MODELAMIENTO DE DATOS PARA LA RECOMENDACIÓN DE PROGRAMAS DE FORMACIÓN POSGRADUAL Y COMPLEMENTARIOS BASADOS EN LA INTELIGENCIA DE DATOS PARA LA CORPORACIÓN UNIVERSITARIA COMFACAUCA - UNICOMFACAUCA

BRAYAN HERNÁN NAVIA ORDOÑEZ
HARLINSON LUGO MOSQUERA

CORPORACIÓN UNIVERSITARIA COMFACAUCA-UNICOMFACAUCA
FACULTAD DE INGENIERÍAS
DEPARTAMENTO DE SISTEMAS
POPAYÁN 2021

MODELAMIENTO DE DATOS PARA LA RECOMENDACIÓN DE PROGRAMAS DE FORMACIÓN POSGRADUAL Y COMPLEMENTARIOS BASADOS EN LA INTELIGENCIA DE DATOS PARA LA CORPORACIÓN UNIVERSITARIA COMFACAUCA - UNICOMFACAUCA

BRAYAN HERNÁN NAVIA ORDOÑEZ
HARLINSON LUGO MOSQUERA

TRABAJO DE GRADO

DIRECTOR: MG. JULIÁN RENE MUÑOZ
Co-DIRECTOR: PhD. GINETH MAGALY CERÓN

CORPORACIÓN UNIVERSITARIA COMFACAUCA-UNICOMFACAUCA
FACULTAD DE INGENIERÍAS
DEPARTAMENTO DE SISTEMAS
POPAYÁN 2021

Nota de aceptación:

El director de la opción de grado y el jurado evaluador del trabajo han aprobado el presente documento de grado, cumpliendo con los requisitos exigidos por la Corporación Universitaria Comfacauca de Popayán, para optar al título de ingeniería sistemas.

JULIAN RENE MUÑOZ BURBANO
_____
Director opción de grado

HELDER YESID CASTRILLON COBO
FRANCISCO JAVIER OBANDO
_____
Jurado evaluador

Agradecimientos

Como primero queremos dar gracias a Dios por darnos la oportunidad de culminar con éxito el presente proyecto de grado, añadido a esto queremos agradecer a nuestros padres por ser los principales motivadores y promotores de esta meta que hoy está a punto de cumplirse, también y no menos importante queremos agradecer a nuestros tutores que nos acompañaron durante el proceso de construcción y desarrollo del proyecto, nuestro éxito se debe gran parte a su excelente vocación docente y su disposición para ser unos mentores constantes en el cumplimiento de nuestros objetivos académicos, no queremos dejar sin reconocer a cada uno de los compañeros, docentes, administrativos y en general todo el personal que se involucró durante todo el proceso de desarrollo del proyecto como también de nuestra vida lectiva en la corporación universitaria Unicomfacauca. Para nosotros es un honor hacer parte de esta institución tan comprometida con la excelente calidad de sus estudiantes y egresados.

Por último queremos dar las gracias al personal encargado de egresados y todo el personal que hizo parte activa del proyecto, ya que su apoyo fue crucial para la consecución de los objetivos de este proyecto.

Resumen

Se realizó un modelamiento de datos para la recomendación de formación pos gradual y complementario, basado en la inteligencia de datos para la Corporación Universitaria Unicomfacauca. El proyecto se procedió porque la universidad Unicomfacauca en su actualidad requiere la reincorporación de sus egresados y es de suma importancia ofrecer cursos o formaciones pos gradual que estén de acuerdo a sus preferencias y a su perfil amiento. Para ello es necesario modelar a los usuarios y por consiguiente modelar los datos que se tienen del usuario.

Con un sistema basado en la inteligencia de datos, se puede recomendar de una manera más eficiente los programas de posgrado esto con base a los algoritmos y técnicas de recomendación, con un algoritmo híbrido que permite recomendar lo más popular con un filtrado colaborativo y con un filtrado basado en conocimiento, al mismo tiempo si el usuario no brinda los datos correspondientes o no completos el algoritmo se basa en conocimientos previos con personas que tienen un perfil parecido; se realizó un mapeo y revisión de literatura enfocada en estas técnicas de algoritmos, se analizó el problema basado en el diseño centrado en el usuario esto en cuanto a lo que se necesitaba en la Universidad, se modelaron los datos con base en la caracterización de los usuarios que se obtuvieron y por consiguiente esta arrojo un conjunto de datos el cual se le realizó una limpieza, una sustracción de datos y una transformación de datos para dejarlos acorde a una minería de datos basado en los siguientes algoritmos:

J48, KNN y CBR, de los cuales se pudo obtener unos resultados en porcentajes los cuales fueron significativamente buenos, sin embargo, se eligió un algoritmo híbrido porque si el usuario no brinda sus datos completos el algoritmo híbrido puede

recomendarle el más popular y si el usuario brinda sus datos completos basados en conocimiento como también se puede discriminar o recomendar de las dos formas, además se le recomienda por reglas pensando en que no se le repita una recomendación que ya ha hecho al usuario a un curso que ya haya tomado; con lo cual se concluye que este modelado de datos facilitara la implementación de un sistema de recomendación en Unicomfacauca para un trabajo a futuro considerando que este algoritmo híbrido tuvo una eficiencia superior al 80% igualmente facilitara las decisiones del usuario ofreciéndole una variante adicional que son sus preferencias. Con este modelado se le facilita a la universidad implementar su sistema con base en los datos que cuenta la institución.

Palabras claves: J48, KNN, CBR, Algoritmo híbrido, Dataset, Inteligencia de datos, minería de datos.

Abstract

A data modeling for the recommendation of postgraduate and complementary training, based on data intelligence for the Corporación Universitaria Unicomfacauca. The project was carried out because Unicomfacauca University currently requires the re-incorporation of its graduates, and it is of utmost importance to offer courses or postgraduate training that are according to their preferences and their profile, for this it is necessary to model the users and therefore model the data that we have about the user. With a system based on data intelligence, it is possible to recommend in a more efficient way the postgraduate programs based on algorithms and recommendation techniques, with a hybrid algorithm that allows to recommend the most popular with a collaborative filtering and with a filtering based on knowledge, at the same time if the user does not provide the corresponding data or not complete the algorithm is based on previous knowledge with people who have a similar profile; a mapping and literature review focused on these algorithms techniques was performed, the problem was analyzed based on the user-centered design in terms of what was needed at the University, the data were modeled based on the characterization of the users that were obtained, and therefore this characterization yielded a Dataset which was performed a cleaning, a subtraction of data and a transformation of data to leave them according to a data mining based on the following algorithms:

J48, KNN and CBR, of which it was possible to obtain some results in percentages which were significantly good, however a hybrid algorithm was chosen because if the user does not provide their complete data the hybrid algorithm can recommend the most popular and if the user provides their complete data based on knowledge as can also be

discriminated or recommended in both ways, in addition it is recommended by rules thinking about not repeating a recommendation that has already made the user to a course that has already taken; with which it is concluded that this data modeling will facilitate the implementation of a recommendation system in Unicomfacauca for a future work considering that this hybrid algorithm had an efficiency higher than 80% and will also facilitate the user's decisions by offering an additional variant that are their preferences. With this modeling, it is easier for the university to implement its system based on the data that the institution has.

Keywords: J48, KNN, CBR, hybrid algorithm, Dataset, data intelligence, data mining.

Tabla de contenido

Lista de tablas

Lista de imágenes

Lista de figuras

Glosario

Algoritmo J48: El algoritmo C4.5 (J48) es uno de los más utilizados en la minería de datos. Se basa en la entropía de la información que permite a medir la perturbación presente en un conjunto de datos.

Algoritmo de aprendizaje: Son piezas de código que ayudan a las personas a explorar, analizar y encontrar significados en conjuntos de datos complejos.

Anaconda: Anaconda es una distribución abierta y gratuita de los lenguajes Python y R, que se usa en ciencia de datos y aprendizaje automático.

Atributo: Cualidad o característica de una persona o cosa, especialmente algo que es parte esencial de su naturaleza.

Ciencia de datos: La ciencia de datos combina varias áreas como estadísticas, métodos científicos, inteligencia artificial (IA) y análisis de datos para extraer valor de los datos.

Clúster: Es un conjunto de datos que se agrupan en conglomerados que contienen características similares. Estas agrupaciones son útiles para explorar datos, identificar anomalías en los datos y crear predicciones.

Clustering: Es una técnica utilizada en la minería de datos (dentro del área de la inteligencia artificial) para identificar automáticamente las agrupaciones (clúster) de elementos según una medida de similitud entre ellos. Esta técnica también se conoce como segmentación.

CRISP-DM: Es un método probado para guiar su trabajo de minería de datos. Proporciona una descripción general del ciclo de vida de la minería de datos.

Dataset: Un conjunto de datos tabulares en cualquier sistema de almacenamiento de datos estructurados, una base de datos de fuente única que se puede vincular a otras, cada columna del registro representa una variable y cada fila corresponde a los datos que se están tratando.

Egresados: Persona natural que ha completado y aprobado satisfactoriamente todo el plan de estudios regulado de una carrera o carrera, pero que aún no ha obtenido el título académico.

Empleabilidad: Es la capacidad de adaptar las circunstancias, **h**a**bil**idades, competencias y conocimientos profesionales y personales a las necesidades del mercado.

Inteligencia de datos: Son el análisis empresarial, la minería, la visualización, las herramientas y la infraestructura de datos y las mejores prácticas las que ayudan a las organizaciones a tomar decisiones basadas en datos.

Jupyter notebook: Es una aplicación web de código abierto que nos permite crear y compartir código y documentos. Es un entorno informático interactivo, que permite a los usuarios experimentar con el código y compartirlo.

K-means: K-Means (traducido como K-Medias en español), es un método de agrupamiento o clustering.

Marketing: Es la ciencia y el arte de investigar, producir y entregar valor para satisfacer las necesidades de un mercado objetivo con fines de lucro. Definir, medir y cuantificar el tamaño del mercado identificado y el beneficio potencial.

Matriz de confusión: Es una herramienta que permite visualizar el desempeño de un algoritmo utilizado en el aprendizaje supervisado.

Minería de datos: Es el proceso de detectar la información procesable de los conjuntos grandes de datos. Emplea el análisis matemático para deducir los patrones y tendencias que existen en los datos.

Pentaho: Es una plataforma de Business Inteligente (BI) centrada en procesos y orientada a soluciones que incluye los componentes necesarios para implementar soluciones basadas en procesos como minería de datos, ETL e informes para mejorar las capacidades de captura y análisis de datos.

Posgrados: Corresponden al curso de estudios avanzados completado después de la licenciatura. Las titulaciones asociadas a un título universitario son una maestría, una especialización o un posgrado.

Python: Python es un lenguaje de programación fácil de leer y escribir debido a su gran parecido con el lenguaje humano. Además, se trata de un lenguaje multiplataforma de código abierto, por tanto, gratuito, que permite desarrollar software sin límites.

Prospecto: Es un usuario que encaja con las características de un usuario ideal que

Dispone de los medios para tomar sus propias decisiones.

Software: Software, software o soporte lógico es el nombre que se le da al sistema formal de un sistema informático que incluye todos los componentes lógicos necesarios que permiten realizar determinadas tareas.

Weka: Es un software que contiene una colección de herramientas de visualización y algoritmos para el análisis de datos y el modelado predictivo, así como una interfaz gráfica de usuario para un fácil acceso a las funciones.

Capítulo primero

1. Introducción.

Dadas las dinámicas actuales donde se generan grandes cantidades de datos e información en las distintas áreas de trabajo que involucran las empresas y organizaciones, incluidas las instituciones de educación superior, se trabaja la oportunidad de obtener beneficio de la información que conciben grandes conjuntos de datos para brindar mejores producto y servicios. Actualmente, el área de egresados de la corporación universitaria Unicomfacauca realiza ofertas académicas para sus egresados por medio del portal web de la universidad y de las redes sociales, estas propuestas muchas veces no están enfocadas a las reales necesidades de los egresados de la corporación y en consecuencia se evidencia la falta de un interés más alto por parte de los mismos.

Es por ello, que el área de egresados busca generar información efectiva para brindar mejores servicios relacionados con la mejora en procesos académicos y de esta forma ofrecer programas o cursos de extensión para mejorar las habilidades de los egresados. Dicha área efectuó un estudio donde se obtiene información basada en encuestas efectuadas a los egresados, esta propuesta pretende, por medio de inteligencia de datos analizar la información y generar un modelo basados en los datos de estos formularios que permitan obtener un perfila miento de egresados a través de la aplicación de diferentes técnicas, herramientas e investigaciones y de esta manera encontrar un algoritmo adecuado que recomiende con efectividad los programas o cursos de extensión a los egresados según sus intereses.

1.1 Formulación del problema.

El área de egresados de la Corporación Universitaria Comfacauca - Unicomfacauca, realiza seguimiento y actualización de información a los egresados de la corporación, para brindar ofertas académicas como cursos de extensión y programas de especialización para de esta manera estar acordes a las necesidades del mercado. Actualmente, estos procesos se publicitan por medio del portal web de la universidad y de las redes sociales autorizadas, lo que ocasiona que dichas propuestas no están enfocadas a las reales necesidades de los egresados y por ende se evidencia la falta de un interés más alto por parte de estos, las ofertas no están orientadas al usuario, y hace que la información en muchos de los casos no sea de su interés, motivo por el cual la presente propuesta pretende hacer un estudio detallado sobre la información de los egresados como sus intereses, áreas de interés entre otros, que permitan establecer ofertas con resultados más cercanos a sus intereses y con un alto grado de aceptación. Por otro lado, los métodos de atracción actualmente no son enfocados y son poco atractivos a la hora de incentivar a los egresados.

En el contexto que proporciona el mundo moderno donde las organizaciones generan grandes volúmenes de datos, es necesario sacar provecho de esta información para convertirla en un activo de valor, esta propuesta pretende obtener modelos de datos por medio de técnicas de inteligencia de datos, que permitan a los usuarios de área de egresados interactuar y aprovechar el conocimiento que estos pueden generar. Donde se busca hacer un análisis del desarrollo profesional de los egresados y de esta forma generar perfiles para el óptimo desarrollo de los procesos mercadeo. Con lo anterior se hace necesario para la Corporación reinventarse en estos procesos donde una buena

alternativa es la aplicación de inteligencia de datos para obtener perfiles académicos que puedan identificar a los prospectos y por ende se puede obtener información relacionada con sus intereses o sus preferencias académicas para la toma de decisiones en los programas o cursos de extensión.

1.2 Justificación.

Los programas pos graduales y de formación complementaria ofrecidos a través de las distintas redes sociales y la página web de la corporación, tienen como finalidad actualizar en conocimientos a los egresados a través de la oferta de estos, sin embargo, se evidencia en la investigación realizada por un tercero contratado por la universidad donde se realizó un análisis de información tras aplicar un formulario a los egresados, donde se evidenció que la mayoría de los egresados no se reintegran a la universidad por la falta de cumplimiento de sus expectativas y necesidades específicas, también basado en el mismo estudio se pudo analizar que gran parte de los egresados no laboran en temas relacionados con su profesión o siendo el caso, su programa de formación pos gradual cursado, esto indica una falta de enfoque más propicio para las necesidades específicas de los egresados, y afinidad en el área profesión a la que se dedica.

Actualmente, la universidad no cuenta con una técnica de recomendación que aporte de manera activa en alguno de las áreas que involucran a los diferentes egresados de esta, por esta razón se requiere ejecutar un análisis de técnicas de recomendación que permita identificar cuál es el algoritmo con mayor eficiencia para hacer este tipo de recomendaciones.

El desarrollo de este proyecto de investigación es importante para poder implementar a futuro un sistema de recomendación, pero todo esto re requiere de modelar los datos,

teniendo en cuenta que la actualidad no se cuenta con ese modelado de datos, ya que no han sido considerados y gestionados de una manera efectiva en la universidad con el fin de hacer ofertas más precisas en las preferencias de los usuarios.

1.3 Pregunta de investigación.

¿Cómo generar modelos de datos que mejoren el proceso de identificación de prospectos para cursos o programas de extensión aplicando técnicas de inteligencia de datos basados en la información de egresados de la Corporación Universitaria Comfacauca - Unicomfacauca?.

1.4 Objetivos.

1.4.1 Objetivo general.

Obtener un modelo de datos que permita la identificación a los prospectos de los egresados en un 80% o más de precisión con el fin de mejorar la oferta académica pos gradual y cursos de extensión, aplicando inteligencia de los datos.

1.4.2 Objetivos específicos.

- Caracterizar la información necesaria que permita la identificación de los perfiles de los egresados.
- Generar un modelo de datos que permita identificar los programas de extensión más propicios según los perfiles de egresados.
- Evaluar el modelo de datos obtenidos para garantizar la mejora de la oferta en programas de extensión para los egresados o prospectos identificados.

1.5 Organización del documento.

El documento se distribuye de la siguiente forma:

- Capítulo primero: Este capítulo contiene la introducción al documento así como la formulación del problema donde se presenta el contexto y la justificación encontrada, también contiene la descripción de los objetivos generales y específicos del proyecto.
- Capítulo segundo: Este capítulo está compuesto por el marco contextual, que permite identificar la familiarización del entorno donde se desarrolla la investigación y las áreas que involucra esta, también se encuentra el marco conceptual que consiente en la descripción de los conceptos, términos, técnicas, herramientas y metodologías utilizas en el desarrollo de la investigación, por último se encuentra el estado del arte, asiente el fundamento histórico en procesos similares realizados en investigaciones, artículos, y libros que permiten dar un sustento teórico a la investigación propuesta.
- Capítulo tercero: Este capítulo contiene la justificación de las metodologías empleadas y el desarrollo de los diferentes procesos para alcanzar los objetivos propuestos.
- Capítulo cuarto: Este último capítulo presenta la evaluación, resultados, conclusiones y aportes al futuro que obtuvieron en la investigación y desarrollo de los objetivos recomendados.

Capítulo segundo

2. Estado del arte, marco conceptual y marco contextual.

2.1 Estado del arte.

Según (M.Gomez,2020) en su guía del estado del arte expone el concepto, las características y los objetivos de un estado del arte y menciona que “el estado del arte es la búsqueda, lectura y análisis de la bibliografía encontrada en relación un tema que se quiere investigar”, teniendo en cuenta la temática en la que se ve envuelta el proyecto de investigación en el capítulo uno, se realizó una investigación enfocada en esta temática identificada, y soluciones o proyectos de investigación similares a lo que se propone en la justificación y objetivos definidos, para esto se tiene en cuenta una organización de la búsqueda, en primera instancia se determina las temáticas principales a investigar donde se identifican palabras claves y tema principal, como tercera tarea se determinan los límites espacio temporales añadido a esto se establecen los repositorios académicos en donde se realizara la búsqueda y la orientación profesional correspondiente, seguido de esto se presentan los resultados donde se construye una lista de las referencias encontradas y se realiza un nuevo límite con el fin de escoger las referencias más importantes e idóneas a considerar dentro de la ejecución del proyecto de investigación actual.

Por último se socializan los resultados más relevantes su propósito y la retroalimentación que se identificó con el proyecto de investigación a desarrollar.

2.1.1 Temática de interés.

- **Tema principal:**

Tema académico: inteligencia de datos en temas universitarios o específicamente egresados.

- **Subtemas:**

- Aplicación de técnicas de inteligencia de datos en datos de egresados.
- perfilamiento de egresados.
- Caracterización de datos para aplicación de técnicas de minería de datos.
- Mejoras en la oferta pos gradual bajo un sistema de recomendación.
- Técnicas de aprendizaje supervisado.
- Big data en el ámbito universitario.
- Sistemas de recomendación basada en casos.
- Aplicación de minería de datos en datos de egresados.

- **Palabras claves:**

- Big data
- Minería de datos
- Sistema de recomendación
- Tratamiento de datos
- Aprendizaje supervisado
- Algoritmos de clasificación
- Caracterización de perfiles

2.1.2 Limitaciones espacio temporal.

Se determinó como límite temporal el lumbral de búsqueda desde el año 2017, sin embargo, al analizar los resultados no se encontró el desarrollo de un proyecto de investigación similar, por lo que se toma la decisión de ampliar el lumbral de búsqueda desde el año 2012, en cuánto a los límites del tipo de fuente, se incluyeron todas las categorías como lo son: tesis de grado, tesis de magíster, tesis doctoral, artículos de revista, artículos web, póster, ponencias, congresos, guías, libros, monografías, patentes. Se eligieron los idiomas escogidos para la búsqueda: inglés, español, portugués.

- **Repositorios académicos:**

- Google scholar
- Scielo
- Springer Link
- RefSeek
- HighBeam research.

- **Lista de referencias:**

Como resultado de la búsqueda se cuenta con una cantidad de setenta y siete referencias entre todas las categorías investigadas, de las cuales se tomaron como referencias principales catorce de ellas que corresponden en su mayoría a proyectos de investigación desarrollados con la temática principal identificada y los subtemas propuestos, en la siguiente imagen se visualiza el modelo de la lista de referencias que fue construida por los investigadores del proyecto.

