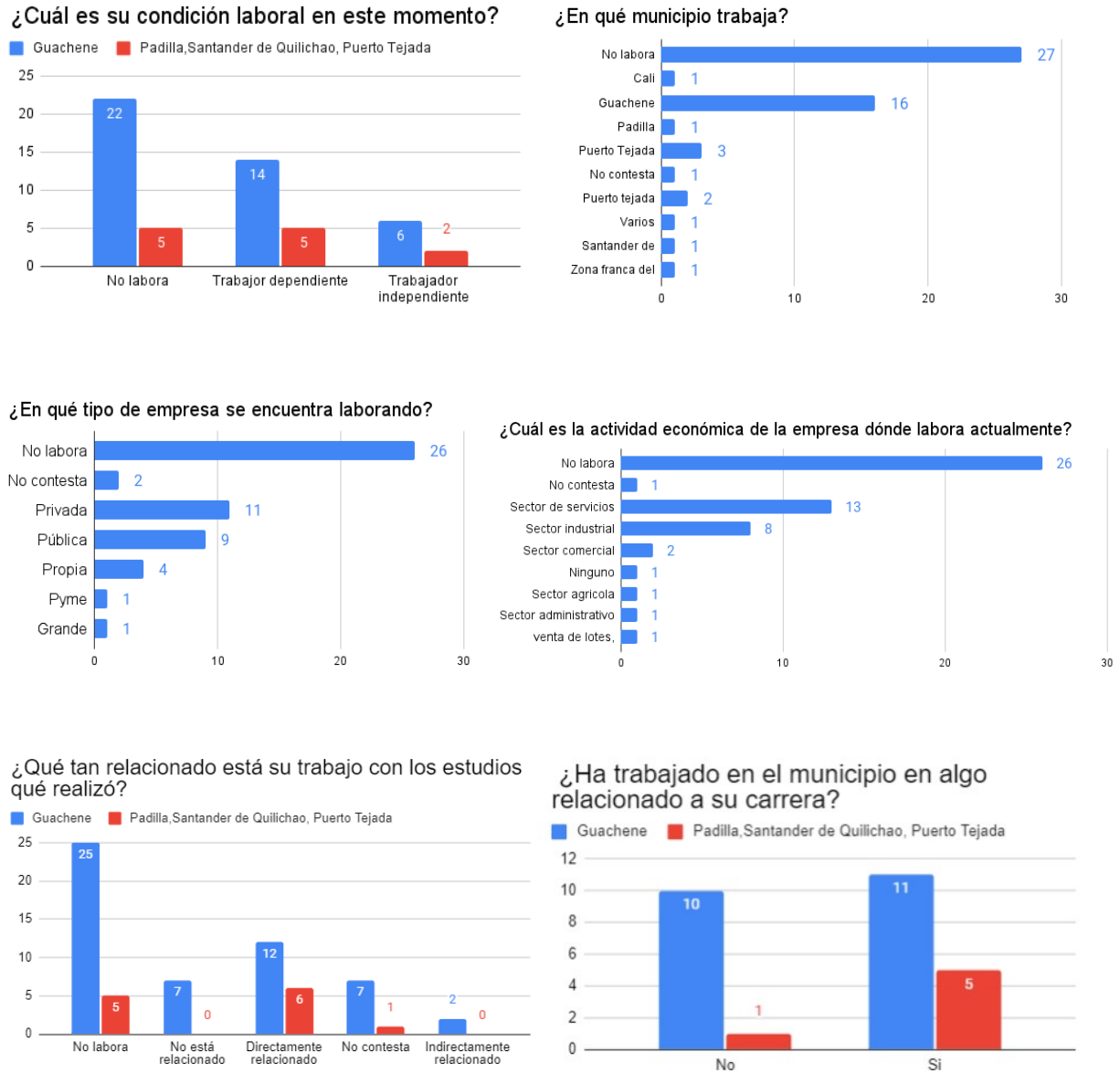
Figura 14 *Información del vínculo familiar y estado civil.*

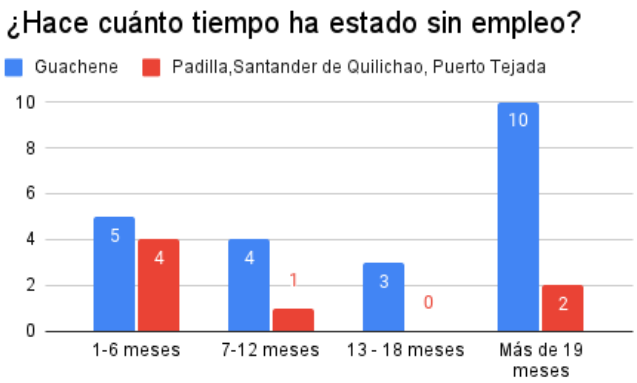


En información laboral se recolectó la condición laboral de los encuestados para lograr determinar en primera instancia, si las carreras que participaron en la Alianza CERES Salto-Afro, Norte del Cauca en la cual culminaron sus estudios, algo a resaltar es que el 55,9% de las personas han trabajado en algo relacionado a su carrera, esto nos puede dar un acercamiento a establecer aspectos que aportaron de manera positiva en el proyecto, pero de igual forma da una

alerta el conocer que el 22,2% de las personas que no laboran se encuentran desempleadas desde hace más de 19 meses como muestra en la Figura 15.