NOMBRE	DOI	URL	PDF
Aplicación web para la elaboración de perfiles de consumidor basada en minería de datos y arquitectura cloud para el apoyo al proceso de conversión de leads en la asociación AIESEC en Perú.	Teabajo de grado	http://54.165.197.99/handle/20.500.12423/2721	D:\TODO\ Documentos 2019 2020\proyecto de
Plataforma web con integración de minería de datos y redes sociales para el seguimiento a graduados del programa de ingeniería de sistemas de la universidad de Cundinamarca, extensión Facatativá	Trabajo de grado	http://repositorio.ucundinamarca.edu.co/handle/20.500.12558/2084	\Articulos mineria de datos\ Plataforma web
Aplicación de mineria de datos como una metodolgia para estimar las principales causas de desempleo y subempleo profeional de los egresadps	Trabajo de grado	http://repositorio.upao.edu.pe/handle/upaorep/6947	D:\TODO\ Documentos 2019 2020\proyecto de
Aplicando metodos y tecnicas de la ciencia de los datos a datos universitarios	ISBN: 978-987-3984-85-3	http://sedici.unlp.edu.ar/handle/10915/76941	
Las empresas impulsadas por analisis y IA prosperan en la era With		https://bit.ly/2Qlule7	
Segmentacion del comportamiento del cliente entre provedores de servicios moviles usando algoritmo k-means	https://zenodo.org/badge/DOI/10.5281/zenodo.1467663.svg	http://bit.ly/2oDNhJV	
Modelo de prediccion de la morosidad en la otorgamiento de credito financiero aplicando metologia CRIPS-DM trabajo		f/2020/09_01/3ecvsi1598957269.pdf?X-Amz-Content-Sha256=UNSIGNED-PAYLOAD&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=LB63ZNJ2Q66548XDC8M5%2F20210217%	D:\TODO\ Documentos 2019 2020\proyecto de
Sistema de gestion y soporte de toma de decisiones basdo en algoritmos de bayer y cluster para mejorar los procesos analiticos del area comecial de una empresa educativa.	Trabajo de grado	http://hdl.handle.net/20.500.12423/648	D:\TODO\ Documentos 2019 2020\proyecto de
Software de data mining: realiza análisis de datos más efectivos	Articulo pag web	https://bit.ly/2q5XM9z	

Imagen 1. Repositorio de referencias encontradas.

Los proyectos de investigación realizados por diferentes grupos de personas con el fin de obtener un título de pregrado o posgrado sobre el análisis de datos han sido valiosos para la toma de decisiones, ya que proporcionan mejoras en los programas académicos como también en la oferta de programas de posgrado. Cada vez están más alineados a las necesidades globales y locales de cada profesional y para esto se han desarrollado diferentes investigaciones, artículos y guías de referencia en cuanto al desarrollo, implementación, ventajas y desventajas de las técnicas utilizadas que sugieren su respectivo análisis de datos. Para la finalidad de este proyecto, se basa en los siguientes artículos.

Dider león González Arroyave, (2019) en el trabajo de grado para optar al título de maestría en ingeniería, propone un sistema de recomendación que estandariza los conjuntos de datos teniendo en cuenta un conjunto de métricas definidas para unos formatos propuestos, para esto se exploran diferentes algoritmos, y selecciona el que mejor lo hace según el conjunto de datos. Este proyecto presenta una un prototipo del sistema sugerido y realiza una compresión de la estandarización de los conjuntos de datos entrada sugeridos para el tratamiento de datos en sistemas de recomendación con tres conjuntos de datos diferentes. Es importante mencionar que esta recomendación se basa en la premisa de que cada cliente requiere un sistema de recomendación, y este tiene u conjunto de datos y características específicas.

Como resultado del análisis hecho al anterior proyecto socializado, se identifica un tema muy valioso para tener en cuenta dentro del desarrollo de la investigación y la modelo propuesto en los objetivos, y es el tema de requisitos para el desarrollo o construcción de los sistemas de recomendación, como lo son los aspectos de captura y actualización de los datos de entrenamiento, la evolución, selección e hibridación de algoritmos dependiendo de los datos con los que se cuente y las interfaces en las que se integra o se despliega la recomendación. Este sistema de recomendación establece la normalización y estandarización de los datos a la hora de utilizar los algoritmos de recomendación sirviendo esta temática con el desarrollo de la metodología elegida para el desarrollo del modelo, también se tienen en cuenta bases investigativas y conceptos aplicados.

El proyecto de grado realizado por Murcia, (2019) propone el desarrollo de un sistema de recomendación que permita reconocer la afinidad entre un perfil profesional y una oferta laboral publicada por una empresa, para realización de este proyecto se usó el proceso de minería de datos y conceptos técnicos que se utilizaron a partir de la minería de texto específicamente sobre las habilidades, requerimientos y funciones de las ofertas laborales y el perfil profesional y las competencias de algunos usuarios del sistema de vinculación laboral como también se tuvieron en cuenta las competencias compartidas a través de la red social LinkedIn. Es importante destacar el diseño y las alternativas de diseño propuestas para el sistema de información, también se tiene en cuenta en este proyecto una recolección, selección y transformación de datos seguido a esto la amplia descripción de la implementación del sistema de información teniendo en cuenta la relación con la base de datos del portal de vacantes profesionales.

Del anterior proyecto es importante destacar el enfoque muy similar que se le da al sistema de recomendación al que se tiene en la propuesta del proyecto actual, sin embargo, su enfoque en netamente el tema profesional de los egresados, que se ve como una alternativa o propuesta futura de aplicación para el modelo a desarrollar, también es relevante para los investigadores el manejo de la recolección, selección y trasformación de datos realizada porque ejemplifica de manera detallada las descripciones y manejo que se le da a los atributos y registros del conjunto de datos a tratar y la importancia que tiene este para una óptima construcción de un sistema de recomendación en general. Una fase importante a destacar dentro del proyecto referenciado es la implementación del sistema de información y como se tienen en

cuenta diferentes integraciones con los sistemas de relación, específicamente la plataforma de empleo vinculada con la Universidad de los Andes, como también las consideraciones a tener en cuenta para óptima implementación de un sistema de información en general, que puede servir, para una propuesta futura de implementación del modelo a construir en un sistema de información en la Corporación universitaria Unicomfacauca.

En su trabajo de investigación para obtener el título de máster, Enio Walid Ghobar, (2017) tiene como objetivo principal utilizan un conjunto de datos sobre preferencias de los usuarios, y aplicar de Machine Learning con el fin de construir un sistema de recomendación, a partir de las preferencias de los usuarios, los investigadores proponen utilizar un método híbrido que combina los filtrados colaborativos, y que obtiene como resultado una solución sencilla y de bajo coste, con respecto a otros sistemas de recomendación ya existentes y analizados previamente en el proyecto, también se establece como un valor agregado el poder hacer recomendaciones sin requerir muchos datos de entrenamiento.

Teniendo en cuenta el anterior proyecto de investigación analizado y el enfoque que maneja se empieza a considerar dentro del proyecto de investigación en desarrollo la opción de realizar la recomendación, bajo la construcción de un método de recomendación hibrida que utiliza las técnicas de filtrado colaborativo, que a su vez se identifican como una solución óptima para realizar recomendaciones a partir de datos que se centran en la opción o actualización de datos para los usuarios, también se tiene en cuenta del anterior proyecto las técnicas de agrupamiento que se utilizaron para crear perfiles principales y secundarios que se puedan identificar dentro del

conjunto de datos y a partir de esto basarse en el tipo de recomendación a realizar dentro de la ejecución del sistema de recomendación, otro punto importante que se tiene en cuenta en este proyecto, es la definición de una recomendación alternativa para casos de perfiles que no se cuente con su opinión completa o para los casos de “comienzo en frío”. Por último y muy crucial a destacar en este proyecto es la descripción de la muestra de aproximación de recomendaciones que brinda una métrica estadística y un conjunto de datos experimental para determinar la validez de las recomendaciones ejecutadas.

Javier Díaz, (2016) propone mayor énfasis en el análisis y aplicación de diferentes técnicas no supervisadas y de técnicas de visualización de datos masivos, con el objetivo de identificar las características más relevantes de los alumnos de la facultad de informática de la Universidad Nacional La Plata, Argentina. A su vez, se espera poder contribuir en el área educativa a través de la determinación de perfiles dinámicos de los alumnos de esta facultad, en lo que se refiere a su interacción con recursos educativos de acceso libre y su interacción con las redes sociales. Dichos perfiles podrán ser utilizados para caracterizar su comportamiento actual y asistirlo en forma automática, a fin de que puedan alcanzar el comportamiento esperado. Del anterior proyecto citado, es importante destacar la minería de datos masivos, considerando también las fuentes de texto libre y las no estructuradas, que en diferentes respuestas del formulario que se va a utilizar para el análisis de datos de este proyecto, correspondan a gran cantidad de textos. Las líneas de investigación recomendadas en este artículo tienen que ver en su mayoría con: estudio de técnicas de agrupamiento aplicables a datos masivos; estudio de distintas técnicas de

procesamiento aplicables a minería de textos; estudio, análisis y comparación de diferentes técnicas de visualización para grandes volúmenes de datos; revisión y análisis de técnicas específicas de Learning Analytics, las cuales son importantes para tener en consideración en el desarrollo de la investigación, sobre la generación de posibles soluciones a problemas en la etapa de análisis de datos.

En su investigación, Ubillús, (2016) tiene como objetivo diseñar un modelo de evaluación para la pertinencia de maestrías en ingeniería y aplicarlo a un caso concreto. Se trata de maestrías que ya se encuentran activas donde se pretende diseñar un modelo que defina el concepto de pertinencia de una maestría en ingeniería de sistemas, donde se haga una revisión bibliográfica y consultando a expertos en los temas. Se identifican dos tipos de pertinencia local y global que a su vez tienen otras dimensiones que involucran la pertinencia social, relacionadas con las necesidades de la comunidad local, regional y nacional.

Esta investigación aporta una serie de conceptos importantes que tienen que ver a su vez con las ofertas de los posgrados, donde se pretenden usar las técnicas de minería de datos, pero en primera instancia el levantamiento de los datos y el filtrado de estos debe estar ligado a la pertinencia ya mencionada. Esta investigación proporciona una perspectiva de cómo analizar el concepto de pertinencia de la educación superior y determinar características como el contexto socioeconómico, los modelos de evaluación de los programas de posgrado y en general los indicadores objetivos que cubran las necesidades e intereses de los estudiantes y las partes interesadas en la oferta académica en programas de posgrado.

Según la el trabajo de grado realizado por Silva, (2015) uno de los proyectos escogidos para el desarrollo de esta investigación, busca la implementación de un formulario para la estructura de presentación de los proyectos de rediseño en la oferta académica vigente a nuevas ofertas en cuanto a pregrado, propuesta por la comisión permanente de universidades y escuelas politécnicas del CES. Con el desarrollo y la implementación de este formulario se podrá obtener información precisa de las carreras y equipamiento con los que cuenta la universidad para tomar decisiones y ejecutar proyecciones para los años siguientes. Teniendo en cuenta su objetivo, se consideran importantes las estructuras del modelo de base datos para el prototipo del formulario, que brindará información relevante en este proyecto ofreciendo datos más precisos y necesarios para un proyecto de rediseño de oferta académica vigente y nuevas ofertas, en este caso en el ámbito de posgrado.

La investigación expuesta por David Luis la Red Martinez, (2015) propone la utilización de técnicas de minería de datos sobre información del desempeño de los alumnos de la asignatura mencionada, con el propósito de caracterizar los perfiles de alumnos con rendimiento académico y de aquellos que no tienen ese buen rendimiento. La determinación de estos perfiles permitirá a un futuro definir acciones específicas interesadas a revertir el bajo rendimiento académico, una vez detectadas las variables asociadas del mismo.

En este artículo, se describen los modelos de datos y minería de datos que se utilizaron y se comentan los principales resultados. Aporta de manera detallada la utilización de este modelo que tiene como propósito caracterizar los perfiles, lo cual es una parte primordial en el desarrollo del proyecto a llevar a cabo. Se tiene en cuenta

la utilización de la metodología, es decir, cuál fue el modelo usado para determinación de perfiles según el rendimiento académico. Para la realización de la siguiente investigación sobre la obtención de datos, se tendrán en cuenta las tablas de datos proporcionadas a través de formularios de preguntas que respondieron los egresados, previamente caracterizadas dándole un ámbito preciso a la aplicación del modelo a implementar.

En su investigación, Leon lipe, (2013) propone la gestión de información de a través de la técnica de agrupamiento (clustering), empleando técnicas de minería de datos, para egresados del programa de ingeniería de sistemas de la Universidad Católica de Santa María. Esta gestión está constituida por un tratamiento de datos que busca tener una mayor obtención de conocimiento, el cual se inicia con la selección de datos de diferentes fuentes. Se desarrolla un proceso de extracción, transformación y carga de los datos de los egresados, continuando con el filtrado, la segmentación y agrupación de la información con características y patrones similares que van a permitir conocimientos básicos para hacer un análisis y obtener una buena toma de decisiones que beneficié al programa de ingeniería de sistemas.

En el proyecto encargado por ProCalidad educación superior elaborado por la empresa consultora Enacción SAC, con su nombre “Diseño de un sistema de seguimiento de egresados y una estrategia para la implementación de dicho sistema” propone, la realización de un sistema de información Web para el seguimiento de egresados teniendo en cuenta la bolsa de trabajo, encuestas y reportes, donde se plantea el escenario actual del manejo actual que se le da al seguimiento de los egresados es de manera manual e involucra procesos engorrosos de forma digital, es

por este que se plantea el uso de las tecnologías de información actuales para sistematizar y dar mejora a los procesos de actualización de datos de los egresados, generación de reportes entre otros.

Uno de los temas principales que te identificaron previamente para la búsqueda investigativa se refiere al manejo que se le da a los datos que involucran a los egresados y todos los procesos en que estos se ven intervenidos, es por eso que este proyecto tiene conceptos de relevancia y pertinencia de la oferta académica pos gradual y la determinación de diferentes consideraciones a tener en cuenta para la óptima realización de este proceso, también el mapeo de los procesos y las oportunidades de mejoras identificadas en el anterior proyecto sirven como base metodológica y conceptual para abordar el tema de la actualización del seguimiento de los datos a los egresados como un análisis y seguido de esto una propuesta de mejora en la concepción de modelo a desarrollar.

2.2 Marco conceptual.

2.2.1 Inteligencia de datos.

Actualmente, la generación de datos en las organizaciones ha creado la necesidad de aprovecharlos de manera efectiva en diferentes procesos que involucren al desarrollo y evolución de las organizaciones, en este caso, una corporación educativa de educación superior, que busca tener una fidelización para sus miembros, que son en su gran mayoría estudiante y egresados. Para esto, son necesarios estudios que revelan patrones o algoritmos para obtener un mayor beneficio, por lo que es necesario hablar de inteligencia de datos.

La inteligencia de datos involucra un conjunto de técnicas y tecnologías que permiten explorar grandes volúmenes de datos que pueden ser, por un lado, de manera automática, es decir, la obtención de resultados de manera instantánea que dependen de unos requisitos específicos; y, por otro lado, se pueden alcanzar de manera semiautomática, es decir, que se utilizan ciertas técnicas para la creación de nuevos algoritmos y que estos se desarrollen de manera enfocada en las condiciones que se le den. La finalidad de estas técnicas y tecnologías como se nombró anteriormente es identificar patrones que expliquen el comportamiento de los datos, que permitan llegar a conclusiones y poder transformar los datos en información relevante para el desarrollo de las organizaciones, donde se pueden identificar mejoras y soluciones que contribuyan de manera activa en la consecución de los objetivos del negocio.

En el siguiente proyecto se usaron y se analizaron los resultados de diferentes técnicas y tecnologías que permitieron cumplir con los objetivos del proyecto, a continuaciones se describe el concepto de las técnicas y tecnologías usadas para un mejor análisis y proyección de lo planteado para esta investigación.

2.2.2 Minería de datos.

Según Juan Miguel Moine, (2011) la minería de datos es el proceso de extraer conocimiento útil, comprensible y novedoso de grandes volúmenes de datos, con el objetivo de encontrar información oculta o implícita que no es posible obtener mediante métodos estadísticos convencionales. En un proceso que puede estar involucrado en diferentes escenarios donde se plantea un problema o una oportunidad, cuya finalidad es buscar patrones y relaciones que puedan llegar a una solución de las mismas, como

también escenarios donde se aproveche el análisis de los datos y obtener nuevos conocimientos.

Actualmente, este proceso se utiliza particularmente en el ámbito empresarial y organizacional. Todo esto teniendo en cuenta la aplicación de algoritmos específicos que pueden llegar a identificar patrones y tendencias para predecir comportamientos futuros que están ocultos y generar modelos que permitan analizar el estado actual del escenario, al cual se le quiere aplicar el proceso. Bedoya (2016) revela que los patrones mencionados pueden ser grupos diferenciales en los registros (análisis de clúster), registros inusuales (detecciones de anomalías) y dependencias entre datos (reglas de asociación).

2.2.3 Descubrimientos en conocimientos de bases de datos (Knowledge Discovery in Databases-KDD).

La minería de datos hace parte del proceso de descubrimiento en conocimientos de bases de datos. El proceso descrito en la imagen 4, se basa en hacer uso de los diferentes algoritmos que permitan adquirir conocimiento a partir de un volumen de datos, teniendo en cuenta la especificación de parámetros adecuados. Este proceso involucra en una de sus etapas la minería de datos con el fin de obtener mejor comprensión, por lo que se lleva a cabo de forma interactiva y repetitiva, es decir, que los actores involucrados, según el contexto, deben hacer parte de todo el proceso, ya que son los que determinan el enfoque del proceso y delimitan los datos a utilizar. A continuación, se presentan las etapas del proceso de extracción de conocimiento:

- Selección de datos
- Minería de datos

- Técnicas de evaluación y mejora de modelos
- Difusión y uso del conocimiento extraído

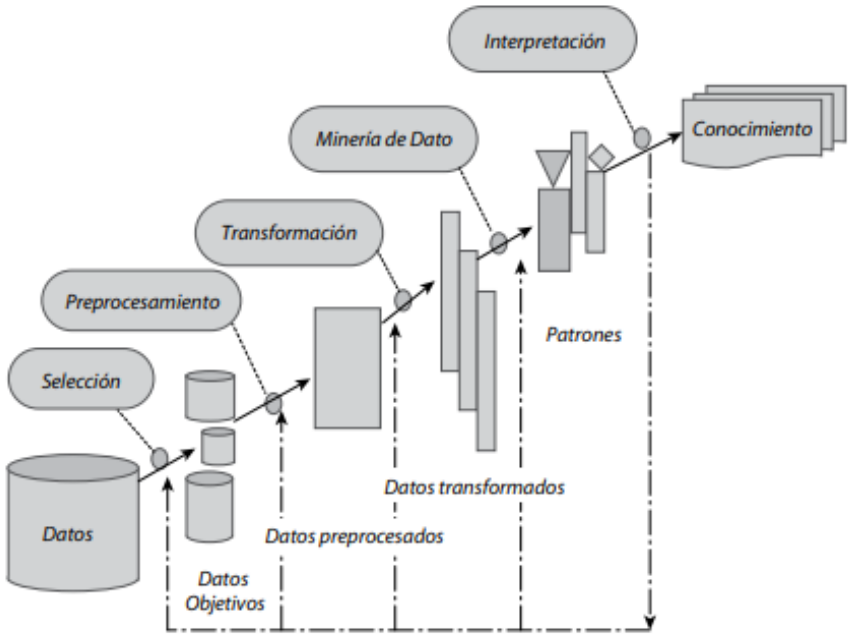


Imagen 2. Timarán-Pereira, (2016) El proceso de descubrimiento de bases de datos.

2.2.4 Ciencia de los datos (Data science).

En el proceso de minería de datos se utilizan técnicas ya creadas, con el fin de descubrir de manera automatizada información útil sin necesidad de definir un problema u oportunidad previa. Estas técnicas se encuentran en constante evolución gracias al avance tecnológico y la información en varios campos que abarcan las bases de datos. Entre las áreas más destacadas dentro de la investigación se encuentran: reconocimiento de patrones, estadística, inteligencia artificial, sistemas expertos, recuperación de información entre otros. También se tiene en cuenta las tareas de

minería de datos, donde cada una de ellas tiene sus propios requisitos y cada información obtenida puede ser muy distinta a la otra. Estas tareas pueden ser predictivas o descriptivas; las predictivas contienen la clasificación y la regresión, y las descriptivas se encuentran con las reglas de asociación y correlaciones.

2.2.5 Técnicas y algoritmos de minería de datos.

En el proceso de minería de datos se utilizan técnicas ya creadas, con el fin de descubrir de manera automatizada información útil sin necesidad de definir un problema u oportunidad previa. Estas técnicas se encuentran en constante evolución gracias al avance tecnológico y la información en varios campos que abarcan las bases de datos. Entre las áreas más destacadas dentro de la investigación se encuentran: reconocimiento de patrones, estadística, inteligencia artificial, sistemas expertos, recuperación de información entre otros. También se tiene en cuenta las tareas de minería de datos, donde cada una de ellas tiene sus propios requisitos y cada información obtenida puede ser muy distinta a la otra. Estas tareas pueden ser predictivas o descriptivas; las predictivas contienen la clasificación y la regresión, y las descriptivas se encuentran con las reglas de asociación y correlaciones.

2.2.6 Algoritmo J48.

Carlos Hernán Cardona Taborda, (2016) indican que el algoritmo J48, es un algoritmo que hace parte de lo que se conoce como árboles de decisión, estos entran en lo que se conoce como clasificación supervisada, en otras palabras, se tiene una variable dependiente o una clase, el algoritmo va a determinar el valor de la clase para casos nuevos. Este va a recorrer desde su nodo raíz hasta que se terminen los nodos hijos y este analiza cada uno de ellos hasta que cada uno de ellos tenga o contenga la misma

clase. Por otro lado, Rodolfo Mosquera, (2016) señalan que el algoritmo J48 es un clasificador de tipo árbol que consiste en generar gráficos a partir de datos, el cual permite una toma de decisiones respecto a las reglas generadas. Por consiguiente, se le conoce como un algoritmo de clasificación estadístico.

2.2.7 Clustering.

Schiaffino, (2018) Indica que clustering es un método descriptivo que divide los datos en grupos con objetos similares que a la vez simplifican la información. Esta técnica tiene como característica principal su naturaleza flexible que permite que se combine con otras técnicas y esto genera sistemas híbridos. Para Lorente-Leyva, (2018) La técnica clustering es capaz de crear grupos y diferenciarlos entre otros dentro del clúster, esto quiere decir, que las observaciones que pertenecen a un grupo están muy cercanas entre sí y están apartadas de las observaciones que están en otro clúster. Como técnicas de clustering existen cuatro tipos: 1) jerárquicos; 2) basados en redes; 3) basados en densidad; 4) particionamiento. Por ejemplo, el K-means, es una técnica de particionamiento que proporciona la creación de un conjunto de clústeres; sin embargo, por lo general su escala no es favorable para los que tienen grandes conjuntos de datos.