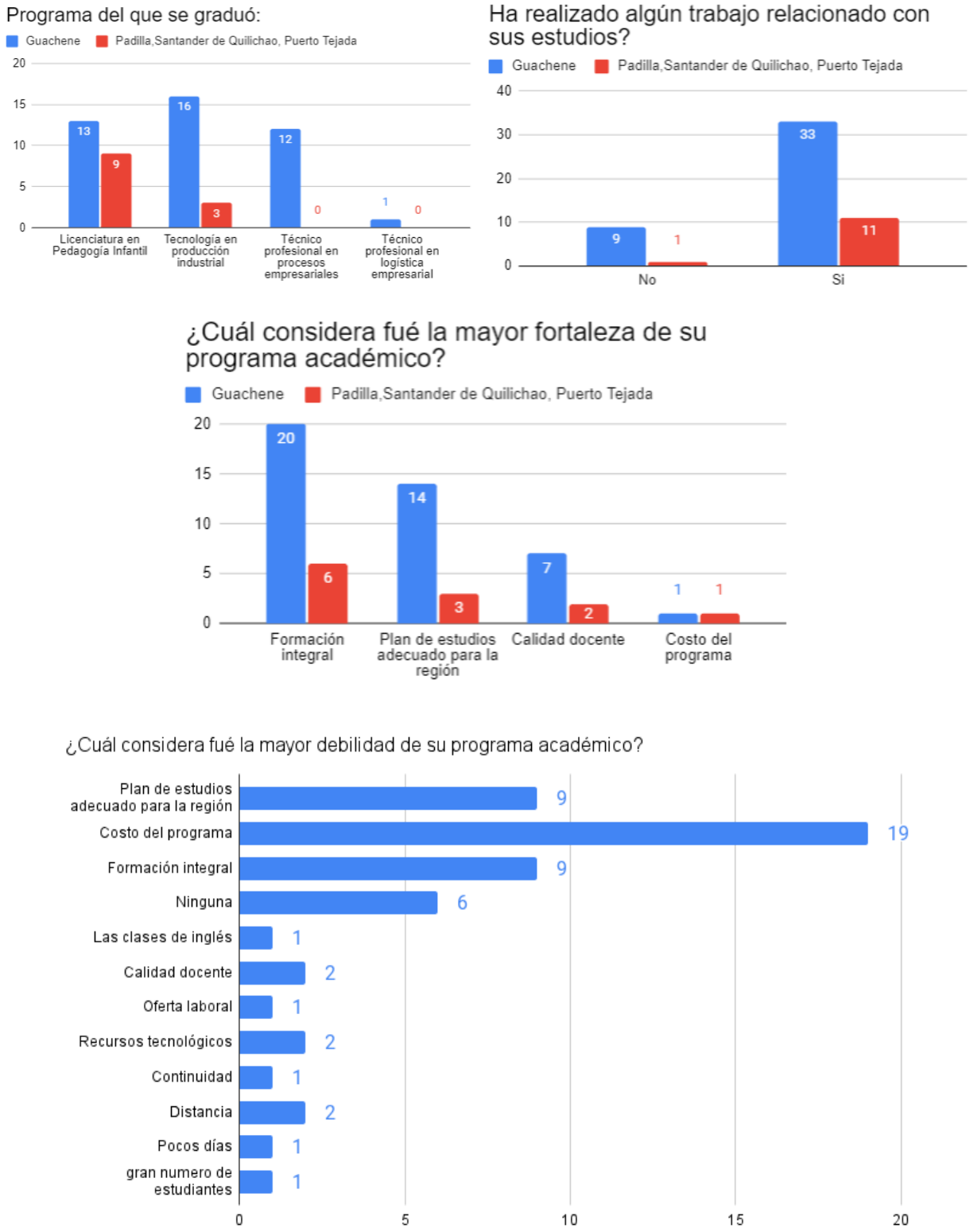
Figura 15 *Información laboral.*





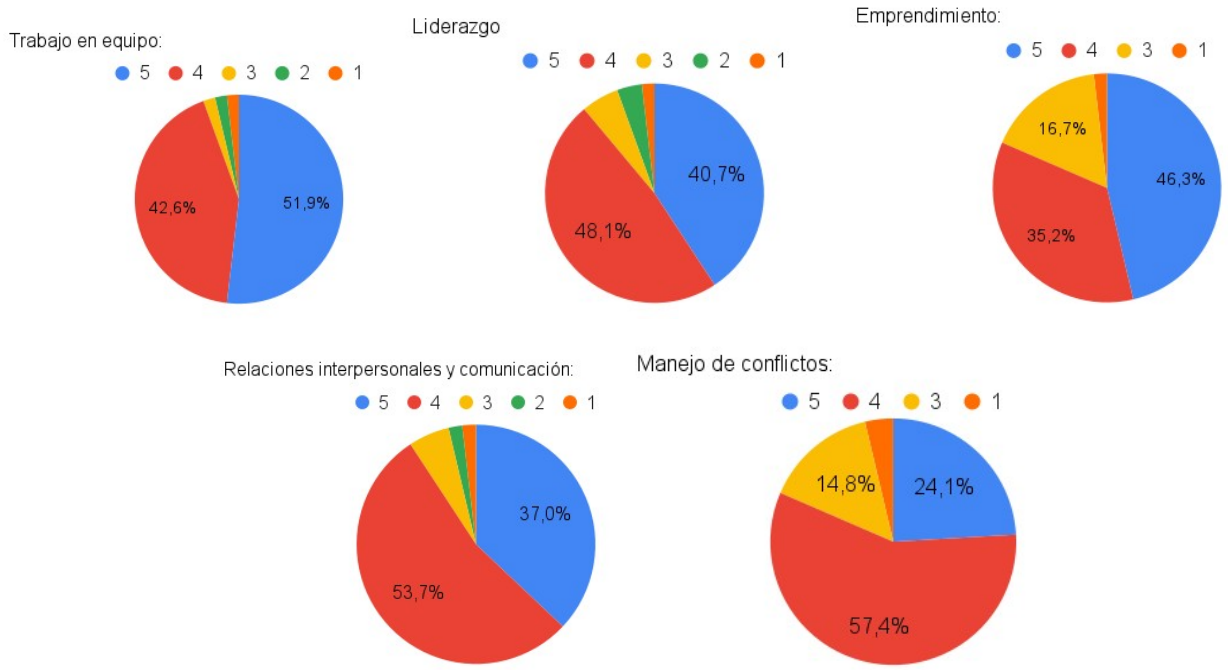
Como información académica se obtuvieron las perspectivas de cada encuestado con respecto al programa academico finalizado o que se encuentra en curso, entre los datos se muestra que el programa con la mayor participación en la encuesta es Licenciatura en pedagogia infantil con un 40,74%, continua Tecnologia en producción industrial con el 35,19% de los datos, Técnico profesional en procesos empresariales 22,22% y por ultimo Técnico profesional en logística empresarial con un 1,85% de los datos. Como debilidad del programa academico los estudiantes encuestados resaltan el costo del programa con un 35,19% algo que se tiene que entrar analizar con respecto a los beneficios ofrecidos por el proyecto, continuando se encuentran el plan de estudios adecuado para la región y formación integral con un 16,67% cada uno de ellos y el 11,11% de los estudiantes considera que el programa academico no tuvo debilidad como se muestra en la Figura 16.

Figura 16 *Información académica.*



En la encuesta la perspectiva que tuvieron sobre algunas de las competencias del programa académico, evaluado de 1 a 5, siendo 1 insuficiente, 2, deficiente, 3 aceptable, 4 bueno y 5 excelente. Los estudiantes respondieron que el trabajo en equipo se trabajó excelente con un 51,9% y bueno con un 42,6%, el liderazgo obtuvo un 48,1% de bueno y 40,7% de excelente, por otra parte emprendimiento obtuvo el 46,3% de excelente, 35,2% de bueno y un 16,7 de aceptable, en relaciones interpersonales la mayoría selecciono que su programa académico fue bueno con un 53,7% y un 37% considera que fue excelente, por último el manejo de conflictos obtuvo un 57,4% en bueno, 24,1% de excelente y un 14,8%, en aceptable como se muestra en la Figura 17, observando los datos se obtiene una calificación positiva en las competencias evaluadas en la encuesta, esto nos servirá para poder comparar y evaluar los programas académicos ofrecidos en el proyecto Alianza CERES Salto-Afro, Norte del Cauca.

Figura 17 *Evaluación de competencias en el programa académico que cursó*



Realizando un consolidado de la información obtenido por los municipios, se logra visualizar que en Guachené frente a los municipios que lo rodean como lo son Padilla,Santander de Quilichao y Puerto Tejada esto lo podemos ver en la Figura 18, son bastante similares teniendo en cuanta la diferencia en el numero de estudiantes encuestados. En ambos se encuentra una mayor población de mujeres, predominando en su gran mayoria la raza afrodecendiente, mantienen una población en los mismos rangos de edades entre los 18 y 35 años de edad, la información del vinculo familiar es una de las mas semejantes, ya que cuentan con igualdad de condiciones, la mitad de los encuestados contestaron que sí tenian hijos al iniciar el programa y la otra mitad contestaron que no, manteniendo que la mayor parte son cabeza de hogar.

Este patron no continua en la información laboral, encontrando un mejor porcentaje de personas empleadas en los municipios de Padilla,Santander de Quilichao y Puerto Tejada a comparación de Guachené en el cual la mayor parte de los encuestados se encuentran desempleados hace 13 meses o mas, de igual manera se evidencia un mayor grupo de persona laborando en trabajos relaciónados a su carrera en estos municipios.

4. Resultados

A través de las encuestas se recolecto 54 datos de estudiantes y egresados, de las diferentes regiones del norte del cauca tal como lo son Padilla, Santander de Quilichao, Puerto Tejada y Guachené, tomando como participación en su gran mayoría en los programas de licenciatura en pedagogía infantil y tecnología en producción, estos datos permitieron a través de la selección de atributos que se lograra establecer las características esenciales dependiendo de la clase seleccionada, en un primer intento se utilizó la combinación Ranker con InfoGainAttributeEval seleccionando como clase la columna de situación académica actual, arrojando como atributos de valor el municipio de residencia, rango de edad, persona cabeza de hogar y la asignatura preferida, dando como resultado para el algoritmo ZeroR 54,717% de instancias correctamente clasificadas, esto da el punto de comparación para las siguientes ejecuciones como es el caso del algoritmo PART, el cual arrojo como resultado 54,717% el mismo que el ZeroR lo cual no es buena señal ya que el valor predictivo de este algoritmo es nulo, el algoritmo J48 continuo con el mismo porcentaje de instancias correctamente clasificadas, dando como interpretación que las características seleccionadas no permiten

predecir el resultado de la clase seleccionada, esta información se ve resumida en la tabla 2, en la cual se muestra esta primera ejecución de los algoritmos en el conjunto de datos obtenidos.