2.2.8 Algoritmo K-NN.

Castro, Coelho y Farias (2019) Aseguran que KNN se utiliza para la clasificación de objetos y su operación se relaciona con la determinación de K vecinos más cercanos de un objeto dado para fines de clasificación. Este enfoque tiene tres elementos clave: un conjunto de registros etiquetados en capacitación básica; una métrica para calcular la distancia entre registros; y el número de vecinos más

cercanos. Para la clasificación de un nuevo conjunto de datos sin un identificador, se calculan las distancias entre el nuevo conjunto de datos y los conjuntos de datos marcados de la base de entrenamiento, Se identifica los próximos vecinos del nuevo conjunto de datos y sus identificadores son usados como parámetros para etiquetar el nuevo elemento. KNN determina que el punto con la distancia más corta representa la clase de salida óptima. Para este propósito, se presenta una base de datos de entrenamiento al algoritmo, que consta de atributos. Entre los cálculos, la distancia euclidiana se destaca como una de las formas más utilizadas en KNN.

Riquelme., (2013) aseguran que el algoritmo de KNN es La búsqueda en los sistemas modernos de recuperación de información tiene como objetivo encontrar objetos relevantes sobre la base de la similitud o la distancia a un objeto de consulta en particular. En este contexto, la búsqueda de los vecinos más cercanos es un enfoque generalizado que busca los objetos más cercanos a la consulta en función de una medida de distancia, por ejemplo, minería de datos, aprendizaje automático y algoritmos de optimización. Entre las características que los distinguen de otros métodos se encuentran: formación rápida y sencilla; Robustez frente a datos de entrenamiento ruidosos En este sentido, una técnica que se utiliza a menudo para la búsqueda de KNN es el algoritmo de vecino más cercano K. El algoritmo de KNN es un método de clasificación clásico con diversas aplicaciones en áreas como el reconocimiento de patrones y la minería de datos; La clasificación KNN ha demostrado su eficacia en aplicaciones de reconocimiento de patrones estadísticos.

2.2.9 Sistema de recomendación baso en casos CBR.

Quiroz Martínez, (2020) Un sistema de recomendación, Es un algoritmo que obtiene información sobre preferencias de sus usuarios y ayuda a los usuarios a identificar la información de aprendizaje más interesante de un grupo mucho más grande de información, Los sistemas de recomendación están basados en técnicas como el filtrado colaborativo, contenido o híbrido, Tos los sistemas de recomendaciones necesitan de una información sobre los usuarios y objetos de aprendizaje para poder así brindar una información de calidad.

Vences-Nava Rodrigo, (2019) Los sistemas de recomendaciones es bueno que se obtengan o es recomendable que se tenga información previa para que así mismo el usuario se dé cuenta de que la información que brinda este algoritmo sea razonable con sus gustos, esto aumentaría la confiabilidad para las recomendaciones de los ítems desconocidos.

2.2.10 Mayoría pondera (weidthing majority).

El algoritmo de mayoría ponderada, propuesto en el año 1994 por Littlestone y Warmuth, ha sido utilizado en diferentes investigaciones por un conjunto de enfoques que se basan en la estrategia y metodología de este algoritmo. Consiste, principalmente, en combinar N predictores o clasificadores a menudo llamados *expertos* que comienzan con un peso, pero se reducen siempre que uno de ellos se anuncie incorrectamente. Seguidamente, para hacer una descripción general, el método toma un voto ponderado de las predicciones de expertos y predice la clase con mayor peso, es decir, que este algoritmo tiene la capacidad de ajustar sus ponderaciones con el tiempo y cada vez que se crea un nuevo clasificado, se elimina el más antiguo.

2.2.11 Votación por mayoría ponderada (weidthing majority voting).

De acuerdo con J. Zico Kolter, (2007) el algoritmo mantiene un enfoque del algoritmo weidthing majority, sin embargo, el aprendizaje en este algoritmo implica la combinación de predicciones hechas por los diferentes clasificadores, en este se determinan los pesos según las prestaciones de cada de uno de los clasificadores, al final, la predicción para cada instancia se realiza en función de los votos ponderados. Aquí es importante mencionar que, se existen tres etapas en este algoritmo: la primera, es capacitar a los clasificadores en el conjunto de entrenamiento; la segunda es determinar los pesos de los clasificadores que utilizan el conjunto de validación, en esta etapa el clasificador genera una decisión que apunta a la etiqueta de clase predicha de una instancia única y luego se evalúan para actualizar sus pesos y la última fase consiste en combinar las salidas de los clasificadores individuales considerando por sus pesos.

2.2.12 Distancia euclidiana.

Torres, (2019) argumenta que la distancia euclidiana tiene una propiedad llamada euclidiana que está contenida por un espacio euclidio, el espacio euclidio como tal es un espacio vectorial que tiene definido un producto interno, el producto interno es un número que siempre va a salir real y positivo y los valores de la distancia euclidiana son propios de una matriz que no producen valores negativos. En este mismo sentido Tabaghi, Dokmaníc y Vetterli (2020) manifiesta que la distancia euclidiana es una matriz que ayuda para la localización de distancias, las matrices de la distancia euclidiana son para algoritmos basados en programación semi-definida el cual se utiliza para resolver lo que son problemas de localización de trayectorias y los modelos de trayectorias elegidas son para aproximaciones en las variables, es decir que la distancia euclidiana construye una

matriz de confusión con el fin de comparar los valores de dos conjuntos de datos para determinar su similitud.

2.2.13 Dejar uno por fuera (Leave one out of cross-validation).

Gelman, (2018) declara que leave out of cross validation es la validación cruzada de un conjunto de datos, este conjunto de datos se dividen en dos que son conjunto de entrenamiento y conjunto de prueba. El conjunto de entrenamiento es para ajustar el modelo y el conjunto de prueba se emplea para evaluar él Dataset en cuanto a su productividad del modelo. La validación cruzada se afirma que cuando dos modelos candidatos tienen la misma información del conjunto de entrenamiento este beneficia mucho a la información que sale del resultado del algoritmo.

Así mismo Rodrigo, (2021) plantea que la validación cruzada, se ajusta con frecuencia para ajustar los parámetros de los modelos como por ejemplo los vecinos más cercanos y este se utiliza para la validación final del modelo de datos o Dataset, La validación más utilizada es para evaluar la capacidad de generalización de un modelo predictivo y por ende evitar el sobre ajuste; después de esto se aplica a todo el conjunto de datos que es el conjunto de aprendizaje. También dice que este algoritmo es un sub-muestreo aleatorio repetido, pero como tal al realizar el muestreo de datos no hay manera de que se superpongan los conjuntos de datos de prueba.

2.2.14 Matriz de confusión.

(Francisco Javier Ariza-López, 2017) Expresan que las matrices de confusión son la forma más común y estándar de informar sobre la precisión de productos derivados de la clasificación de datos, es una tabla de contingencia que sirve como herramienta estadística para el análisis de observaciones. Como estándar para informar sobre la

precisión temática de cualquier producto de datos, por supuesto, esta misma herramienta se puede utilizar para evaluar la calidad temática de cualquier tipo de datos. Una matriz de confusión es un conjunto de valores que tienen en cuenta el grado de similitud entre observaciones: un conjunto de datos bajo control, la matriz de confusión contiene las cantidades correspondientes a los elementos bien clasificados. Una matriz de confusión le brinda una vista completa de la distribución de acordes y errores en clases, pero es difícil de administrar de una manera simple, por lo que hay índices derivados diferentes para resumir su información por un valor o por un conjunto reducido por valores.

Salla., et al. (2018) denominan que la matriz de confusión en la validación, robusta a cualquier distribución de datos y tipo de relación, evalúa rigurosamente la validez y proporciona información adicional sobre el tipo y las fuentes del error, como media, mínima, máxima y varianza del error o la magnitud del error. Cuando se dispone de clasificaciones discretas subyacentes, un enfoque alternativo es validar la precisión de la clasificación con matrices de confusión. El método de la matriz de confusión puede revelar más errores a medida que analiza mediciones más detalladas. El análisis de la matriz de confusión, incluso si el dispositivo mide solo unas pocas clases principales. Esto proporciona información valiosa sobre situaciones propensas a errores que deben abordarse en un mayor desarrollo del dispositivo. Cómo toda la información disponible sobre el comportamiento real se puede utilizar en el análisis de matriz de confusión, aunque el dispositivo mide solamente unas pocas clases principales. Esto proporciona información relevante sobre las situaciones propensas a errores que deben abordarse.

2.2.15 Algoritmos de aprendizaje.

Vite Cevallos & Carvajal Romero, (2020) revelan que los algoritmos de aprendizaje como soporte para el entrenamiento de modelos que reaccionan a la predicción de datos en diversos campos de la ciencia, utiliza métodos supervisados y no supervisados, como soporte para los procesos de análisis de datos. La fase de predicción de datos es el resultado de entrenar el modelo y analizar los parámetros que permiten que los datos sean más rigurosos. Su estructura requiere el análisis de varios algoritmos que permiten evaluar un conjunto de datos para determinar si se trata de un problema de clasificación o de regresión. Para resolver problemas de predicción, se pretende entrenarlos con diferentes algoritmos que permitan al modelo predecir el nuevo conjunto, empleando técnicas de métodos supervisados que permitieron clasificar los datos del estudio, selección del algoritmo del árbol de decisión, que clasifica correctamente la información y facilita la predicción del comportamiento.

2.2.16 Algoritmos de aprendizaje supervisado.

Lerache, barrionuevo y sattolo (2020) denotan que un algoritmo de aprendizaje supervisado. Se llama supervisado porque necesita un conjunto de datos previamente marcados y clasificados para su entrenamiento. Sobre este conjunto de datos conocido como "conjunto de datos de entrenamiento", el algoritmo hace predicciones y las compara con las etiquetas, con el error recibido, y ajusta el modelo a través de iteraciones sucesivas, logrando así un aprendizaje mucho más progresivo. Hay una variedad de algoritmos con estas propiedades, tales como regresión lineal, regresión logística, redes neuronales, máquina de soporte de vectores, K vecinos más cercanos. Este algoritmo

de clasificación permite estimar, a través de un clasificador, la probabilidad de que un nuevo ejemplo pertenezca a una clase.

2.3 Herramientas.

2.3.1 Weka.

Según Eibe Frank, (2016) denomina que Weka es una herramienta de código abierto y está disponible para los programas de Windows y Linux, contiene básicamente varios algoritmos de aprendizaje automático que están implementados en java, y se establecen en tres categorías: clasificación, regresión y agrupamiento. La cantidad de algoritmo que soporta Weka son setenta y seis para clasificación y regresión que son para resolver problemas de clasificación; ocho algoritmos de agrupamiento y tres algoritmos de reglas de asociación. Esta herramienta es favorable, puesto que la comparación de algoritmos y el desarrollo de investigaciones científicas, además, pueden ser utilizados para identificar cuál de los algoritmos funciona mejor sobre un conjunto de datos. En el mismo sentido Espinoza Mina, (2018) plantea que Weka, es una herramienta de código abierto el cual brinda algoritmos de aprendizaje automático para tareas de minería de datos, contiene herramientas para la clasificación, regresión y clustering. Weka es muy utilizado en la minería de datos y varios estudios confirman que Weka ayuda a darle una perspectiva diferente a los múltiples procesos de control y operación.

2.3.2 Anaconda.

Kadiyala y Kumar (2017) declaran que Anaconda es una librería de Python para solucionar problemas de ciencia de datos, es una plataforma de código abierto que facilita el uso de lenguajes de programación para el procesamiento de datos y análisis predictivo. Esta, además, es una de las plataformas de ciencia de datos más populares

y se puede instalar desde su repositorio. Así mismo Merette., et al (2020) plantea que anaconda es un software gratuito que proporciona herramientas para el diseño de investigación y ciencia de datos, brinda también acceso a varios entornos de programación como, por ejemplo: python. Anaconda va muy bien con el IDE de python lo cual una vez que este tenga información con python no tendrá problema en transferir esta información a otro IDE.

2.3.3 Jupyter notebook.

Merette., et al (2020) menciona que Jupyter notbook es una IDE, este es ejecutado en el navegador predeterminado que facilita que los bloques de código se pueden ejecutar por separado, esta es una gran ventaja, ya que permite utilizar diferentes tipos de textos en el mismo IDE, por lo que las visualizaciones, salida de códigos y los diferentes tipos de operaciones se pueden usar en el mismo lugar. Jupiter notbook es fundamental para programar en Python y realizar análisis de datos. Como está basado en la web es estéticamente muy atractivo y los resultados son de manera organizada.

Según Alejandro Benito Santos, (2018) Jupyter notebook es una herramienta indispensable para el trabajo de ciencia de datos, esto permite que ejecute las tareas típicas en lo que es el análisis de datos, como lo son: importación, exportación, manipulación, transformación, visualización, creación de modelos estadísticos y de aprendizaje automático. Ayuda a documentar y facilitar la comprensión de algoritmos, ya que los resultados parciales y el resultado final denominados como notebook contienen una serie de piezas de códigos y notas, esto permite que se puedan difundir y puedan ser validados por terceros en distintos entornos. Jupyter notebook es de software libre y

es una aplicación web y es escrita en python, esto permite que se puedan difundir y puedan ser validados por terceros en distintos entornos.

2.4 Metodologías utilizadas.

2.4.1 Investigación-acción.

Gladys Patricia Guevara Alban, (2020) afirman que la investigación acción es una integración de conocimiento y la acción, esto hace que los usuarios se involucren, conozcan y se adentren en la realidad del objeto de estudio, para darle solución al problema, los usuarios de esta metodología son los propios actores para generar lo que son los cambios y las transformaciones definitivas, esta metodología es un proceso educativo y es para aprendizaje colectivo, también es para que los usuarios de esta misma tengan oportunidad de utilizar técnicas de investigación y técnicas para recoger información y saber aprovechar los resultados para el beneficio de la solución al problema.

(Ángela García Pérez, 2018) asegura que la metodología investigación acción es un método con eficacia para la asociación de problemas de diferentes características y así mismo como es cualitativa en su desarrollo aborda todo el problema y hace que los usuarios de esta metodología aprovechen al máximo todas características, puesto que los usuarios pueden experimentar todo el conocimiento que esta brinda, toda transformación va a producir un cambio el cual toda la iniciativa y cambio va a generar un nuevo conocimiento en la investigación y esto genera una nueva innovación al momento de encontrar una solución al problema y aplicar la metodología de investigación acción.

2.4.2 Metodología CRIPS_DM.

Según Colombia, https://www.ibm.com/, (2021) CRIPS-DM según sus siglas en inglés Cross-Industry Standard Process for Data Mining es un método para orientar trabajos de minería de datos. El ciclo vital del modelo contiene seis fases y la secuencia de cada una de ellas no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario. El modelo de CRISP-DM es flexible y se puede personalizar fácilmente. En lugar de realizar el modelado, su trabajo se centrará en explorar y visualizar datos para descubrir patrones en datos. Este método permite crear un modelo de minería de datos que se adapte a unas necesidades concretas. En tal situación, las fases de modelado, evaluación y despliegue pueden ser menos relevantes que las fases de preparación y comprensión de datos. Sin embargo, es importante considerar algunas cuestiones que surgen durante fases posteriores para la planificación a largo plazo y objetivos futuros de minería de datos.

2.5 Marco contextual.

2.5.1 Corporación universitaria Unicomfacauca

La Corporación universitaria Comfacauca es una identidad educativa privada, sin ánimo de lucro, fundada en 2002 por la caja de compensación familiar del Cauca, en ese entonces llamándose Instituto tecnológico de Educación superior de Comfacauca ITC, el cual ha venido brindando respuestas a las necesidades del departamento, fomentando la educación vinculada con la cadena productiva de la región y sus necesidades, actualmente es una de las instituciones de educación superior más reconocidas de la región por su alto reconocimiento a sus programas de pregrado, tres de ellos con acreditación de alta calidad, también su vinculación con diferentes instituciones

educativas nacionales e internacionales, es importante mencionar, la amplia oferta de programas de posgrado, como oferta de educación continuada, que atiende principalmente las necesidades de formación de la región. (Comfacauca, 2021). Su sistema de organización y calidad, involucra de manera activa a los egresados de la institución a través de diferentes procesos de comunicación y vinculación, para así evaluar el cumplimiento de los objetivos y la actualización y mejora de estos procesos.

2.5.2 Misión.

Según (Unicomfacauca, 2021) describe la misión como:

La Corporación Universitaria Comfacauca tiene como Misión ofrecer un ambiente de formación integral para el desarrollo profesional de personas que impulsen la productividad en las organizaciones, apoyándose en la docencia, la investigación y la proyección social, en la búsqueda constante de la excelencia, la equidad social y el mejoramiento de las condiciones de vida de los caucanos y colombianos. (p.1)

2.5.3 Visión.

Según (Unicomfacauca, 2021) describe la misión como:

La Corporación Universitaria Comfacauca tiene como Misión ofrecer un ambiente de formación integral para el desarrollo profesional de personas que impulsen la productividad en las organizaciones, apoyándose en la docencia, la investigación y la proyección social, en la búsqueda constante de la excelencia, la equidad social y el mejoramiento de las condiciones de vida de los caucanos y colombianos. (p.1)

2.5.4 Política de calidad.

La institución describe como política de calidad “Propender por la eficiencia y eficacia de nuestros servicios, a través de la mejora continua de nuestros procesos,

fundamentados en la docencia, la investigación y la proyección social, articulados y encaminados a exceder la satisfacción de la comunidad.” (Unicomfacauca, 2021, p.2)

2.5.5 Egresados.

Hasta el momento la corporación universitaria Comfacauca, tiene 4.000 graduados en los diferentes programas Tecnológicos, profesionales y de formación post gradual, desde el año 2004 hasta el año 2021. Hay que destacar que muchos de los programas ofrecidos en los primeros años de fundación de la institución, actualmente no se encuentran ofertados o han sido establecidos como programas académicos de pregrado.

2.5.6 Oficina de egresados y empleabilidad.

En la página de (Unicomfacauca, 2021) se describe lo siguiente:

La Oficina de Egresados y Empleabilidad hace parte de la Dirección de Proyección Social de la Corporación Universitaria Comfacauca, área que tiene como finalidad fortalecer el vínculo de los Graduados Unicomfacaucanos para que exista una relación permanente en beneficio suyo, de la Corporación y la sociedad. El propósito es acompañarlos a lo largo de su vida, favoreciendo su desarrollo personal y profesional.

Una de las herramientas más importantes que utiliza la oficina de egresados y empleabilidad, es la realización de una encuesta semestral, que contiene preguntas relevantes para la información que se desea obtener de los egresados, en temas como, empleabilidad, capacitación, re vinculación, seguimiento, emprendimiento. Que ayudan a generar la información importante para los sistemas de calidad y organización dentro de la institución. (p.1)

2.5.7 Extensión universitaria.

En la página de (Unicomfacauca, 2021) se describe lo siguiente:

> Pensando en el crecimiento personal y profesional de los colombianos, Unicomfacauca desarrolla los siguientes programas pos gradual que permiten la actualización de conocimientos en diversas áreas y la cualificación laboral de empresarios, líder y emprendedor de la región. Así, y con el compromiso de fortalecer dinámicas de desarrollo en el suroccidente del país, Unicomfacauca plantea una parrilla de programas en Especialización y Maestría, tanto propios como en alianza con otras reconocidas Instituciones de Educación Superior nacionales. (p.2)

A continuación, cada uno de los programas en detalle:

- Especialización en Comunicación Organizacional.
- Especialización en Finanzas.
- Especialización en Gerencia del talento Humano.
- Especialización en Gestión de la Cadena de Suministros y Logística.
- Especialización en Sistemas de Información en Gestión y Control de Costos Organizacionales.
- Especialización en Sistemas Inteligentes Aplicados a Internet de las Cosas.
- Especialización en Tecnologías de la Información y Comunicación en Educación.

Maestrías:

- Maestría en Tributación.
- Maestría en Educación desde la Diversidad

2.5.8 Educación continuada.

En la página de (Corporación universitaria Comfacauca, 2021) se describe lo siguiente:

> Promover la Educación Continuada fomenta el sentido de pertenencia de los colaboradores en las empresas y al mismo tiempo optimizar su productividad para asumir con éxito los constantes cambios a los que deben enfrentarse y continuar a la vanguardia del mercado laboral. Por esta razón, Unicomfacauca ha diseñado un portafolio de Cursos, Diplomados, Seminarios y Talleres en las diferentes áreas del conocimiento, contribuyendo así con la formación de profesionales competitivos, creativos e innovadores que aporten al progreso económico, científico, social y cultural de las empresas y la sociedad. A continuación, la oferta actual:
>
> Diplomados:
>
> - Diplomado en Coaching Gerencial.
> - Diplomado en Creación y Gestión de Proyectos Audiovisuales.
> - Diplomado en Docencia Universitaria.
> - Diplomado en E-Commerce.
> - Diplomado en Formación Pedagógica en la primera infancia.
> - Diplomado en Formulación y Evaluación de Proyectos.
> - Diplomado en Gestión Estratégica del Talento Humano.
> - Diplomado en Gestión Pública.
> - Diplomado en Marketing Digital.

- Diplomado en Pensamiento Creativo.
- Diplomado en SAP Bussines one.
- Diplomado en Sistema de Seguridad y Salud en el Trabajo con énfasis en covid-19.
- Diplomado en Social Selling.
- Diplomado en Transformación Digital.