Tabla 2. *Comparación de los algoritmos ZeroR, PART y J48, con los atributos seleccionados por los algoritmos Ranker e InfoGainAttributeEval como clase la situación académica actual.*

	Algoritmos		
	ZeroR	PART	J48
Instancias correctamente clasificadas	54,717%	54,717%	54,717%
Instancias incorrectamente clasificadas	45,283%	45,283%	45,283%

Se utilizó un segundo método utilizando los algoritmos de Greedy Stepwise y ClassifierSubsetEval, al ser un método Wrapper hay que seleccionar un clasificador el cual es el J48 y el PART, los cuales asignaron como variables de mayor relevancia el municipio de residencia, rango de edad actual, tiempo desempleado y fortaleza del programa, dando como resultado del algoritmo ZeroR 54,717% de las instancias correctamente clasificadas este resultado no vario de la ejecución anterior pero servirá como base para comparar los siguientes algoritmos como se muestra en la tabla 3, el algoritmo PART y el J48 obtuvieron 64,1509% y 62,2642% de las instancias correctamente clasificadas respectivamente, de acuerdo a los resultados el método Wrapper selecciono mejor los atributos que el utilizado en primera instancia, en esta ejecución el algoritmo PART se adapta un poco mejor a los datos obtenidos generando como reglas de decisiones:

- Rango edad actual = 24-35: Termino el programa académico que inicio con la Alianza Ceres (18.0/5.0)
- Rango edad actual = 36-45: Decidió no continuar estudiando en el programa que inicio por medio de la Alianza Ceres, pero en el futuro piensa ingresar a estudiar de nuevo (8.0/2.0)

- Fortaleza programa = Plan de estudios adecuado para la región AND Municipio residencia = Guachené: Estudiando actualmente en la Uniajc en la misma carrera en la que inicio en la Alianza Ceres (6.0/1.0)
- : Termino el programa académico que inicio con la Alianza Ceres (21.0/8.0)

Tabla 3. *Comparación de los algoritmos ZeroR, PART y J48, con los atributos seleccionados por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase la situación académica actual.*

	Algoritmos		
	ZeroR	PART	J48
Instancias correctamente clasificadas	54,717%	64,1509%	62,2642%
Instancias incorrectamente clasificadas	45,283%	35,8491%	37,7358%

Figura 18. *Resultado arrojado por el algoritmo PART en el conjunto de datos seleccionado por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase la situación académica actual.*

```
=== Summary ===

Correctly Classified Instances          34               64.1509 %
Incorrectly Classified Instances        19               35.8491 %
Kappa statistic                          0.3722
Mean absolute error                      0.198
Root mean squared error                  0.3286
Relative absolute error                 77.5156 %
Root relative squared error             92.9154 %
Total Number of Instances               53
Ignored Class Unknown Instances                 1

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,364    0,143    0,400      0,364    0,381      0,229    0,640     0,396     Estudiando actualmente en la Uniajc en la misma carr
               0,600    0,047    0,750      0,600    0,667      0,605    0,768     0,510     Decidio no continuar estudiando en el programa que i
               0,000    0,000    ?          0,000    ?          ?        0,942     0,292     Estudiando actualmente en la Uniajc en una carrera d
               0,828    0,458    0,686      0,828    0,750      0,388    0,692     0,692     Termino el programa academico que inicio con la Alia
               0,000    0,000    ?          0,000    ?          ?        0,396     0,019     Decidio no continuar estudiando en el programa que i
Weighted Avg.  0,642    0,289    ?          0,642    ?          ?        0,699     0,569

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  4  0  0  7  0 |  a = Estudiando actualmente en la Uniajc en la misma carrera en la que inicio en la Alianza Ceres
  1  6  0  3  0 |  b = Decidio no continuar estudiando en el programa que inicio por medio de la Alianza Ceres, pero en el futuro piensa ingresar a
  2  0  0  0  0 |  c = Estudiando actualmente en la Uniajc en una carrera diferente, porque decidio cambiar la que incio en la Alianza Ceres
  3  2  0 24  0 |  d = Termino el programa academico que inicio con la Alianza Ceres
  0  0  0  1  0 |  e = Decidio no continuar estudiando en el programa que inicio por medio de la Alianza Ceres y no piensa continuar estudiando
```