2.5.9 Programa entrénate.

En la página de (Corporación universitaria Comfacauca, 2021) se describe lo siguiente:

> El programa entrénate es un servicio de carácter gratuito dirigido a estudiantes, egresados y graduados con el fin de fortalecer habilidades blandas, competencias y actualización por área del conocimiento. Fue diseñada desde el área de Egresados y Empleabilidad y Proyección Social y Extensión de Unicomfacauca con el ánimo de enriquecer la formación de los Egresados no Graduados y Graduados de la Corporación, esto para mejorar sus condiciones de empleabilidad y competitividad, identificar sus destrezas para impulsar el desarrollo personal y profesional, y propiciar espacios para el intercambio de conocimiento y experiencias entre profesionales. Con este programa se busca:

- Identificar y fortalecer habilidades blandas para impulsar el desarrollo personal y profesional de los Unicomfacaucanos.
- Propiciar espacios para el intercambio de conocimientos y experiencias entre profesionales.

- Fortalecer la red de contactos con perfil profesional para los egresados y graduados de la Corporación.

Esta es una estrategia que busca actualizar en conocimientos a los Graduados de la familia Unicomfacauca y acercar a estudiantes de últimos semestres de nuestros programas académicos a entornos empresariales o de emprendimiento, abordando temas de interés en sus áreas de conocimiento. Así, este curso se dio con el objetivo de mejorar las capacidades de nuestros egresados en perfiles que demandan el tratamiento y la transformación de los datos para obtener información y conocimiento que ayuden a la toma de decisiones a nivel organizacional. De este modo, la inteligencia empresarial o inteligencia de negocios, siendo una habilidad que cada día gana más protagonismo a nivel empresarial, y que también tiene un alta, pero a la vez, escasa demanda de perfiles, se ha convertido en un punto de debate entre los profesionales de la ingeniería y empresarios que buscan mejorar su operación apoyada en la toma de decisiones acertadas.

Cursos:

- Curso en Creación de Piezas Publicitarias.
- Curso en Finanzas para pequeños productores agropecuarios.
- Curso de Matemáticas para la Universidad.
- Curso en Metodologías ágiles de Coach Ontológico.
- Curso en Pre-Icfes.

Talleres:

- Taller de Ortografía y Redacción para el entorno Laboral.

Capítulo tercero

3. Metodología

Aranda, (2008) afirma en su artículo, que la importancia de la metodología en los trabajos de investigación para obtener un título de educación superior, los estudiantes universitarios deben producir un trabajo académico con reglas y regulaciones específicas que emanan de cada facultad y reflejan el espíritu de investigación inspirado por la institución a lo largo de sus estudios. Toda investigación tiene una serie de características fundamentales a través de las cuales se extrae información para lograr los objetivos propuestos. Si se trata de una investigación científica, el método es riguroso y técnico, está orientado a ampliar el conocimiento científico y no necesariamente tiene que tener aplicación práctica; la investigación solo puede ser una forma de descubrir una cuestión. La metodología de la investigación no únicamente puede verse como una herramienta de conocimiento para afrontar lo desconocido, sino que puede verse como un paradigma de investigación real desde la construcción del proyecto de investigación, la investigación en sí y la presentación de los resultados.

3.1 Tipo de investigación.

En las metodologías existen 2 clases de metodologías o también se les conoce como rutas las cuales son: cualitativa y cuantitativa.

Teniendo en cuenta a Piza Burgos, (2019) En su artículo argumenta que, las metodologías cualitativas En muchos casos las técnicas y herramientas utilizadas para lograr el objetivo propuesto son los mismos, pero es importante distinguir que, aunque existe una estrecha relación entre ellos, tienen significados diferentes. La metodología

de la investigación cualitativa requiere el reconocimiento de diferentes contextos para poder captar las posibles perspectivas del fenómeno en estudio, y para ello no basta con emplear un solo método, sino la articulación de varios con sus respectivas herramientas o instrumentos, sus ventajas y limitaciones. Depende del investigador decidir cuáles se adaptan mejor a su tema de estudio, para lo cual necesita un amplio conocimiento. La pluralidad metodológica posibilita una visión más global y holística del objeto de estudio, ya que cada método nos ofrece una perspectiva diferente.

La metodología de investigación cualitativa consiste en un grupo de técnicas que utilizan una variedad de herramientas para recopilar datos y construir una teoría informada. El investigador juega un papel fundamental en la elección de las técnicas a utilizar, ya que debe evaluar las características del escenario en el que se realiza la investigación, las características de las personas y las limitaciones de tiempo y recursos que puedan existir. La combinación de métodos y técnicas permite obtener una mayor riqueza y variedad de información. Sus resultados ayudan a establecer la validez. En la medida en que los participantes de la investigación perciban el problema como resuelto y el investigador tenga las actitudes suficientes para recopilar toda la información e interpretar sus sentimientos, aumentan la credibilidad de los resultados.

En este sentido hacia la investigación propuesta es cualitativa, ya que se aborda de una encuesta realizada a egresados y una entrevista a los encargados del área educación continua de la universidad Unicomfacauca, los egresados debieron llenar un formulario el cual fue entregado a sus correos electrónicos, las preguntas que se les hizo a los egresados fueron de manera cualitativa, puesto que eran preguntas abiertas y otras de selección múltiple el cual se describen en opiniones de los egresados, la utilización

de estas dos herramientas para partir con la investigación propuesta concluye en por qué esta proyecto de investigación tiene un enfoque cualitativo en primera instancia.

Lo expuesto por Rodríguez, (2017) Indica en su artículo que La Metodología cuantitativa Puede definirse como un conjunto de procedimientos o reglas generales por las que se investiga el objeto de estudio de la ciencia. Los métodos cuantitativos estudian hechos observables, medibles y replicables donde se utilizan modelos estadísticos con precisión matemática y codificación numérica. La investigación cuantitativa destaca las siguientes características importantes tales como: problema, hipótesis, variables, etc. herramientas que recopilan información y miden variables altamente estructuradas. La combinación de métodos y técnicas permite una mayor riqueza y variedad en la información obtenida. Triangular los resultados, ayuda a lograr la validez. En la medida en que los participantes de la investigación perciban el problema a resolver y el investigador tenga las aptitudes suficientes para recolectar toda la información e interpretar sus sentimientos, contribuirán a la credibilidad de los resultados.

Teniendo en cuenta lo anterior mente mencionada la metodología cuantitativa aplica en este proyecto, ya que al obtener respuestas de lo que fue las entrevistas y formularios se tuvo en cuenta que al no tener toda la información requerida anteriormente se obtuvo por hacer porcentajes en valores nulos, valores no especificados y porcentajes en respuesta y no respuestas. Para obtener valores más exactos de manera cuantificada, también se utilizó la técnica de validación cruzada (cross validation) que da como resultado un porcentaje de fiabilidad en cuento a la ejecución del algoritmo elegido. En la construcción del modelo se necesita de un buen análisis estadístico basaba en los

resultados de las pruebas que proyectaban una calidad en el modelo construido, gracias a estos análisis estadísticos se puede tomar decisiones en el modelo construido.

Citando a Chaves Montero, (2018) Da a conocer en su libro que, Los métodos mixtos se utilizan cada vez con más frecuencia porque se complementan entre sí y, además, generan teorías, aumentan la confiabilidad, validez y comprensión de los resultados. Los exámenes mixtos se pueden realizar en paralelo o secuencialmente, según el objetivo del estudio. Utilizan una metodología mixta en un estudio integrando sistemáticamente métodos cuantitativos y cualitativos en un estudio individual para tener una visión más completa del fenómeno, ya que permiten la comparación de frecuencias, factores y resultados.

Los métodos de investigación mixtos enriquecen la investigación a partir de la triangulación con mayor amplitud, profundidad, variedad, riqueza interpretativa y sentido de compresión. La Implementación de la metodología mixta para lograr una mayor eficiencia investigadora. Con este método mixto se examina con más detalle una situación concreta, ya que los instrumentos de ambos métodos en su cooperación, brindan información que permite comprender la realidad examinada y analizarla para su posterior transformación. La investigación cuantitativa permite justificar necesidades, descubre problemas, los relaciona y cuantifica. Por otro lado, la investigación cualitativa proporciona la base para el contenido examina las causas, caracteriza la operación y enriquece los cambios hipotéticos de solución. Finalmente, se puede constatar que la investigación mixta aumenta la posibilidad de ampliar las dimensiones de la investigación y que la comprensión del fenómeno en estudio es mayor y más profunda. [pag 165, 166 y 182 cap 8]

Por esta razón en este proyecto de investigación se desarrolla una investigación mixta teniendo en cuenta lo anteriormente descrito sobre los tipos de investigación existente basándonos en la opinión de autores de conocimiento y la aplicación de estas en el proyecto con el fin de conseguir los objetivos propuestos.

3.2 Diseño de la metodología.

En la presente investigación se utiliza la investigación-acción, que es una metodología que aporta a diferentes problemáticas, sé específica que requiera solución y que afecta a un determinado problema, sea una asociación, comunidad, escuela o empresa o servicio. “Constituye un método idóneo para emprender cambios en las organizaciones” Roberto Hernández Sampieri, (2020) Es usada por los investigadores que han identificado un problema en su centro de trabajo, en este caso en el alma mater, y desean contribuir a investigar y mejorar o solucionar determinado problema.

En el caso específico del desarrollo de este proyecto se enfocará en la manera práctica, que involucra consigo la metodología cíclica de la investigación-acción:

- Identificación del problema
- Observación
- Reflexión

Además, se empleará como metodología apoyo en el desarrollo de los modelos de datos la metodología CRIPS-DM. Que describe, un Proceso estándar de la industria para la minería de datos. Presenta una guía de referencia en el desarrollo de proyecto de minería de datos. Permite estructurar el proceso en seis fases que a continuación se verán desarrolladas con el objetivo de comprender en profundidad el contexto a estudiar

y permitir tener un diagnóstico acertado teniendo en cuenta el tema y el objetivo a desarrollar. La metodología Investigación acción junto con la metodología CRIP-DM:

Fase 1 - Identificación del problema: que debe ser compartida y acordada de manera consensuada, las preocupaciones posibles o de mayor interés a investigar por parte del grupo que participan en la experiencia de investigación y según la metodología CRIPS-DM se utilizan técnicas para la comprensión del negocio que involucran, determinación del contexto, determinación de objetivos del negocio, objetivos de data mining para una completa compresión del negocio y tener todos los requerimientos establecidos para esta primera fase.

Fase 2 – observación inicial: Se realiza una exploración estado actual donde se investiga la literatura científica con respecto a antecedentes, para posteriormente diseñar un plan de acción general, análisis de los datos esto en cuanto a la investigación, en cuanto al proceso de data Mining en esta fase se realiza la compresión de los datos generando informes que describen el estado y características de los datos que se tienen también se lleva a cabo la preparación de los datos, donde se construye una base de datos, estructurada según criterios de selección para adaptarlos a las técnicas de minería de datos.

Fase 3 – Reflexión-acción: Finalmente, se realiza un proceso de reflexión mediante el cual se pueda hallar sentido a los procesos, los problemas, las limitaciones y las condiciones en los que se ha manifestado toda la acción estratégica del plan ejecutado, este plan de acción debe definir el cambio estratégico que apunte a la mejora factible a alcanzar, expresa las circunstancias que tiene lugar al cambio planteado con el fin de

valorar las acciones y efectos, desde una reflexión crítica se pueda hacer una mejora con el desarrollo cíclico de la metodología.

Previamente a esta reflexión se desarrollan las últimas etapas de la metodología de desarrollo data mining que involucra el modelado más apropiado para el desarrollo del proyecto el método de evaluación del mismo y posteriormente su despliegue e implantación que genera consigo los resultados y los informes de los alcances obtenidos.

3.3 Actividades y resultados relacionados en el desarrollo del proyecto.

Tabla 1. Actividades y resultados en el desarrollo del proyecto

OBJETIVO	ACTIVIDAD	INSTRUMENTO	PRODUCTO A OBTENER
Caracterizar la información necesaria que permita la identificación de los perfiles de los egresados.	Identificación del problema	• Estudio e indagación del problema. • Investigación de literatura científica con respecto a antecedentes. •Criterios de indagación y recolección de datos según metodología CRISP-DM.	Detección y diagnóstico del problema a investigar.
	Exploración inicial del estado actual		
	Comprensión del negocio		Objetivos del negocio, objetivos de data mining y sus respectivos criterios de éxito.
	Realizar el plan de proyecto		Plan de proyecto.
Generar un modelo de datos que permita identificar los programas de extensión según los perfiles de egresados.	Comprensión de los datos	• Identificación de calidad, estructura y características según metodología CRISP-DM.	Dataset aplica técnicas de minería de datos.
	Exploración de los datos		

	Construcción de Modelo data mining	• Técnicas de minería de datos para construcción de "Dataset". • Técnicas de modelado más apropiadas para el proyecto.	Modelo de técnica de data mining para identificar y perfilar egresados según criterios del proyecto.
Evaluar el modelo de datos obtenidos para garantizar la mejora a los programas de extensión para los egresados o prospectos.	Evaluación del modelo. Despliegue e implantación del modelo Reflexión	•Método de evaluación de modelo, despliegue y presentación de resultados según metodología CRISP-DM. • Reflexión y acción.	Comprobación de los resultados, despliegue y validación del modelo según criterios del proyecto.

3.4 Aplicación de la metodología CRIPS-DM aplicada para la identificación de perfiles de egresados y sus prospectos en la oferta académica de posgrado.

Dentro de las metodologías y técnicas utilizadas por parte de expertos en proyectos de tratamiento inteligente de datos, según Galan, V. (2015). CRISP-DM (Cross Industry Standard Process for Data Mining) se describe como la metodología más completa y más utilizada para el óptimo desarrollo de los proyectos de inteligencia de datos, esta metodología está estructurada en seis fases, como se describe en la imagen 3: Comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implantación, cada fase está compuesta por un conjunto de tareas y actividades. Algunas de las fases son bidireccionales, es decir que de una fase específica se puede volver a una fase anterior para poder revisar. Esta metodología se aplicará en el desarrollo del proyecto para alcanzar los objetivos propuestos gracias a su

guía general para el desarrollo del trabajo de inteligencia de datos que se pretende realizar dentro del proyecto.

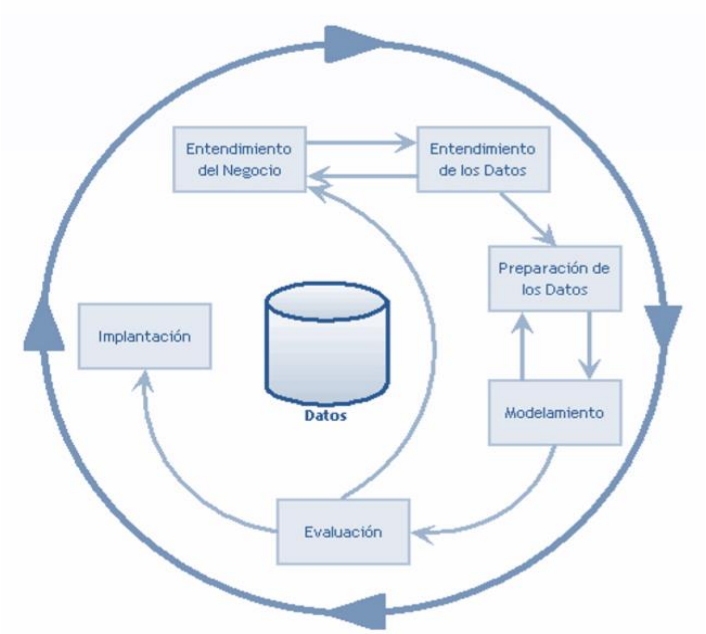


Imagen 3. Ciclos del proceso CRIPS-DM.Galan, (2015). Secuencia del proceso CRISP-DM [figura 4]

3.4.1 FASE 1 Comprensión del negocio:

La comprensión del negocio se divide en las siguientes cuatro tareas o subprocesos:

3.4.2 Contexto.

Actualmente, la corporación universitaria Comfacauca cuenta con un aproximado de cuatro mil graduados hasta el momento desde su inicio como alma mater, estos egresados hacen parte fundamental en la mejora de los procesos de educación para la universidad, para esto la universidad realiza semestralmente una encuesta a los egresados realizada a su posterior graduación, estas encuestas han venido cambiando su estructura a lo largo de los años y en diferentes tablas de Excel que alojan las

respuestas de los egresados, también se cuenta con una base de datos proporcionada a partir el estudio hecho a través de una encuesta realizada por un externo que arroja información importante sobre los egresados.

Sin embargo, no se cuenta con ningún estudio en profundidad que pueda detallar el comportamiento de los egresados y sus prospectos para así se puedan hacer conclusiones, a través de la identificación de patrones o predicciones que aporten para la re vinculación de los egresados, la actualización de la oferta académica por parte del área de mercadeo y el fortalecimiento en conocimientos y la educación integral de los egresados la corporación universitaria Comfacauca. Se desea ejecutar un aprovechamiento de los datos que actualmente se tienen de los egresados de la corporación universitaria Unicomfacauca, Para esto se pretende actualizar la oferta académica, y la capacitación de los egresados en las necesidades y/o falencias que se hayan identificado en diferentes temas de aprendizaje, esto a través del programa “entrénate”, que ofrece la corporación para los egresados. Se cree que bajo la aplicación de minería de datos se puede aprovechar los datos que se tienen de los egresados y llegar a obtener información relevante, para el área de egresados, calidad, mercadeo, bienestar institucional y empleabilidad de la universidad.

3.4.3 Objetivos del negocio.

El objetivo del negocio está enfocado en construir un modelo de identificación de perfiles con una precisión de un 80% o más, que permita a su vez identificar los prospectos de los egresados para así ofrecer los programas de extensión más propicios según los perfiles encontrados con el fin de aplicar la estrategia para la re vinculación de los egresados.

3.4.4 Criterio de éxito del negocio.

Realizar aportes para la actualización de la oferta académica en programas de extensión para los egresados, como aportes al programa entrénate en los temas de interés para el óptimo desarrollo del programa, esto se realizará de una manera enfocada gracias a la identificación de los prospectos de cada egresado y las necesidades de uno. De esta manera lograr la re vinculación de los egresados y ejecutar un óptimo tratamiento de los datos que se tienen de estos.

3.4.5 Valoración de la situación.

Bajo los aportes que se lograron obtener luego de una entrevista con la persona encargada del área de egresados y empleabilidad, se pudo conocer que los datos de los egresados son de vital importancia para los procesos de calidad y mejora institucional para la corporación, puesto que a través de estos procesos se puede llevar una estadística que beneficia tanto al egresado como a la corporación universitaria.

3.4.6 Inventario de recursos.

Actualmente, se cuenta con una base de datos que consta de 4 tablas Excel que contienen las respuestas a las encuestas por parte de los egresados desde el año 2016 al 2021 donde hay un número mayor a mil doscientos registros y 19 campos de interés enfocado en los objetivos del negocio, cabe destacar que las preguntas de las encuestas han variado al transcurrir de los años y donde actualmente se cuenta con un formulario actualizado que resume los campos que puedan arrojar información de mayor interés.

Se contará con datos de prueba obtenidos a través del software de análisis de datos Power BI de Microsoft. Para la construcción del modelo se emplea el software Weka, una herramienta de minería de datos ampliamente usada y aprobada a nivel internacional. El

lenguaje de programación Python será utilizado para desarrollar scripts de inteligencia de datos para lograr el alcance del proyecto.

Los recursos hardware con los cuales se dispone son los siguientes con sus correspondientes características:

PC1.

- Marca: Hp
- Modelo: HP All-in-one
- Procesador: AMD A6-7310 APU with AMD Radeon R4 Graphics
- Memoria RAM: 4,00 GB (3,43 GB)
- Tipo de sistema: Sistema operativo de 64 bits, procesador x64

PC2.

- Marca: Lenovo
- Modelo: Lenovo i300
- Procesador: Intel Core i5
- Memoria RAM: 4.00 GB procesador x64
- Tipo de sistema: Ubuntu 20.04 LTE

En cuanto a la construcción de los modelos se utilizarán las siguientes herramientas software, como librerías y entornos que se ven involucrados en la construcción y ejecución de los modelos.

- Librería pandas para python.
- Librería numpy para pyhthon.
- Weka.

- Anaconda entornó.

- Jupyter notebook.

3.4.7 Riesgos y contingencias

Si bien se cuenta con una base de datos amplia que está conformada por 4 tablas de Excel con las respuestas de los egresados a los diferentes formularios que se han realizado desde el año 2016 al 2021, esta información no está siendo tratada, para esto es necesario construir un conjunto de datos, para obtener la información que se requiere con el fin de alcanzar los objetivos. Se espera obtener un modelo de datos basado en inteligencia de datos con un 80% o más de precisión que permita la identificación de perfiles y los prospectos de estos.

Tabla 2. Tabla de riesgos y contingencias.