Con los resultados que se encuentran en la figura 18, podemos determinar por el momento con los datos que se encuentran a disposición de semillero, que la edad es un factor importante a la hora de clasificar la situación académica, los resultados obtenidos se pueden interpretar que las personas con un rango actual entre 24 y 35 años, terminaron el programa

académico que inicio, por otro lado las personas que tenían un rango de edad entre 36 y 45 años decidieron no continuar el programa académico, pero desean en un futuro volver a vincularse a los estudios, igualmente Guachené al ser un municipio predominante en los datos y con mayor participación en la convocatoria, se ve relacionado con los estudiantes que creen en el plan de estudio adecuado para la región como una fortaleza del programa.

En la siguiente ejecución se trató como clase de valor el atributo del tiempo de desempleo, se realizó la selección de atributos nuevamente con los métodos Greedy Stepwise y ClassifierSubsetEval, usando como clasificador el algoritmo J48, el cual arrojo como atributos de valor el municipio de residencia, barrio, si el estudiante trabajo profesionalmente, programa académico, si conocía las posibilidades que le brindaba el programa, fortaleza del programa y sus debilidades, obteniendo resultados cero predictivos en sus ejecuciones como se puede observar la comparación en la tabla 4, todos los algoritmos obtuvieron desempeños por debajo de las instancias incorrectamente clasificadas, por lo cual su valor predictivo es nulo y no se obtiene información valiosa al respecto, el algoritmo con el resultado más representativo fue el J48 el cual el resultado se encuentra en la figura 19.

Tabla 4. *Comparación de los algoritmos ZeroR, PART y J48, con los atributos seleccionados por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase el tiempo desempleado.*

	Algoritmos		
	ZeroR	PART	J48
Instancias correctamente clasificadas	41,3793%	24,1379%	44,8276%
Instancias incorrectamente clasificadas	58,6207%	75,8621%	55,1724%

Figura 19. *Resultado arrojado por el algoritmo J48 en el conjunto de datos seleccionado por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase el tiempo desempleado.*

```
=== Summary ===

Correctly Classified Instances          13               44.8276 %
Incorrectly Classified Instances        16               55.1724 %
Kappa statistic                          0.0812
Mean absolute error                      0.3356
Root mean squared error                  0.4373
Relative absolute error                 94.7141 %
Root relative squared error            103.8178 %
Total Number of Instances               29
Ignored Class Unknown Instances                25

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,917    0,824    0,440      0,917    0,595      0,133    0,648     0,322     Mas de 19 meses
                 0,000    0,000    ?          0,000    ?          ?        0,528     0,094     7-12 meses
                 0,222    0,100    0,500      0,222    0,308      0,164    0,332     0,143     1-6 meses
                 0,000    0,000    ?          0,000    ?          ?        0,365     0,071     13 - 18 meses
Weighted Avg.    0,448    0,372    ?          0,448    ?          ?        0,494     0,200

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 11  0  1  0 |  a = Mas de 19 meses
  4  0  0  0 |  b = 7-12 meses
  7  0  2  0 |  c = 1-6 meses
  3  0  1  0 |  d = 13 - 18 meses
```

Por otro lado este mismo conjunto de variables se vieron favorecidas en los algoritmos de agrupamiento, llevándose a cabo la ejecución del algoritmo simpleKmeans se encontró que lograba encontrar en dos y tres grupos un número significativo de instancias correctamente agrupadas que los demás, se utilizaron dos grupos para mejorar la eficiencia del algoritmo al tener que dividir en menos grupos los datos, el resumen obtenido en la figura 20 informa la clasificación del algoritmo, la comparación de cada uno de las ejecuciones variando conjuntos y su porcentaje de instancias correctamente agrupadas resultantes se encuentran en la tabla 5, se muestra como el algoritmo disminuye su porcentaje de instancias correctamente agrupadas aumentando el número de grupos.

Figura 20. *Resultado arrojado por el algoritmo simpleKmeans en el conjunto de datos seleccionado por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase el tiempo desempleado.*

```
Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      22 ( 41%)
1      32 ( 59%)


Class attribute: tiempo desempleado
Classes to Clusters:

 0 1  <-- assigned to cluster
 4 8 | Mas de 19 meses
 1 3 | 7-12 meses
 4 5 | 1-6 meses
 2 2 | 13 - 18 meses

Cluster 0 <-- 1-6 meses
Cluster 1 <-- Mas de 19 meses

Incorrectly clustered instances :       17.0     31.4815 %
```

Tabla 5. *Comparación de resultados obtenidos del algoritmo simpleKmeans con los atributos seleccionados por los algoritmos Greedy Stepwise y ClassifierSubsetEval como clase el tiempo desempleado.*

	Número de grupos		
	2	3	4
Instancias correctamente agrupadas	68,5185%	68,5185%	66,6667%
Instancias incorrectamente agrupadas	31,4815%	31,4815%	33,3333%

Los resultados arrojados por cada uno de estos grupos tabla 6, muestra que los más representativos son de 1 a seis meses desempleado, mostrando como datos agrupados los municipios de Padilla, Santander de Quilichao y Puerto Tejada, específicamente el barrio la esperanza, en el programa académico licenciatura en pedagogía infantil, esto evidencia una dificulta que se tiene para el programa de generar empleos en estas regiones. Por otra parte en más de 19 meses desempleados se encuentran agrupados los estudiantes de Guachené, en este caso el barrio agrupado es Jorge Eliecer Gaitán, siendo mucho más evidente la problemática pero

hacia el programa académico de Tecnología en producción, en ambos conjuntos se encuentra relacionada la característica que los estudiantes no han trabajado profesionalmente.

Tabla 6. *Resultado del algoritmo simpleKmeans utilizando 2 conjuntos para encontrar agrupaciónes en los datos.*

	Grupos	
	1 a 6 meses	Más de 19 meses
Municipio residencia	Padilla, Santander de Quilichao y Puerto Tejada	Guachené
Barrio	La esperanza	Jorge Eliecer Gaitan
trabajo profesionalmente	No	No
programa académico	Licenciatura en Pedagogía Infantil	Tecnología en producción
Conocía las posibilidades que le brindaba el programa	Si	Si
Fortalezas del programa	Formación integral	Formación integral
debilidades del programa	Costo del programa	Costo del programa

5. Discusión

Los datos arrojados por las diferentes técnicas de análisis de datos en los resultados mostrados en las tablas 3, dan a interpretar que, la edad de los estudiantes de la región ha afectado la culminación de los estudiantes, siendo esto un factores a tener en cuenta en futuras convocatorias para lograr establecer planes de acción o acompañamiento para las personas que se encuentren en rangos de edades avanzadas y evitar deserciones académicas altas como ocurre en esta ocasión.

De igual forma es pertinente establecer mejores asociaciones entre la academia y la industria para evitar números de desempleos que se prolongan en años, cada región obtiene diferentes problemáticas de la industria que los rodea que necesitan solucionar, en algunas ocasiones las carreras que se ofertan no cubren esta necesidad por esta razón se requiere ejecutar mejores planes de prácticas laborales escuchando las empresas que rodean las áreas y fortaleciendo las competencias del pensum académico enfocándolas a las necesidades que estas tenga.

Es necesario seguir aumentando el volumen de datos de las encuestas realizadas, ya que de esto dependerá obtener mejores resultados y ofrecer mejores planes de desarrollo en cada uno de los municipios que tenga el proyecto Alianza CERES Salto-Afro, Norte del Cauca y la Institución Universitaria Antonio José Camacho, obteniendo en cada resultado nuevos análisis e interpretaciones, esto evitara obtener resultados vacíos y facilitara la entrega de mayor información de valor en cada una de los atributos de los datos.