Riesgo	**Descripción**	**Contingencias**
Mal entendimiento de los datos	Si se realiza un mal entendimiento de los datos esto puede causar problemas y retrasos dentro del desarrollo del proyecto, como también se podría obtener resultados incoherentes con los objetivos y alcance del proyecto.	Ejecutar entrevistas con expertos en el entendimiento de los datos y las necesidades que estos involucran.
Optima Selección de datos	Si bien se cuenta con bases de datos similares correspondientes a los formularios de satisfacción realizados a los egresados desde el año 2016 al año 2021, los campos no son similares dado que, cada base de datos cuenta con un número, correspondiente a los	Para esto se ejecutó una investigación en primer lugar de comparar y filtrar los campos similares en las dichas bases de datos. En segundo punto, se realizó un segundo filtro correspondiente a los intereses y objetivos de negocio del proyecto, dejando así solo los campos que aportan al proyecto.

	campos de las bases de datos.	
Limpieza y calidad de datos	Similar al problema anterior, si bien se cuenta con bases de datos similares correspondientes a los formularios de satisfacción efectuados a los egresados desde el año 2016 al año 2021, los campos no son similares dado que, cada base de datos cuenta con un formato, y categorización diferente.	Para esto se realizó un proceso de categorización para cada campo escogido en el anterior problema descrito.
Fallo en recursos computacionales	La falla de los equipos de cómputo donde se desarrolle el modelo y la construcción del Dataset, podría ocasionar retrasos y problemas de calidad de los datos.	Se tendrán dos copias de la construcción del modelo como de la base de datos a tratar, una alojada en la nube y la segunda en un dispositivo de almacenamiento portátil.
La técnica metaheurística no es la adecuada	Escoger una técnica de inteligencia de datos que no arroje los resultados esperados puede causar el incumplimiento de los objetivos de minería de datos.	Realizar diferentes variantes dentro del desarrollo del modelo para la construcción de una técnica híbrida y que esté alineada con los parámetros correspondientes al alcance de los objetivos.
La cantidad de datos no sean suficientes	Si luego de hacer la selección de los datos la cantidad de datos no es suficiente para realizar una minería de datos, puede causar retrasos en el proyecto y unos resultados no exactos para el alcance de los objetivos.	Se tendrán en cuenta técnicas de simulación de datos en el tema de construcción de Dataset para él Óptimo desarrollo del modelo y su despliegue.

3.4.8 Costos y beneficios

Como aportes del proyecto se podrá actualizar la oferta académica para posgrados y ofrecerle de una manera enfocada a los prospectos de los egresados, que permitirá la re vinculación de los egresados y óptimo aprovechamiento de métodos de mercadeo. Los costos de este proyecto se pueden ver en la tabla 2.

Tabla 3. Tabla de recursos

Tabla/justificación	Recursos		Total
	Asesores	Estudiantes	
Personal	4’320.000	2’725200	7’045.200
Equipo	0	600.000	600.000
Software: java, Python, Linux, open office, Anaconda.	0	0	0
Viajes y salidas de campo	0	0	0
Bibliografía		150.000	150.000
Materiales	0	350.000	350.000
Servicios técnicos	0	450.000	450.000
Publicaciones	0	0	0
Administración	0	0	0
Comunicaciones	750.000	1’800.000	2’550.000
Otros	0	0	0
TOTAL	5’070.000	6’075.200	11’145.200

3.4.9 Objetivo de la minería de datos

Se pretende generar un modelo que recibirá como entrada un conjunto de datos, que corresponde a respuestas específicas de los egresados previamente escogidas bajo los objetivos del negocio, a formularios de satisfacción realizados desde el año 2016 a 2021.

- Se pretende identificar los perfiles de los egresados con una exactitud de un 80% o más de confiabilidad.
- Recomendar según el perfil del egresado previamente identificado, el programa de formación pos gradual o complementario más idóneo a cursar, por parte de un egresado.

3.4.10 Criterios de éxito de minería de datos.

La técnica implementada en el modelo debe identificar los perfiles de los egresados, basándose en los criterios de medición, para esto se utilizaran medidas de evaluación como, la validación cruzada.

3.4.11 Generación de un plan de proyecto.

A continuación se presenta el plan de proyecto basado en la metodología CRISP-DM aplicada al proyecto inteligencia de datos, para dar cumplimiento a sus objetivos, que gracias a su confiabilidad entre los expertos de la materia y pensando en la consecución del alcance del proyecto es la metodología más óptima. La descripción de sus fases de manera breve son las siguientes:

Fase I. Comprensión del negocio: Esta fase contiene la comprensión de los objetivos y requisitos del proyecto desde una perspectiva del negocio, es muy importante convertir este conocimiento adquirido del negocio en un problema de inteligencia de datos para la consecución de un plan preliminar que tenga como objetivo el alcance de los objetivos del negocio.

Fase II. Comprensión de los datos: Esta fase comprende la recopilación inicial de los datos con el fin de conocerlos y familiarizarse con ellos, se identifica su calidad y se identifican las relaciones más evidentes para establecer la primera hipótesis.

Fase III. Preparación de los datos: En esta fase se procede a la preparación de los datos para adaptarlos a las técnicas de inteligencia de datos que se van a emplear en la siguiente fase, para esto se realizan tareas de selección de datos, limpieza de datos, generación de variables adicionales, integración de datos. La fase del modelado está muy relacionada con la preparación de los datos, ya que según la técnica elegida los datos deben ser procesados de una determinada manera.

Fase IV. Modelado: En esta fase se seleccionan las técnicas de modelado más apropiadas para el proyecto esto siguiendo unos criterios necesarios que den como resultado un modelo óptimo.

Comprender: se comprende el reto que allá establecido, todos los miembros realizan aportes acerca de la solución del problema, con la finalidad de estar alineados a partir de la misma base de conocimientos.

Fase V. Evaluación: En esta fase se evalúa el modelo teniendo en cuenta el cumplimiento de los criterios de éxito del problema, se verifica el porcentaje de fiabilidad, y se revisan el proceso para corregir de ser necesario un paso anterior, en esta fase se involucran diferentes herramientas para la interpretación de los resultados.

Fase VI. Despliegue o implementación: En esta fase se transforma el conocimiento obtenido en acciones dentro del proceso del negocio, el analista podrá recomendar acciones basadas en la observación del modelo y sus resultados.

Tabla 4. Cronograma de ejecución de las fases descritas.

METODOLOGÍA CRIPS-DM	FECHAS DIVIDIDA EN MESES							
	abril	mayo	junio	julio	agostó	septiembre	octubre	noviembre
Comprensión del Negocio								
- Determinar objetivos del negocio - Evaluación de la situación - Determinar los objetivos de DM - Realizar plan de proyecto	X	X						
Comprensión de los datos								
- Recolectar los datos iniciales - Descripción de los datos - Exploración de los datos -Verificar la calidad de los datos		X	X	X				
Preparación de los datos								
- Seleccionar los datos - Limpiar los datos - Construir los datos - Integrar los datos - Formateo de los datos			X	X				
Modelado								
- Escoger la técnica de modelado - Generar el plan de prueba - Construir el modelo - Evaluar el modelo					X	X		
Evaluación								
- Evaluar los resultados - Prueba y validación del modelo - Revisar el proceso							X	X
Despliegue								
- Informe final - Análisis y resultados. - Obtención del modelo -Conclusión -Trabajos futuros								X

3.5 FASE 2 Comprensión de los datos:

3.5.1 Recolectar datos iniciales.

Los datos proporcionados para el problema son compartidos por parte del departamento de empleabilidad y egresados de la corporación universitaria Comfacauca, estos datos vienen de 4 tablas de Excel que contienen las respuestas de los egresados a los formularios de evaluación y satisfacción obtenidos desde el año 2016 a 2021. A continuación se describe el conjunto de datos adquiridos y el método para obtenerlos en la tabla 5.

Tabla 5. Conjunto de datos adquiridos

Nombre tablas	**Número de registros**	**Número de campos**	**Método para obtener los datos**
Encuesta egresados 2016-2017	981	182	Formulario encuesta egresados generada, y realizada los años 2016-2017
Encuesta egresados 2018	524	67	Formulario encuesta egresados generada, y realizada el año 2018
Encuesta egresados 2019-2020	1192	49	Formulario encuesta egresados generada y realizada los años 2019-2020
Encuesta egresados 2021	1402	48	Formulario encuesta egresados generada y realizada 2019-2020-2021
Proyecto evaluación de impacto en formación de pregrado en los egresados de la Corporación universitaria Comfacauca, cohortes 2014 a 2018.	389	368	Seguimiento a egresados de la corporación universitaria Comfacauca, para los graduados desde el año 2014-2018

3.5.2 Especificar los criterios de selección.

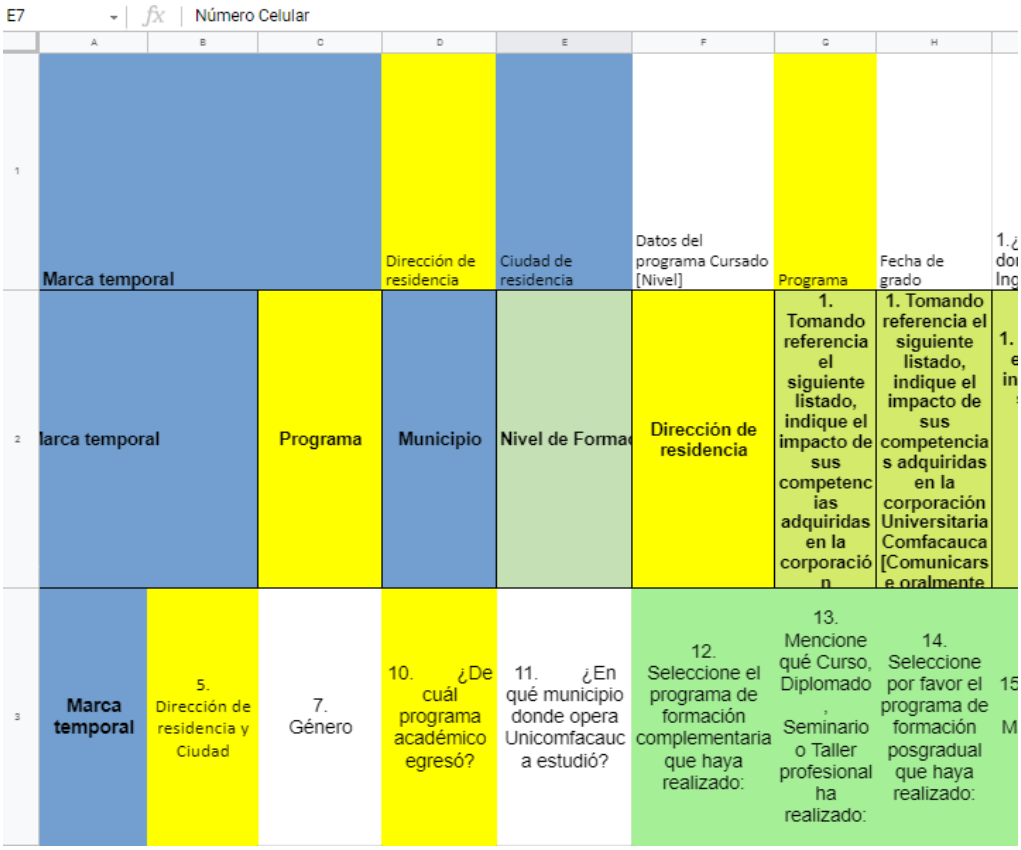
Los criterios de selección están basados en los objetivos de negocio del proyecto que a su vez se basan en dos aspectos importantes para su elección, el primero, la entrevista realizada con el personal encargado de los egresados y empleabilidad de la corporación universitaria Comfacauca. El segundo basado en la investigación realizada a las bases de datos existentes, los problemas identificados y las posibles soluciones encontradas para la construcción de un modelo que cumpla con los objetivos de minería de datos. También es valioso destacar que estos atributos fueron seleccionados para cumplir los objetivos del proyecto en la medida que cumplían con las características suficientes para que proporcionen información importante a través de los datos.

- La Imagen 4. evidencia el primer paso escogiendo todos los atributos o campos involucrados en las tablas de Excel correspondientes a las encuestas realizadas semestralmente desde el año 2016 al 2021.

	A	B	C	D	E	F	G	H	I	J	
1	**Marca temporal**	Nombres y Apellidos Completos	Cédula de Ciudadanía	Lugar de expedición:	Fecha de Nacimiento	Lugar de Nacimiento	Dirección de residencia	Ciudad de residencia	Teléfono fijo	Número Celular	Correo electro
2	**Marca temporal**	**Nombres y Apellidos Completos**	**Cédula**	**Programa**	**Municipio**	**Nivel de Formación [.]**	**Dirección de residencia**	**Departamento de residencia**	**País**	**Teléfono de residencia**	**Cel**
3	**Marca temporal**	1.Nombre y apellidos completos	2.Número de Identificación	3.Número telefónico de celular	4.Dirección de correo electrónico	5.Dirección de residencia y Ciudad	7.Género	10.¿De cuál programa académico egresó?	11.¿En qué municipio donde opera Unicomfacauca estudió?	12.Seleccione el programa de formación complementaria que haya realizado:	13.Me ne qu Cursc Diplor Semir Taller profe ha realiz
4	**Marca temporal**	1.Nombre y apellidos completos	2.Número de Identificación	3.Número telefónico de celular	4.Dirección de correo electrónico	5.Dirección de residencia y Ciudad	7.Género	10.¿De cuál programa académico egresó?	11.¿En qué municipio donde opera Unicomfacauca estudió?	one el programa de formación complementaria que haya realizado:	ne qu Cursc Diplor Semir Taller profe ha realiz

Imagen 4. Selección de campos

- En segundo paso se escogieron los campos de interés para poder identificar características importantes para el análisis de los perfiles. Dejando así, por un lado, los campos que no son necesarios e impertinentes para lograr los objetivos del negocio como se muestra en la Imagen 5.

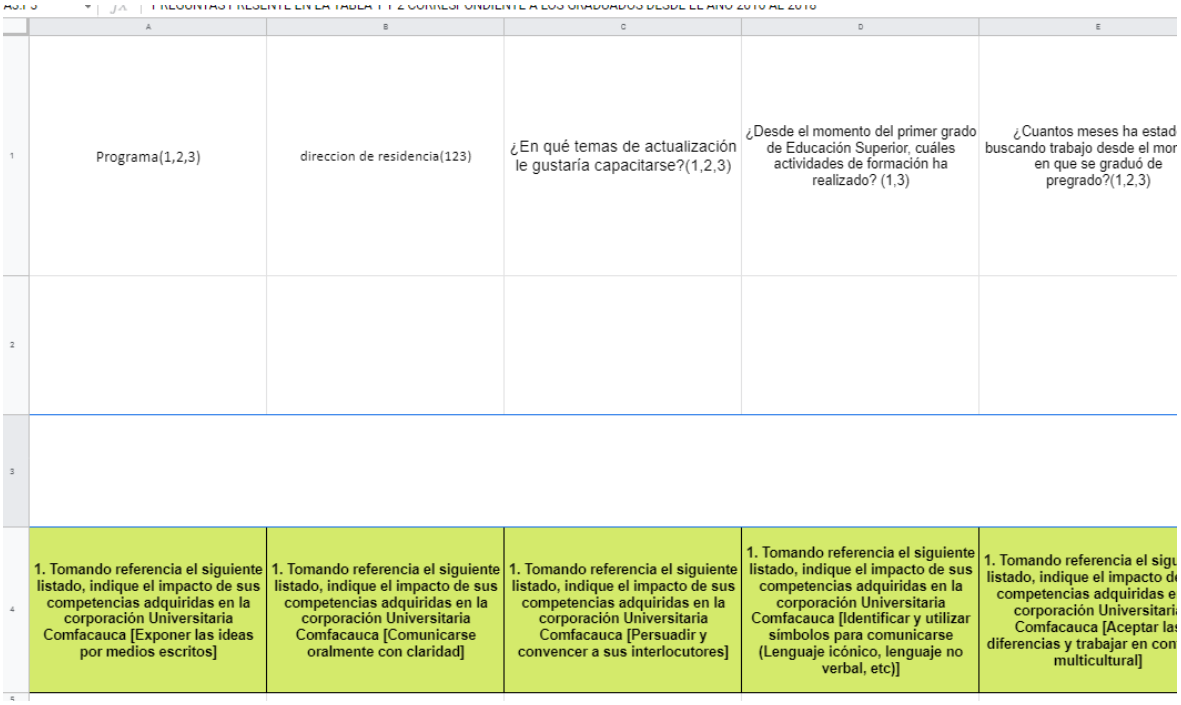


E7 | Número Celular

	A	B	C	D	E	F	G	H	
1	Marca temporal			Dirección de residencia	Ciudad de residencia	Datos del programa Cursado [Nivel]	Programa	Fecha de grado	1.¿ dor Ing
2	larca temporal		Programa	Municipio	Nivel de Forma	Dirección de residencia	1. Tomando referencia el siguiente listado, indique el impacto de sus competencias adquiridas en la corporación	1. Tomando referencia el siguiente listado, indique el impacto de sus competencias adquiridas en la corporación Universitaria Comfacauca [Comunicarse oralmente	1. e in
3	Marca temporal	5. Dirección de residencia y Ciudad	7. Género	10. ¿De cuál programa académico egresó?	11. ¿En qué municipio donde opera Unicomfacauca estudió?	12. Seleccione el programa de formación complementaria que haya realizado:	13. Mencione qué Curso, Diplomado, Seminario o Taller profesional ha realizado:	14. Seleccione por favor el programa de formación posgradual que haya realizado:	15 M

Imagen 5. Filtrado de campos

- Por último se filtró los campos anteriormente escogidos en una tabla nueva tabla de Excel para así dejar de lado los atributos y campos impertinentes que no aportan a la realización del proyecto como se muestra en la imagen 6.



	A	B	C	D	E
1	Programa(1,2,3)	direccion de residencia(123)	¿En qué temas de actualización le gustaría capacitarse?(1,2,3)	¿Desde el momento del primer grado de Educación Superior, cuáles actividades de formación ha realizado? (1,3)	¿Cuantos meses ha estad buscando trabajo desde el mor en que se graduó de pregrado?(1,2,3)
2					
3					
4	1. Tomando referencia el siguiente listado, indique el impacto de sus competencias adquiridas en la corporación Universitaria Comfacauca [Exponer las ideas por medios escritos]	1. Tomando referencia el siguiente listado, indique el impacto de sus competencias adquiridas en la corporación Universitaria Comfacauca [Comunicarse oralmente con claridad]	1. Tomando referencia el siguiente listado, indique el impacto de sus competencias adquiridas en la corporación Universitaria Comfacauca [Persuadir y convencer a sus interlocutores]	1. Tomando referencia el siguiente listado, indique el impacto de sus competencias adquiridas en la corporación Universitaria Comfacauca [Identificar y utilizar símbolos para comunicarse (Lenguaje icónico, lenguaje no verbal, etc)]	1. Tomando referencia el sigu listado, indique el impacto d competencias adquiridas e corporación Universitari Comfacauca [Aceptar las diferencias y trabajar en con multicultural]

Imagen 6. Filtrado de campos

3.5.3 Descripción y exploración de los datos.

En este informe en el primer punto se obtiene a tener muy en cuenta que las 4 bases de datos (encuestas graduados - 2016 y 2017, encuestas graduados 2018, encuesta graduados 2019 y 2020, encuesta 2021.) que brindó la universidad de Unicomfacauca todas son encuestas dirigidas para los egresados, como tal esta es la relación más grande que hay, por otra parte, se tienen preguntas, hay preguntas las cuales se asemejan o son de igual pregunta esas son las relaciones que hay en estas 4 bases de datos. El volumen de estos datos es de 1401 encuestados y las preguntas son 47 en total, pero en realidad las preguntas que interesan y son de utilidad son 16 preguntas. Las preguntas tienen una complejidad y es que tienen datos vacíos (sin responder), ya que algunos egresados no las han respondido.

Tabla 6. Análisis de volumen de datos.

Nombre de los Campos	**Número de registros**	**Número de no respuestas**
Programa	4100	638
¿En qué temas de actualización le gustaría capacitarse?	4100	1806
¿Desde el momento del primer grado de Educación Superior, cuáles actividades de formación ha realizado?	981	2
¿En el futuro, le gustaría cursar otros estudios en la Corporación?	981	2
¿Principalmente, qué otros estudios le gustaría cursar en la Corporación?	1506	2
Seleccione el programa de formación complementaria que haya realizado:	2594	24
¿Qué ha pensado hacer en los próximos 10 años? (múltiple respuesta)	981	2
Mencione qué Curso, Diplomado, Seminario o Taller profesional ha realizado:	2594	1026
Seleccione por favor el programa de formación post gradual que haya realizado	2594	80
Mencione qué Especialización, Maestría o Doctorado ha realizado	2594	1149
Nombre la Universidad donde realizó dicho posgrado mencionado anteriormente:	2594	2008
¿Qué aspectos cree usted que debe mejorar Unicomfacauca para la formación de los próximos profesionales?	2594	20
¿En qué aspectos podría fortalecerse el programa que usted estudió?	981	2

Tabla 7. Comprensión y comprobación de los atributos.

Nombre de los Campos	**Descripción de los campos**	**Categorización.**
Programa	Es el indicativo que referencia de qué programa académico egreso y cuál sería su rama en un posgrado.	Cadena
¿En qué temas de actualización le gustaría capacitarse?	Permite identificar en qué temas está interesado el egresado.	Cadena
¿Desde el momento del primer grado de Educación Superior, cuáles actividades de formación ha realizado?	Saber si los egresados siguieron sus estudios de posgrado o si terminaron la carrera y ya dejaron su proceso de formación.	Cadena
¿En el futuro, le gustaría cursar otros estudios en la Corporación?	Saber si la universidad tiene buena oferta académica o si son del gusto de los egresados.	Cadena
¿Principalmente, qué otros estudios le gustaría cursar en la Corporación?	Saber qué gustos tienen los egresados en cuanto a su posgrado.	Cadena
Seleccione el programa de formación complementaria que haya realizado:	Saber que opciones tienen los egresados y por ende saber qué brindarle en la oferta académica	cadena
¿Qué ha pensado hacer en los próximos 10 años? (múltiple respuesta)	Tener en cuenta la opinión del egresado y saber si quiere ser un posgrado y por ende brindar un posgrado de su agrado.	Cadena
Mencione qué Curso, Diplomado, Seminario o Taller profesional ha realizado:	Saber que opciones tienen los egresados y por ende saber qué brindarle en la oferta académica	cadena
Seleccione por favor el programa de formación pos gradual que haya realizado	Saber si el egresado ha realizado un posgrado y saber si se le puede brindar uno.	Cadena
Mencione qué Especialización, Maestría o Doctorado ha realizado	Saber si el egresado ha realizado un posgrado y saber si se le puede brindar uno.	Cadena

Nombre la Universidad donde realizó dicho posgrado mencionado anteriormente:	Saber el por qué lo hizo allá y por ende mirar los precios y ofertas que se brinda en dicha universidad	cadena
¿Qué aspectos cree usted que debe mejorar Unicomfacauca para la formación de los próximos profesionales?	Saber qué expectativas tienen los egresados con la universidad y por ende saber su opinión y tener en cuenta falencias y su oferta académica	cadena
¿En qué aspectos podría fortalecerse el programa que usted estudió?	Mejorar la formación en cuanto a su educación proyectándose a lo que en realidad le gusta al estudiante y al egresado.	Cadena

3.5.4 Entrevista a experto de dominio

Entrevista 1. Entrevista realizada a gestor encargado de los egresados y la oficina de empleabilidad 2019-2020.

En la entrevista que se le realizó al gestor de egresados, que es el encargado de la oficina de posgrados de la universidad de Unicomfacauca, cuenta que la comunicación con los egresados es muy poca, puesto que muchos de los egresados o no utilizan el número de celular que dejaron o no abren el correo electrónico estudiantil, por otro lado, cuenta que estas encuestas que se le realizan a los egresados la deben tener en cuenta porque los pares académicos están al pendiente de esas encuesta porque les interesa saber si el egresado se encuentra laborando una vez graduados o si el trabajo en el que esta ejerce su profesión; para que los egresados tenga una re vinculación con la universidad, la universidad hace unas encuestas y los que respondan las encuestas entran a sorteos de premios tales como: viajes, celulares, etc.

En el tema de pares académicos cuenta el gestor de egresados que estos tienden a profundizar más en los egresados, puesto que como dije anteriormente el tema laboral

y hacen preguntas como: ¿cuántos están egresados están trabajando? ¿Cuántos egresados están desempleados?

Por otra parte, cuenta que lo más importante para la oficina de egresados es el tema laboral y en qué otros temas les gustaría capacitarse. La universidad tiene convenio con la oficina de empleo y ahí es donde los egresados obtienen la oportunidad de aplicar su hoja de vida. Además, socializa un programa que llamado "entrénate" que está enfocado en ayudar a los egresados en las falencias que tienen en cuanto se desempeñan en su labor.

En un resumen en la universidad lo más considerable es saber en qué está elaborando el egresado en qué campo se desempeña y en cómo la universidad por medio las encuestas puede ayudar a los egresados en el ámbito laboral, por otro lado, también les interesa mucho los temas de posgrados (diplomado, curso, especialización, maestría) porque esto les ayuda a los programas de la universidad abrir esos cursos en la universidad.

3.5.5 Informe de calidad de datos.

Los datos obtenidos en cada campo (en este caso preguntas de las encuestas), en su mayoría son satisfactorias, esto quiere decir que existe una respuesta. Hay que mencionar que de igual manera sus respuestas son coherentes con las preguntas realizadas en los campos, pero también existen errores, algunos de caligrafía como de ortografía. También existen algunos valores omitidos en las respuestas de los campos, sobre todo en las últimas preguntas efectuadas en las encuestas. Asimismo existen una cantidad mínima de respuestas incoherentes a las preguntas representadas no con la misma taxonomía. A continuación se presenta el proceso para identificar los campos

omitidos y los campos en blanco. Se identificaron los campos omitidos o campos en blanco con color azul claro, como se muestra en la siguiente imagen.

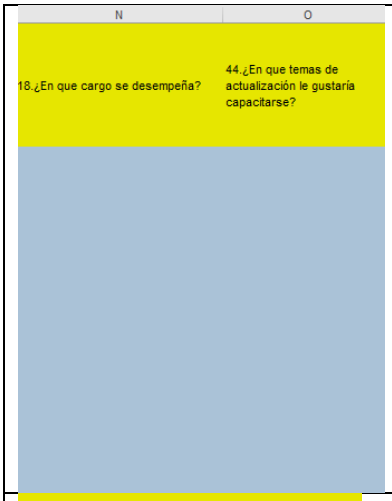
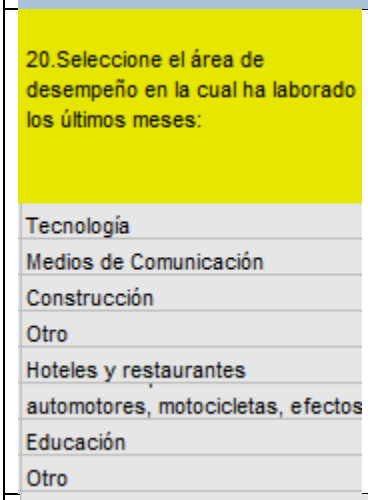
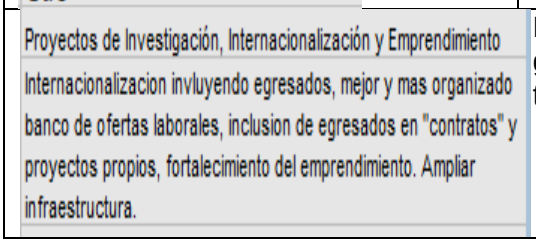
	B	C	D	E	F	G
1	10.¿De cuál programa académico egresó?	12.Seleccione el programa de formación complementaria que haya realizado:	13.Mencione qué Curso, Diplomado, Seminario o Taller profesional ha realizado:	14.Seleccione por favor el programa de formación posgradual que haya realizado:	15.Mencione qué Especialización, Maestría o Doctorado ha realizado:	17.¿Está trabajando actualmente?
2	Ingeniería de Sistemas	Ninguna		Ninguna		No
3	Comunicación Social y Periodismo	Ninguna		Ninguna		Sí
4	Contaduría Pública	Ninguna		Ninguna		Sí
5	Contaduría Pública	Diplomado	Diplomado en niif	Ninguna		Sí
6	Tecnología en Gestión Gastronómi	Diplomado	Diplomado en parrilla y maridaje	Ninguna	No	No
7		Ninguna	datos	Especialización	datos	No
8	Ingeniería de Sistemas	Ninguna		Ninguna		No
9	Comunicación Social y Periodismo	Diplomado	comunicación organizacional	Ninguna		Sí
10	Contaduría Pública	Ninguna		Ninguna		No
11	Ingeniería de Sistemas	Ninguna		Ninguna		Sí
12	Ingeniería Industrial	Curso	de gestión de calidad NTC ISO 9001 e hice	Ninguna		No
13	Tecnología en Comunicación Soci	Diplomado	Diplomado en televidencia crítica	Ninguna		Sí
14	Comunicación Social y Periodismo	Diplomado	Marketing Digital	Ninguna		Sí
15	Comunicación Social y Periodismo	Diplomado	Gestión Estratégica de la Comunicación	Ninguna		No
16	Ingeniería Industrial	Ninguna		Ninguna		No
17	Contaduría Pública	Curso	Administración de recursos humanos	Ninguna		No
18	Ingeniería Industrial	Diplomado	Lean six sigma	Ninguna		Sí
19	Ingeniería Industrial	Ninguna		Ninguna		No
20	Comunicación Social y Periodismo	Diplomado	Revelado y retoque de fotografía	Ninguna		No
21	Ingeniería Industrial	Diplomado	Lean six sigma	Ninguna	Ninguna	Sí
22	Ingeniería de Sistemas	Diplomado	Windows y Linux	Ninguna		Sí
23	Ingeniería de Sistemas	Ninguna		Ninguna		Sí
24	Comunicación Social y Periodismo	Ninguna		Ninguna		Sí
25	Tecnología en Gestión Gastronómi	Diplomado	Diplomado de parrilla y maridaje	Ninguna	Ninguna	No

Imagen 7. Detección campos vacíos

- Lista de los resultados de la verificación de calidad de datos; si existen problemas de calidad, liste las posibles soluciones. Las soluciones a los problemas de calidad de datos generalmente dependen tanto del conocimiento de los datos y como del negocio.

Tabla 8. Falencias encontradas.

Evidencia	Problemas de calidad	Posibles soluciones
13.Mencione qué Curso, Diplomado, Seminario o Taller profesional ha realizado: Diplomado en niif Diplomado en parrilla y maridaje datos comunicación organizacional	Registros omitidos o en blanco.	Teniendo en cuenta la pregunta del campo correspondiente, se deduce que los campos en blanco corresponden a la no realización de educación complementaria realizada, para lo cual se pondrá el valor de "ninguna". De igual manera se sustituirá para las preguntas similares que no tengan respuesta alguna.
21.¿En cuál empresa laboró por última vez? 2018 Corporación Universitaria Comfacauca Aserhi SAS esp Hotel camino real GITCAM	Incoherencia en las respuestas.	Para las respuestas o campos donde la respuesta sea incoherente, se reemplazaran como campos en blanco o nulos.
20.Seleccione el área de desempeño en la cual ha laborado los últimos meses: Tecnología Medios de Comunicación Construcción Otro Hoteles y restaurantes automotores, motocicletas, efectos Educación	Problemas de ortografía y caligrafía en las respuestas.	Para las respuestas que contengan faltas de ortografía y caligrafía, se corregirán estos errores y se hará una clasificación de las respuestas en general para la próxima construcción de una Dataset.

	Gran cantidad de campos vacíos	Para las preguntas en las cuales existe una gran cantidad de campos vacíos u omitidos se determinará la viabilidad de tener en cuenta estos campos y en el caso de ser útiles, se establecerá una respuesta, teniendo en cuenta sus respuestas en los formularios de los anteriores años.
	Respuestas que causan ruido en el análisis de datos.	Se deberá clasificar las respuestas que no aporten un sentido a los datos en este caso “otros”.
	Respuestas con gran cantidad de texto.	Para este tipo dee respuesta se clasificaron por una opinión establecida para tener una coherencia en los datos obtenidos.

3.6 FASE 3 Preparación de los datos:

3.6.1 Selección de los datos.

Inicialmente, se tenía un conjunto de datos el cual contenía todas las variables de las tablas que nos fueron brindadas por el área de egresados de la Universidad Unicomfacauca. Se selecciona un subconjunto de datos considerando la calidad que tienen los datos, el volumen o los tipos de datos que están relacionados con las técnicas

de minería de datos, estos datos fueron seleccionados por criterios previamente definidos que se enfocan en la información académica de los egresados el cual dará las respuestas necesarias para cumplir con los objetivos.

En este conjunto de datos se tuvo en cuenta primero que todo el tipo de variables (preguntas) que se tenían y qué relación se obtenían de ellas, después de una larga observación en la cual con la ayuda de los objetivos se iban a escoger unas variables las cuales iban tomando forma a lo que se quería llegar.

En la siguiente imagen se puede detallar unos colores los cuales ayudaron a la identificación de que variables estaban presentes en las tablas, el color amarillo significaba que esa variable estaba presente en las tres tablas, el color azul significaba que estaba presente esa variable en las 2 primeras tablas y el color verde, significaba que estaba presente en la segunda y tercera tabla.

			Dirección de residencia	Ciudad de residencia		
Programa	Municipio	Nivel de Formación [.]	Dirección de residencia			
		5. Dirección de residencia y Ciudad	7. Género	10. ¿De cuál programa académico egresó?	11. ¿En qué municipio donde opera Unicomfacauca estudió?	12. Seleccione el programa de formación complementaria que haya realizado:

Imagen 8. Relación de variables en las diferentes tablas.

Con este proceso se logra identificar las relaciones entre las tablas y la cantidad de registros que se podían ir obteniendo.

Programa(1,2,3)	direccion de residencia(123)	¿En qué temas de actualización le gustaría capacitarse?(1,2,3)	¿Desde el momento del primer grado de Educación Superior, cuáles actividades de formación ha realizado? (1,3)

Imagen 9. Ejemplo de las relaciones entre las variables en las tablas.

Así de esta manera se identifica una relación y por ende se tenía una mejor visión de los objetivos y de cómo se podía ir llegando a ellos. Teniendo en cuenta los pasos anteriores se hicieron una integración de las tablas con las variables que eran de importancia y las cuales se asemejan a los objetivos del proyecto, así mismo se hizo un conjunto de datos con las variables y registros de todas las tablas que se tenían. A si como se muestra en la siguiente imagen.

Cedula	programa	¿Qué ha pensado hacer en los próximos 10 años?	¿Desde el momento del primer grado de Educación Superior, cuáles actividades de formación ha realizado?	¿En el futuro, le gustaría cursar otros estudios en la Corporación ?	¿Principalmente, qué otros estudios le gustaría cursar en la Corporación ?	Seleccione el programa de formación complementaria que haya realizado	Mencione qué Curso, Diplomado, Seminario o Taller profesional ha realizado	Seleccione por favor el programa de formación posgradual que haya realizado	Mencione qué Especializac ón, Maestría o Doctorado ha realizado
4617757	Tecnología en Maquinaria e Instrumentació n Industrial	Iniciar una nueva carrera universitaria, Trabajar en Colombia	Diplomados	Si me gustaría	Diplomados				
106173316	Ingeniería de Sistemas	Estudiar un posgrado en Colombia, Trabajar en Colombia, Crear una empresa	Especialización	Si me gustaría	Maestría				
106169354	Tecnología en Agroambiental	Estudiar un posgrado fuera de Colombia, Trabajar en Colombia, Crear una empresa	Seminarios/Cursos, Universitarios, Especialización	Si me gustaría	Diplomados				
106171479	Ingeniería de Sistemas	Iniciar una nueva carrera universitaria, Estudiar un posgrado en Unicomfacauca	Tecnológicos	Si me gustaría	Universitario				

Imagen 10. Integración de tablas.

En esta integración de tablas los espacios en blanco son porque los registros son de las bases de datos de las cohortes (2019-2020) y (2021). Aquí ya se sabía el enfoque y los campos que había cierta información que era importante para él proyecto. De esta

manera y con los expertos en el tema ya se podían ir mirando variable por variable cuáles son las más propicias para un buen Dataset.

En la siguiente imagen se va a detallar la integración de unas variables las cuales se unificaron solo en una variable, puesto que las respuestas y la pregunta se podían homologar y quedar solo una variable. Como se puede observar en la imagen los campos que tienen el mismo color son los que se integraron.

¿Desde el momento del primer grado de Educación Superior cuáles actividades de formación ha realizado?	Seleccione el programa de formación complementaria que haya realizado	Seleccione por favor el programa de formación posgradual que haya realizado	¿En el futuro, le gustaría cursar otros estudios en la Corporación?	¿Principalmente, qué otros estudios le gustaría cursar en la Corporación?	Mencione qué Curso, Diplomado, Seminario o Taller profesional ha realizado	Mencione qué Curso, Diplomado, Seminario o Taller profesional ha realizado	Mencione qué Especialización, Maestría o Doctorado ha realizado
Diplomados	null	null	Si me gustaría	Diplomados	null	null	null
Especialización	null	null	Si me gustaría	Maestría	null	null	null
Universitarios	null	null	Si me gustaría	Diplomados	null	null	null
Tecnológicos	null	null	Si me gustaría	Universitario	null	null	null

Imagen 11. Integración de variables.

En la siguiente imagen se va a observar cómo quedó en el conjunto de datos cómo se va reduciendo el número de variables para el conjunto de datos final. Las variables con color son los campos que se integraron y quedaron de esa manera

cedula	programa	¿Que ha pensado hacer en los proximos 10 años?	¿Que ha pensado hacer en los proximos 10 años?	¿Que ha pensado hacer en los proximos 10 años?	seleccione el programa de formacion posgradual o complementario que ha realizado	mencione que formacion complementaria o posgradual realizo	¿en el futuro le gustaria cursar otros estudios en la corporacion?	¿principalmente que otros estudios le gustaria cursar en la corporacion?	nombre la universidad donde realizo dicho posgrado mencionado anteriormente
1061733163	ingenieria de sistemas	estudiar un posgrado en colombia	trabajar en colombia	crear una empresa	especializacion	null	si me gustaria	maestria	null
1061693544	tecnologia agroambiental	estudiar un posgrado fuera de colombia	trabajar en colombia	crear una empresa	especializacion	null	si me gustaria	diplomados	null
1061714795	ingenieria de sistemas	iniciar una nueva carrera universitaria	estudiar un posgrado en unicomfacauca	null	tecnologicos	null	si me gustaria	universitario	null
1061738030	comunicacion social y periodismo	estudiar un posgrado fuera de colombia	trabajar en colombia	estudiar un posgrado en unicomfacauca	maestria	null	si me gustaria	diplomados	null
1061708094	comunicacion social y periodismo	estudiar un posgrado fuera de colombia	null	null	diplomados	null	si me gustaria	especializacion	null

Imagen 12. Integración final de variables.

Después de este proceso se consultó con los expertos en el tema en el cual se volvió hacer una revisión de las variables que se tienen y se llegó a la conclusión que los campos “¿Qué ha pensado hacer en los próximos 10 años?”, “¿en el futuro le gustaría cursar otros estudios en la corporación?”, “nombre de la universidad donde realizó dicho posgrado mencionado anteriormente”, “¿Qué aspectos cree usted que debe mejorar Unicomfacauca para la formación de próximos profesionales?” y “¿En qué aspectos podría fortalecerse el programa que usted estudió?” no eran de suma importancia en consecuente con los objetivos del proyecto y la información que estas variables brindaban no estaban acorde a lo planteado en el proyecto. Basado en la opinión de los expertos se determinó que el Dataset quedaría de esta estructura.

Resultada categorización.

cedula	programa	seleccione el programa de formacion posgradual o complementario que ha realizado	mencione que formacion complementaria o posgradual realizo	¿principalmente que otros estudios le gustaria cursar en la corporacion?	¿en que temas de actualizacion le gustaria capacitarse?
1061733163	ingenieria de sistemas	especializacion	null	maestria	seguriada informatica
1061693544	tecnologia agroambiental	especializacion	null	diplomados	seguridad y salud en el trabajo
1061714795	ingenieria de sistemas	tecnologicos	null	universitario	proyectos
1061738030	comunicacion social y periodismo	maestria	null	diplomados	lectoescritura
1061708094	comunicacion social y periodismo	diplomados	null	especializacion	comunicacion organizacional marketing digital
1061706978	contaduria publica	diplomados	null	especializacion	finanzas
1061729452	ingenieria de sistemas	especializacion	null	especializacion	base de datos
10297490	tecnologia en electricidad	null	null	especializacion	energias renovables
1061746940	comunicacion social y periodismo	null	null	diplomados	en el buen uso de la tecnologia

Imagen 13. Conjunto de datos.

Cuando se pretendía subir este Dataset para así aplicar los diferentes algoritmos, la variable “Cédula” se determinó que no era de importancia porque su campo no brinda una buena información para el proyecto, puesto que es muy irrelevante y se obtuvo que lo mejor era quitarla del Dataset final.

A partir del campo “mencione qué formación complementaria o posgradual realizó” que se eligió como clase principal, se tomó la decisión de reducir el número de clases que habían, puesto que muchas de ellas no estaban vigentes en la universidad y por esa razón hay un total 17 clases que se obtuvieron en este Dataset, el cual las clases que no estaban vigentes se homologaron a ofertas actuales que se tienen en la corporación Universitaria Unicomfacauca. Algunos de los cursos o seminarios que no se pudieron homologar por las vigentes en la universidad son porque no tenían una correlación y por ende no se les efectuó su respectiva homologación.

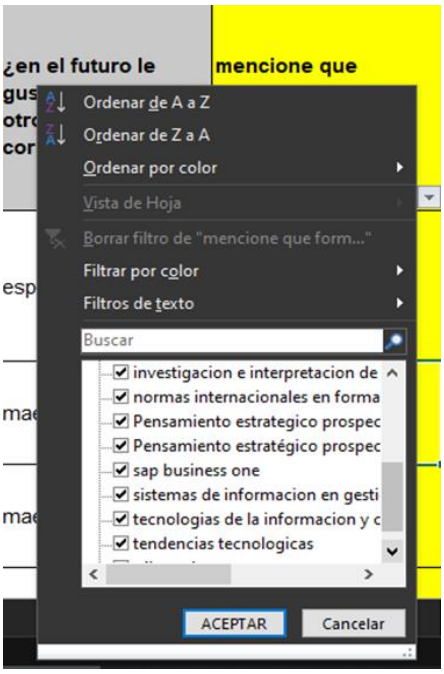


Imagen 14. Tipos de clases seleccionadas y homologadas.

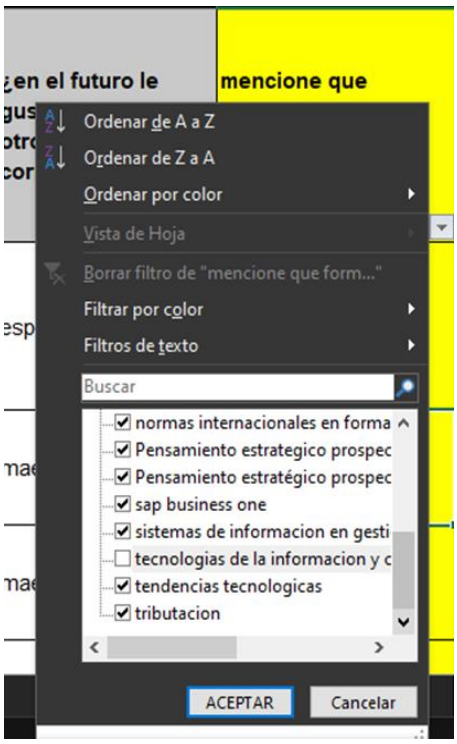


Imagen 15. Tipos de clases seleccionadas y homologadas.