6. Conclusiones.

Con base en los resultados obtenidos en las diferentes aplicaciones de técnicas de análisis de datos, se encontraron patrones y relaciones en la información obtenida del proyecto Alianza CERES Salto-Afro, Norte del Cauca, como se muestra en la tabla 6 las dos agrupaciones generadas, en la cual se evidencia que en el municipio de Guachené se encuentra un déficit de empleo para los programas de tecnología en producción ya que estos estudiantes se encuentran desempleados más de 19 meses, igualmente en los municipios aledaños como lo son Padilla, Santander de Quilichao y Puerto Tejada se encuentra un déficit de desempleo en el programa de licenciatura en pedagogía infantil, llegando a estar los estudiantes entre 1 y 6 meses desempleados, el patrón en común que se encuentran en estos dos conjuntos es el no haber trabajado profesionalmente, lo cual da una alerta y aporta a futuros proyectos a generar convenios efectivos con las empresas que rodean e implementando pensum académicos apropiados a la región, con prácticas profesionales que le aporten experiencia de peso a los egresados de cada programa académico.

En relación a los datos obtenidos en la tabla 3, el algoritmo PART arroja en las reglas de calificación, que los estudiantes presentan un gran índice de deserción en los programas dependiendo de su edad, los rangos de edades más afectados son entre los 36 y 45 años, dando como resultado en los estudiantes que decidieron no continuar estudiando en el programa que inicio por medio de la Alianza Ceres, pero en el futuro piensa ingresar a estudiar de nuevo.

A través de la metodología KDD se logró extraer información de valor en los conjuntos de datos obtenidos del proyecto Alianza CERES Salto-Afro, Norte del Cauca, gracias a esta metodología se establece un ciclo que permite realizar iteraciones entre sus etapas, efectuar una correcta limpieza de los datos, estableciendo la selección de atributos, permitiendo implementar nuevos modelos de análisis de datos, logrando nuevas interpretaciones y su réplica en diferentes proyectos futuros.

Referencias

Angulo, J. F., & Blanco, N. (1994). *TEORÍA Y DESARROLLO DEL CURRÍCULUM*.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. En J. Kogan, C. Nicholas, & M. Teboulle (Eds.), *Grouping Multidimensional Data: Recent Advances in Clustering* (pp. 25-71). Springer. https://doi.org/10.1007/3-540-28349-8_2

Dueñas-Reyes, M. X. (2009). *Minería de datos espaciales en búsqueda de la verdadera información*. 20.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34. https://doi.org/10.1145/240455.240464

Frank, E., Hall, M. A., & Witten, I. (2016). *Data Mining: Practical Machine Learning Tools and Techniques: Vol. Fourth Edition*. Morgan Kaufmann. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

Gutiérrez, J. A., & Molina, B. (2015). Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. *Revista Ontare*, *3*(2), 33-51. https://doi.org/10.21158/23823399.v3.n2.2015.1440

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

K-means Clustering en Machine Learning (K-medias). (2020, mayo 2). *DATA SCIENCE*. https://datascience.eu/es/aprendizaje-automatico/k-means-clustering-en-machine-learning/

Laaksonen, J., & Oja, E. (1996). Classification with learning k-nearest neighbors. *Proceedings of International Conference on Neural Networks (ICNN'96)*, *3*, 1480-1483. https://doi.org/10.1109/ICNN.1996.549118

Menacho Chiok, C. H. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos*, *78*(1), 26-33.

Raval, K. (2012). *Data Mining Techniques* (Vol. 2). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.898.6657&rep=rep1&type=pdf

Rychen, D. S., & Salganik, L. H. (2006). *Las competencias clave para el bienestar personal, social y económico* (J. M. P. Corredor, Trad.).

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A., & Alvarado Pérez, J. C. (2016). *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*. Universidad Cooperativa de Colombia. https://doi.org/10.16925/9789587600490

Zhao, Q., & Bhowmick, S. (2003). Association Rule Mining: A Survey. *Technical Report*, 20.