En la siguiente imagen se puede observar el Dataset final con el que se trabajó para la obtención de los modelos de datos. El color amarillo que se observa la clase principal con la cual el proyecto se basó más que todo en esta clase, puesto que el proyecto nos va a recomendar el nombre del programa posgradual a realizar.

programa	seleccione el programa de formacion posgradual o complementario que ha realizado	¿en el futuro le gustaria cursar otros estudios en la corporacion?	mencione que formacion complementaria o posgradual realizo	¿en que temas de actualizacion le gustaria capacitarse?
tecnologia agroambiental	diplomado	especializacion	gerencia politica y gestion administrativa derechos etnicos equidad de genero cultura y paz	investigacion e interpretacion de cultivos y especies productoras
tecnologia agroambiental	especializacion	maestria	investigacion e interpretacion de cultivos y especies productoras	investigacion e interpretacion de cultivos y especies productoras
comunicacion social y periodismo	especializacion	maestria	tecnologias de la informacion y comunicacion en educacion	marketing digital
comunicacion social y periodismo	diplomado	maestria	gestion de innovacion emprendimiento y marketing	marketing digital
ingenieria de sistemas	especializacion	maestria	tecnologias de la informacion y comunicacion en	seguridad informatica

Imagen 16. Dataset final.

3.6.2 Limpiar los datos.

Para optimizar la calidad de los datos se usó la analítica de datos con Excel con la opción remplazar que garantizo una limpieza en el Dataset y su consistencia con la intención de prepararlos para la siguiente fase que es modelación. A continuación se observa a observar lo mencionado anteriormente. Imagen 17 y la imagen 18, reducción del volumen de datos en cuanto a registros, se observa que se repiten por consiguiente se eliminan esos registros.

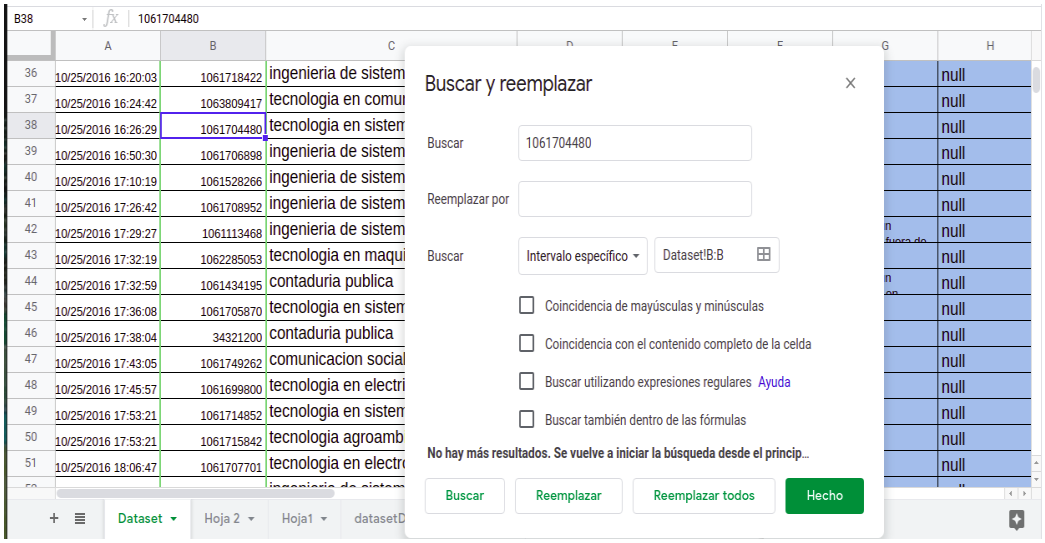


Imagen 17. Reducción de volumen de datos.

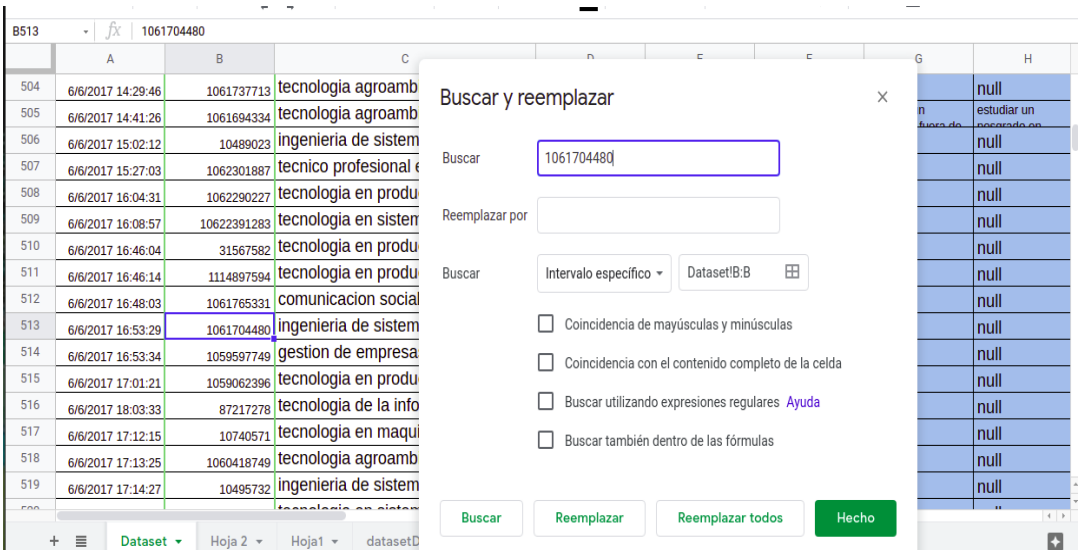
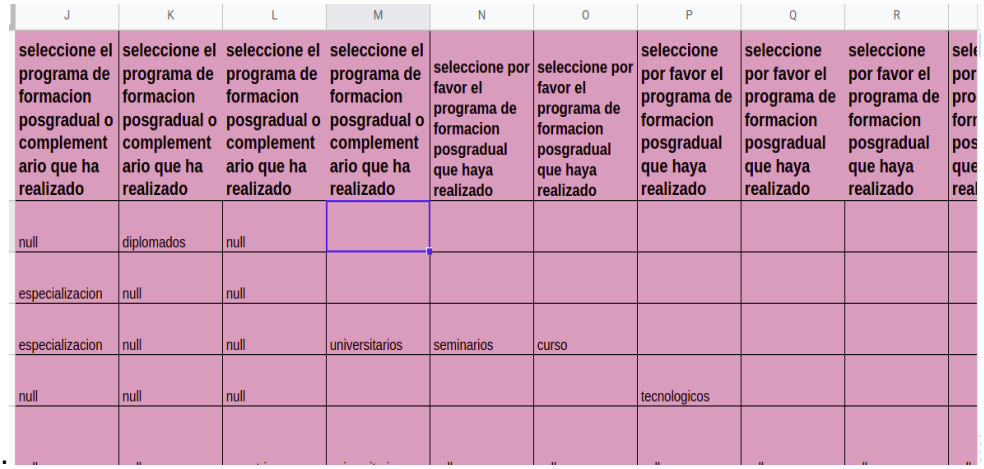


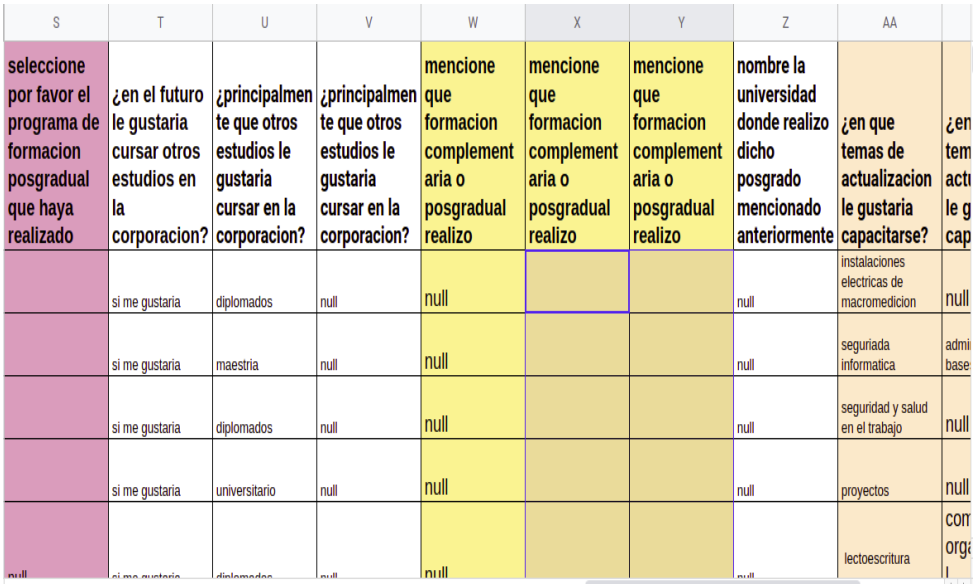
Imagen 18. Reducción de volumen de datos

En la imagen 19 y 20, se observa el tratamiento de valores vacíos a través de la analítica de datos en Excel, donde se toma la decisión de realizar un tratamiento a dichos valores vacíos, remplazándolos por campos nulos.



J	K	L	M	N	O	P	Q	R	
seleccione el programa de formacion posgradual o complementario que ha realizado	seleccione el programa de formacion posgradual o complementario que ha realizado	seleccione el programa de formacion posgradual o complementario que ha realizado	seleccione el programa de formacion posgradual o complementario que ha realizado	seleccione por favor el programa de formacion posgradual que haya realizado	seleccione por favor el programa de formacion posgradual que haya realizado	seleccione por favor el programa de formacion posgradual que haya realizado	seleccione por favor el programa de formacion posgradual que haya realizado	seleccione por favor el programa de formacion posgradual que haya realizado	sel por pro for pos que rea
null	diplomados	null							
especializacion	null	null							
especializacion	null	null	universitarios	seminarios	curso				
null	null	null				tecnologicos			

Imagen 19. Tratamiento de valores vacíos.



S	T	U	V	W	X	Y	Z	AA	
seleccione por favor el programa de formacion posgradual que haya realizado	¿en el futuro le gustaria cursar otros estudios en la corporacion?	¿principalmente que otros estudios le gustaria cursar en la corporacion?	¿principalmente que otros estudios le gustaria cursar en la corporacion?	mencione que formacion complementaria o posgradual realizo	mencione que formacion complementaria o posgradual realizo	mencione que formacion complementaria o posgradual realizo	nombre la universidad donde realizo dicho posgrado mencionado anteriormente	¿en que temas de actualizacion le gustaria capacitarse?	¿en tem act le g cap
	si me gustaria	diplomados	null	null			null	instalaciones electricas de macromedicion	null
	si me gustaria	maestria	null	null			null	seguriada informatica	admi base
	si me gustaria	diplomados	null	null			null	seguridad y salud en el trabajo	null
	si me gustaria	universitario	null	null			null	proyectos	null
null				null				lectoescritura	com orga l

Imagen 20. Tratamiento de valores vacíos.

A continuación, se describe cómo se llevó a cabo el proceso de normalización de los datos en algunos registros que presentan algunas anomalías que pueden generar ruido dentro del análisis de los datos; Un ejemplo de esto se ilustra en las imágenes 21 y 22

donde se puede observar que los nombres de los programas están escritos de una forma abreviada, lo cual no estaría bien, por consiguiente se realiza la normalización de estos datos para corregir de maneras correctas estas y todos los registros que se encuentren con una característica similar.

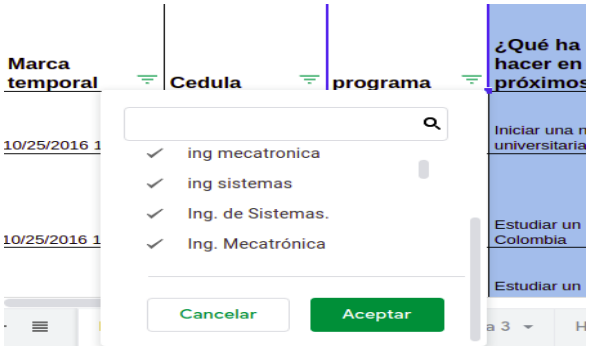


Imagen 21. Normalización de los campos en los registros.

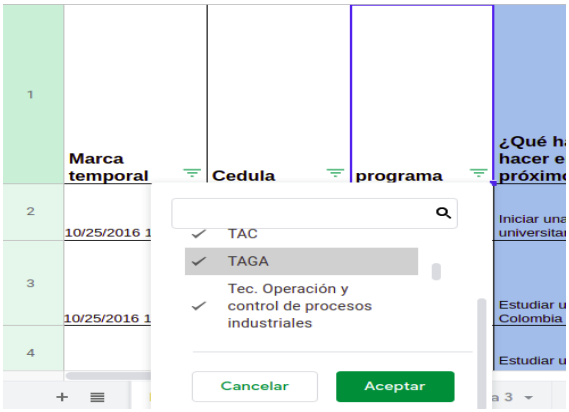


Imagen 22. Normalización de registros

3.6.3 Construcción e Integración de los datos.

La integración de los datos implica la creación de nuevas estructuras a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de

resumen. La generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.

En la siguiente imagen, se puede apreciar que los campos que están de color azul claro, se va a unificar en solo un campo, que se llama “seleccione el programa de formación posgradual o complementario que ha realizado”, se puede detallar que tienen dos campos que son parecidos y en el Dataset quedaría con más columnas, pero con la misma información por consiguiente se toma la decisión de unificar estos campos en una sola columna que brinda la misma información.

M	N	O	P	Q
¿Desde el momento del primer grado de Educación Superior, cuáles actividades de formación	**¿En el futuro, le gustaría cursar otros estudios en la Corporación?**	**¿Principalmente, qué otros estudios le gustaría cursar en la Corporación?**	**Seleccione el programa de formación complementaria que haya realizado**	**Seleccione por favor el programa de formación posgradual que haya realizado**
null	Si me gustaría	Diplomados	null	null
null	Si me gustaría	Maestría	null	null
null	Si me gustaría	Diplomados	null	null
null	Si me gustaría	Universitario	null	null

Imagen 23. Unificación de dos campos para crear un nuevo campo.

En la imagen 24, se observará que las columnas que están de color morado son parecidas, por ende se van a unificar en una solo columna que se llamó “mencione que formación complementaria o posgradual realizo”. Esto con el fin que la información sea

más clara en una sola columna y en el Dataset no quede con más columnas y con la misma información.

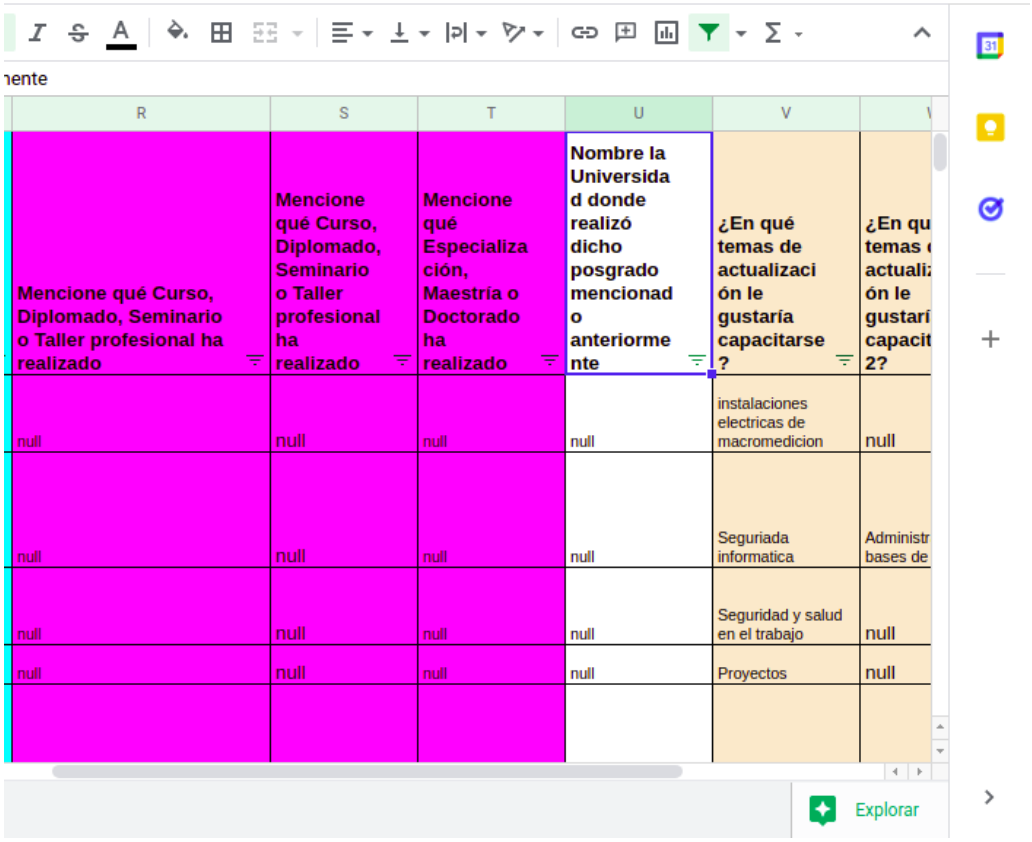


Imagen 24. Unificación de dos campos para crear un nuevo campo.

En la imagen 25, se puede visualizar que hay unas columnas que se llaman “¿Qué ha pensado hacer en los próximos 10 años?”, hablando con los expertos en el tema, se recomienda sacar esta columna del trabajo porque no da una buena información referente este trabajo de grado y en el Dataset no ayudaría con una buena información.

	A	B	C	D	E	F	G	H	I
1	marca		programa	¿Que ha pensado hacer en los proximos 10 años?	¿Que ha pensado hacer en los proximos 10 años?	¿Que ha pensado hacer en los proximos 10 años?	¿Que ha pensado hacer en los proximos 10 años?	¿Que ha pensado hacer en los proximos 10 años?	¿Que ha pensado hacer en los proximos 10 años?
2	10/25/2016 11:38:20	4617757	maquinaria e instrumentacion industrial	iniciar una nueva carrera universitaria	trabajar en colombia	null	null	null	null
3	10/25/2016 14:42:04	1061733163	ingenieria de sistemas	estudiar un posgrado en colombia	trabajar en colombia	crear una empresa	null	null	null
4	10/25/2016 14:45:50	1061693544	tecnologia agroambiental	estudiar un posgrado fuera de colombia	trabajar en colombia	crear una empresa	null	null	null
5	10/25/2016 14:51:52	1061714795	ingenieria de sistemas	iniciar una nueva carrera universitaria	estudiar un posgrado en unicomfacauca	null	null	null	null
6	10/25/2016 14:53:15	1061738030	comunicacion social y periodismo	estudiar un posgrado fuera de colombia	trabajar en colombia	estudiar un posgrado en unicomfacauca	null	null	null
7	10/25/2016 14:54:08	1061708094	comunicacion social y periodismo	estudiar un posgrado fuera de colombia	null	null	null	null	null
8	10/25/2016 14:54:14	1061706978	contaduria publica	estudiar un posgrado en colombia	trabajar en colombia	null	null	null	null

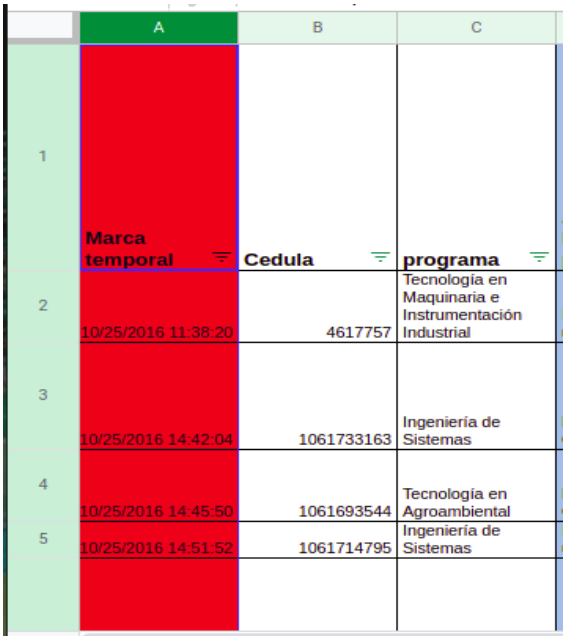
programa Todo

Imagen 25. Eliminación de columnas.

En la imagen 26 y la imagen 27, las columnas en color rojo no brindan información sobre los objetivos del proyecto de grado por ende se llegó a la conclusión que esto sería eliminado, puesto que en el Dataset no sería de utilidad.

W	X	Y	Z	AA
qué as de alizaci e taría acitarse	¿En qué temas de actualizaci ón le gustaría capacitarse 3?	¿En qué temas de actualizaci ón le gustaría capacitarse 4?	¿Qué aspectos cree usted que debe mejorar Unicomfac auca para la formación de	¿En qué aspectos podría fortalecers e el programa que usted estudió?
	null	null	null	difusion
istracion de de datos	null	null	null	Analisis de la informacion, Gestion de bases de datos, Programacion orientada a Objetos
	null	null	null	Mejorar las practicas de laboratorio y salidas de campo
	null	null	null	Actualización
				* Mejorando las estrategias pedagógicas de los docentes

Imagen 26. Eliminación de columnas.



	A	B	C
1	Marca temporal	Cedula	programa
2	10/25/2016 11:38:20	4617757	Tecnología en Maquinaria e Instrumentación Industrial
3	10/25/2016 14:42:04	1061733163	Ingeniería de Sistemas
4	10/25/2016 14:45:50	1061693544	Tecnología en Agroambiental
5	10/25/2016 14:51:52	1061714795	Ingeniería de Sistemas

Imagen 27. Eliminación de columnas.

En la imagen 28, se analizar cómo se crea otro registro en el Dataset, esto se hace porque tiene dos respuestas con el mismo registro y como las respuestas son diferentes entonces se hace el otro registro. Se va a observar en los dos ejemplos que se tienen de color azul y verde.

A	B	C	D	E	F	G
76142404	P 2 3 1 4 2	ingenieria industrial	diplomado	null	lean six sigma	maestria
1061789647	P 1 4 1 3 1 3	ingenieria de sistemas	especializacion	maestria	tecnologias de la informacion y comunicacion en educacion	maestria
1061789647	P 1 4 1 3 5 3	ingenieria de sistemas	seminario	maestria	tendencias tecnologicas	maestria
1061749384	P 1 4 1 3 1	ingenieria de sistemas	especializacion	null	tecnologias de la informacion y comunicacion en educacion	maestria
1061749384	P 1 4 1 3 2	ingenieria de sistemas	diplomado	null	sistemas scada	especializacion
1061814177	P 3 1 4 3 1	ingenieria mecatronica	especializacion	null	tecnologias de la informacion y comunicacion en educacion	maestria
1061814177	P 3 1 4 3 4	ingenieria mecatronica	curso	null	curso cad	especializacion
1061687212	P 3 1 4 3 1	ingenieria mecatronica	especializacion	null	tecnologias de la informacion y comunicacion en educacion	maestria

Imagen 28. Creación de nuevos campos de registros.

3.6.4 Formateo de los datos

Este punto consiste principalmente en la realización de transformaciones sintácticas de los datos sin modificar su significado de tal forma que se permita y se facilite utilizar alguna técnica de minería de datos en concreto, como por ejemplo la reordenación de los campos y/o de los registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación. En las siguientes imágenes se puede analizar cómo estaba la base de datos en cuanto a las letras mayúsculas porque estas hacen ruido a la hora de hacer un proceso de agrupación. En la imagen 29 se puede detallar los campos (programa) que tienen Mayúscula y al tener las letras mayúsculas esto genera ruido al momento de ingresar los datos al Dataset.

Marca temporal	Cedula	programa
10/25/2016 11:38:20	4617757	Tecnología en Maquinaria e Instrumentación Industrial
10/25/2016 14:42:04	1061733163	Ingeniería de Sistemas
10/25/2016 14:45:50	1061693544	Tecnología en Agroambiental
10/25/2016 14:51:52	1061714795	Ingeniería de Sistemas
10/25/2016 14:53:15	1061738030	Comunicación social y periodismo
10/25/2016 14:54:08	1061708094	Comunicación social y periodismo
10/25/2016 14:54:14	1061706978	Contaduría Publica

Imagen 29. Campos con letras mayúsculas.

Se puede observar cómo estaba la base de datos en cuanto a caracteres especiales, esto también hace ruido y se eliminaron esos caracteres, se puede apreciar en la imagen 30. Claro está que en este campo también hay las letras mayúsculas, estas fueron corregidas como en todo el Dataset.

null	Diplomados	null	null	null	null	null	null
Especialización	null	null	null	null	null	null	null
Especialización	null		universitarios	Seminarios/	cursos	null	null
null	null	null	null	null	null	Tecnológicos	null
null	null	Maestría	Universitarios	null	null	null	null
null	Diplomados	null	null	Seminarios/	cursos	null	null
null	Diplomados	null	null	Seminarios/	cursos	null	null
Especialización	Diplomados		null	null	null	Tecnológicos	null
null	null	null	Universitarios	Seminarios/	cursos	null	null

Imagen 30.

En la imagen 31, se puede analizar que hay muchos campos en la base de datos que contienen tildes y por ende también había que corregir esto, porque en el Dataset a la hora de ingresar esta información va a generar ruido y se tendrán errores en el Dataset.

6/5/2019 16:25:5	1062297731	Tecnología en Producción Industrial	null	null	null	null	null	null
6/5/2019 16:38:1	1061776356	Contaduría Pública	null	null	null	null	null	null
6/5/2019 16:40:3	1061813897	Tecnología en Gestión Gastronómica	null	null	null	null	null	null
6/5/2019 16:49:5	1061734723	Contaduría Pública	null	null	null	null	null	null
6/5/2019 17:00:4	1061707701	Ingeniería Mecatrónica	null	null	null	null	null	null
6/5/2019 17:45:5	1061786044	Contaduría Pública	null	null	null	null	null	null
6/5/2019 18:10:5	1062307624	Contaduría Pública	null	null	null	null	null	null
6/5/2019 19:23:3	1061728818	Contaduría Pública	null	null	null	null	null	null
6/5/2019 19:32:2	1061529851	Contaduría Pública	null	null	null	null	null	null
		Tecnología en						

Imagen 31.

En la imagen 32. Se continua haciendo un filtro para mirar los errores tanto de ortografía, caracteres especiales, tildes. Se puede apreciar que hay errores en la ortografía. Esto también hace ruido al momento de subir esta información al Dataset y por ende generaría muchos errores.

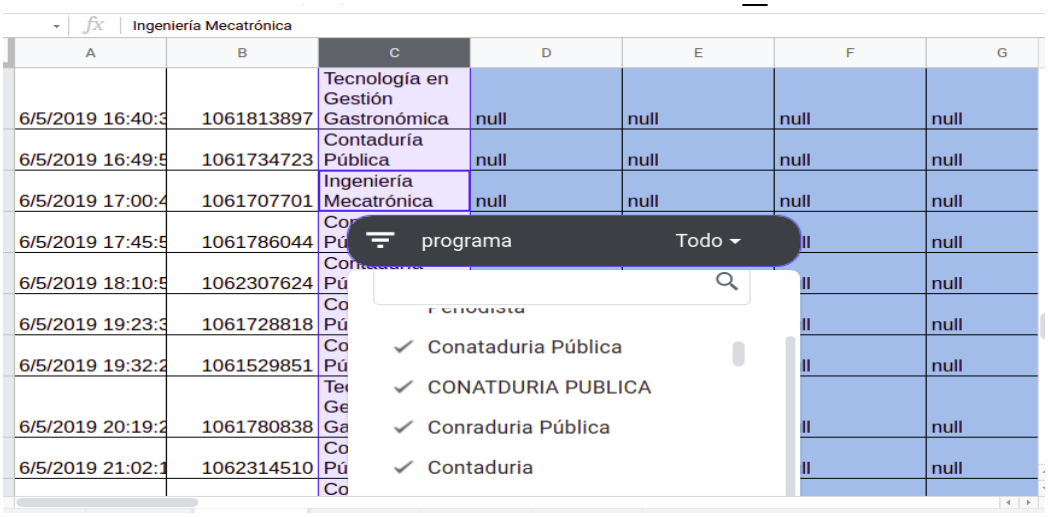


Imagen 32. Corrección de ortografía en los campos.

En la imagen 33, se analiza que tanto errores ortográficos y errores de tildes, esto se va a corregir en solo un dato que sería, “contaduría pública”, esto por si se sube la información al Dataset se tendría muchos errores, puesto que hay muchos campos que tienen estos errores, por ende esto ya no sería un problema en el Dataset cuando lo se vaya a utilizar.

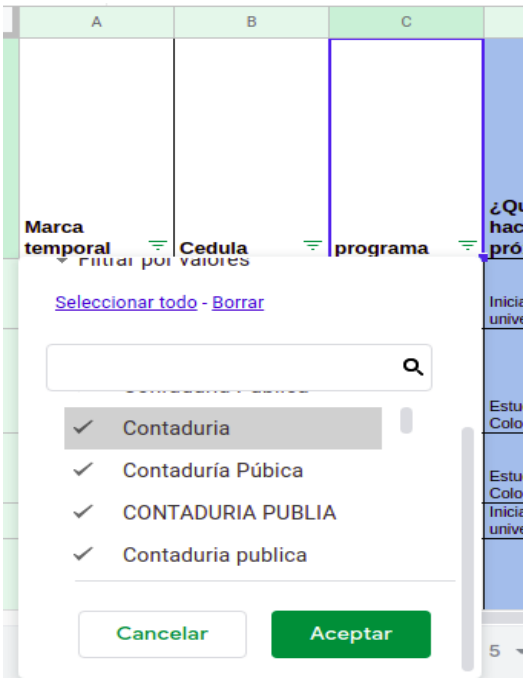


Imagen 33. Corrección de nombre al campo y errores ortográficos.

3.7 FASE 4 Modelado:

3.7.1 Escoger la técnica de modelado.

Esta fase describe la selección de las técnicas de modelado más apropiadas para darle solución a la necesidad que se ha identificado en la investigación de este proyecto. Los criterios de selección de técnicas de modelado más apropiadas para el proyecto se presentan en (Cortina, 2015). Es de gran importancia tener en cuenta el objetivo principal del proyecto y la relación de las diferentes técnicas de inteligencia de datos existentes. A continuación se presenta una matriz de cumplimiento de los criterios de selección de las técnicas de modelado según los objetivos del proyecto. Técnicas de inteligencia de datos escogidas para la consecución del proyecto:

- T1: Clasificación, J48

- T2: Clasificación, KNN
- T3: Filtrado colaborativo Weighting Majoriting
- T4: Filtrado colaborativo más votado Weighting majority voting
- T5: Filtrado basado en conocimiento similaridad con Distancia euclidiana
- T6: filtrado híbrido Similaridad integrado con agrupamiento
- ■: No cumple
- **X** : Si cumple

Tabla 9. Criterios para seleccionar las técnicas

	Ser apropiada para el problema.	Disponer de los datos adecuados.	Cumplir los requisitos del problema.	Tiempo adecuado para obtener un modelo.	Conocimiento de la técnica.
T1	**X**	**X**	■	**X**	■
T2	**X**	**X**	■	**X**	■
T3	**X**	**X**	**X**	**X**	■
T4	**X**	**X**	■	**X**	■
T5	**X**	**X**	■	**X**	■
T6	**X**	**X**	**X**	**X**	**X**

3.7.2 Generar el plan de prueba.

Esta tarea contiene el plan de pruebas pertinentes que verifiquen el cumplimiento de los objetivos del proyecto y la validez del modelo elegido una vez esté construido, teniendo en cuenta esto y los tipos de aplicación de las técnicas de inteligencia de datos que se escogieron, se determinó que la razón de error va a ser la medida de calidad que

le dé validez al modelo probado. Teniendo en cuenta que se eligieron dos herramientas en donde se van a desarrollar los modelos de prueba:

- Aplicación de algoritmos de clasificación a través del software Weka,
- Construcción de un algoritmo de recomendación basado en casos (CBR) en el entorno de desarrollo Jupyter Notebook utilizando la librería panda.

Se genera un plan de prueba para cada modelo de prueba propuesto, este contiene dos criterios de pruebas, criterio de validación y satisfacción, donde en el criterio de validación, como bien se mencionó antes, se basará en los resultados de la razón de error del modelo de prueba, para el caso de la aplicación de los algoritmos de clasificación a través de Weka, y para el sistema de recomendación basado en casos, se empleará la validación cruzada con el algoritmo leave one out of cross-validation. Estas medidas de error las calcula automáticamente Weka al ejecutar los modelos de clasificación escogidos. Para entender mejor estos indicadores se describen a continuación.

Cross-Validation: Esta validación que automáticamente realiza Weka realiza una evaluación cruzada donde se dividirán las instancias en un número de particiones seleccionadas en el parámetro (folds). Técnicamente dado un número n se dividen los datos en n partes y por cada parte se crea un clasificador, con las n=1 partes sobrantes, por último se efectúa una prueba con esta parte, y se desarrolla así de manera repetitiva con cada una de las n partes.

Leave one out-cross validation: Esta validación tiene como principal método dejar uno por fuera para evaluar, es decir, se escoge un subconjunto para validación de una muestra de datos, y el resto del conjunto como conjunto de entrenamiento, es decir,

según el principio de dejar uno por fuera, se deja como sub conjunto una entidad y esta sé válida con todas las entidades restantes, este proceso requiere un aumento en la capacidad de computación para el entrenamiento de los datos establecidos. Gracias a su validación uno a uno se considera un enfoque óptimo que traerá resultados con exactitud.

En cuanto a la decisión de establecer un porcentaje de confiabilidad para la recomendación resultante, especificada en el objetivo general, se tienen en cuenta tres aspectos importantes que se describirán a continuación, esto con el fin de evaluar y presentar un porcentaje de confiabilidad para las recomendaciones resultantes y al final se propone un trabajo futuro encontrado en el apartado 6.3 “Trabajos futuros”.

Fundamento del porcentaje de confiabilidad:

La primera razón y más crucial por la cual se hace referencia al 80% de confiabilidad es por el conjunto de datos que se obtuvo, donde luego de las fases de análisis y de construcciones se evidencia que no es un conjunto de datos totalmente confiable, uno de los análisis resultantes revela que en el atributo clase seleccionado se identificó una preferencia mayor en una clase “tecnologías de la información y comunicación en educación” por esta razón se modificó el objetivo general y se tomó la decisión de establecer un porcentaje de confiabilidad a las recomendaciones. Si bien no se tiene un buen conjunto de datos o uno con calidad de datos confiables, se le da un porcentaje considerablemente alto para un buen resultado de recomendación y se utiliza una métrica “precisión de clasificación” (Classification Accuracy) para de esta manera darle ese porcentaje de confiabilidad a las recomendaciones del modelo escogido.

Como segunda razón se tiene en cuenta un proyecto especificado en el estado del arte, en donde también se emplea una métrica de medición de confiabilidad en un desarrollo de un sistema de recomendación en el cual se explica la métrica y los porcentajes obtenidos en este proyecto, los resultados de la métrica de confiabilidad empleada arrojaron los siguientes resultados 60 y 70% para cada recomendación evaluada, estos fueron considerados muy buenos resultados según Enio Walid Ghobar (2017) Un sistema de Recomendación Basados en Perfiles Generados por Agrupamiento y Asociaciones [magíster]. Universidad Politécnica de Valencia. La métrica empleada en este proyecto se llama Score la cual le da un porcentaje de confiabilidad a los casos probados. En el experimento número 2, de los 571 usuarios han obtenido algún nuevo ítem y su distribución de éxito en la recomendación. Se tiene un perfil 1 de 62% y un perfil 2 con 38%; perfil principal con un 34% y un perfil secundario con 28%. Una breve explicación es Perfil 1: 353 usuarios han obtenido recomendación de un nuevo ítem a través del “Perfil 1”, lo que corresponde al 61.82% de los casos probados. El promedio general del Score ha sido 47.09% y el promedio de los “top 3” salta a 52,53%. Aún basado en este total, podemos verificar que: perfil principal: repitiendo el éxito del experimento 1, la gran mayoría de recomendaciones han sido de 3 ítems o más en casi un 61% de los casos, destacándose el promedio del score del “top 3” de casi 56%, Perfil secundario: Hemos tenido éxito para hacer recomendaciones a través de subgrupos en 158 casos, siendo responsable por 44,75% de las sugerencias del “Perfil 1”, lo que resalta todavía más la efectividad de esta alternativa para la recomendación si comparamos al experimento 1, que ya había sido muy expresiva. Perfil 2: 218 (38,18%) usuarios no recibieron recomendaciones del “Perfil 1”, pero han sido atendidos

completamente por el flujo alternativo del “Perfil 2”, donde un mínimo de 3 ítems nuevos ha sido recomendado a cada uno. También se constata a través de la tabla 10, que el motivo que más ha contribuido para el fracaso en recomendar por el “Perfil 1” corresponde a usuarios que fueron asignados a grupos o subgrupos con ninguna regla de asociación disponible para hacerlo:

Tabla 10. Criterios para seleccionar las técnicas

Motivo	**Cantidad**	**%**
Sin reglas	132	60.55%
Sin recomendación	86	39.45%

El sistema de recomendación que se desarrolló está en un 80% de confiabilidad, Basado en lo anteriormente descrito en el proyecto que se analizó y se tomó como referencia en el estado del arte, también nos da una base de confiabilidad teniendo en cuenta que el presente proyecto de investigación tiene un rango más alto de confiabilidad establecido. Con el fin de obtener un porcentaje de la recomendación del modelo que sea lo más confiable posible y teniendo en cuenta las características anteriormente descritas, se considera un trabajo a futuro descrito ampliamente en el apartado 6.3

“Trabajos a futuro”, para de esta forma obtener una confiabilidad en la recomendación del 100%.

Como ultima razón se tiene en cuenta las diferentes entrevistas con el gestor de egresados que trajeron opiniones importantes para establecer una métrica de confiabilidad, principalmente por el conocimiento que se tiene, de la no estandarización del formulario de actualización de datos de los egresados, que arrojarían resultados no totalmente confiables, pero si bien los datos no eran totalmente confiables se afrontó el reto dentro del desarrollo de este proyecto de establecer esta métrica de confiabilidad para dar un soporte a la hora de evaluar las recomendaciones y presentarlas al gestor de egresados y de esta manera sugerir el modelo construido como una óptima solución al problema de reintegración de los egresados a través de la oferta académica enfocada los perfiles de los egresados, añadido a esto se presenta una prueba de validación y satisfacción encontrada en el apartado 4.2 “Prueba y validación del modelo”.

3.7.3 Construcción del modelo:

3.7.4 Caracterización del modelo de usuario.

Dentro de la construcción del modelo se tiene en cuenta que todas las técnicas elegidas anteriormente tienen un conjunto de parámetros que determinan la caracterización del modelo que se va a construir, la selección de estos parámetros es un proceso analítico que se basa en diferentes criterios de selección y pertinencia para el proyecto. A continuación se presenta la caracterización de un modelo de usuario a partir de las características de un perfil genérico y las características de los atributos del conjunto de datos.

Tabla 10. Tabla de caracterización de usuario, teniendo en cuenta la información general del Dataset.

Información profesional	Información preferencial	Información educativa posgradual
-Programa	-¿En qué temas de actualización le gustaría capacitarse? -¿principalmente qué otros estudios le gustaría cursar en la corporación?	-Seleccione el programa de formación posgradual o complementario que ha realizado. -Mencione el nombre del programa o curso de formación posgradual que haya realizado.

Teniendo en cuenta la información obtenida en la tabla anterior, con el fin de definir de manera previa las relaciones entre cada uno de los elementos característicos, se destinan los ítems en los ejes de la tabla, con el fin de determinar si cada ítem es de tipo dependiente o independiente entre los otros ítems, es importante tener en cuenta todos los ítems utilizados en la aplicación de los modelos, a continuación se muestra la tabla 11 que corresponde a la información general del Dataset donde se puede mostrar las relaciones de dependencia e independencia entre ítems.

Tabla 11. Tabla Comparación entre ítems de la información general del Dataset de prueba.

Eje y		Eje x				
		programa	¿En qué temas de actualización le gustaría capacitarse?	¿principalmente qué otros estudios le gustaría cursar en la corporación?	Seleccione el programa de formación postgradual o complementario que ha realizado.	-Mencione el nombre del programa o curso de formación postgradual que haya realizado.
programa	x					
¿En qué temas de actualización le gustaría capacitarse?		x				
¿principalmente qué otros estudios le gustaría cursar en la corporación?			x			
Seleccione el programa de formación postgradual o complementario que ha realizado.				x	x	
-Mencione el nombre del programa o curso de formación postgradual que haya realizado.				x	x	
					Dependiente	
					Independiente	

En la tabla 11 se puede observar como los ítems “Seleccione el programa de formación posgradual o complementaria que ha realizado” y “mencione el nombre del programa o curso de formación posgradual que haya realizado” son los únicos que tienen una relación, estos ítems corresponden a la información educativa posgradual que puede brindar el Dataset. Para obtener las métricas en la tabla anterior, durante la creación de

las tablas comparativas, surgen unas preguntas para categorizar los ítems que se obtuvieron. Las preguntas para obtener las métricas son las siguientes:

A. **¿Qué Ítems son dependientes o son independientes?**

Significado:

A1. Dependencia: Este representa que Ítem se puede inferir o calcular a partir de uno o más Ítems (no necesariamente independientes).

A2. Independencia: Es el ítem que no puede ser inferido o calculado a partir de otro ítem, por lo que se hace necesario pedirlo al usuario.

B. ¿Qué tipo de relación tienen los ítems dependientes respecto a otros ítems?

Significado:

B1. Relación cualitativa: Es aquella relación de ítems en que el resultado no genera un valor exacto, pero es deducible de hipótesis lógicas planteadas a partir de la experiencia.

B2. Relación cuantitativa: Es aquella que se forma entre los ítems cuyo resultado proviene de una relación exacta.

Respuesta: se identificó que todos los ítems son de tipo cualitativos y tienen una relación cualitativa.

C. ¿Existen Ítems mutuamente dependientes?

Significado:

C1. Dependencia mutua: Un ítem puede ser deducible a partir de otro y la relación se da de igual manera en sentido contrario.

C2. No dependencia mutua: Cuando la inferencia o cálculo de un ítem hacia otro Ítem, se da solo en una dirección y no viceversa.

Respuesta: Según las tablas anteriores se identificó que existe dependencia mutua entre “Seleccione el programa de formación posgradual o complementaria que ha realizado” y “mencione el nombre del programa o curso de formación posgradual que haya realizado”, ya que sabiendo un ítem se podría deducir el otro.

D. Entre los ítems mutuamente dependientes, ¿Cuál tiene más nivel de importancia de acuerdo a su clasificación?

Significado:

D1. De rutina: El ítem puede ser periódico o no.

D2. Periódico: El ítem contiene la misma información a medida que el tiempo transcurre.

D3. No periódico: El ítem no contiene la misma información en diversos puntos del tiempo.

D4. De consistencia: El ítem ofrece una mejor consistencia en la información que otro.

Respuesta: Dentro de los ítems de dependencia mutua, la pregunta “mencione el nombre del programa o curso de formación post gradual que haya realizado” tiene mayor consistencia en la información que brinda, por ende es de consistencia. Finalmente en el

proceso de evaluación de la caracterización de los ítems por medio de las métricas anteriormente definidas, se identificaron los resultados mostrados en la tabla 12, donde el número indica la cantidad de ítems para ser registrados en los atributos.

Tabla 12. Consulado de ítems de acuerdo a métricas de evaluación.

Total Ítems	5
ítems información profesional	1
ítems informacion preferencial	2
ítems informacion educativa postgradual	2
ítems Independientes	3
ítems Dependientes	2
ítems No mutuamente dependientes	4
ítems mutuamente dependientes	1
ítems periodicos	0
ítems no periodicos	5

Los ítems tienen una relación que se realizan a partir de tablas comparativas de información y respetando las métricas que se hicieron en la tabla 11. Una vez los datos de los usuarios sean obtenidos y el agente inteligente establece la forma de aprendizaje, con la ayuda de un algoritmo de clasificación se podrá realizar el proceso de inferencia, esto se hace con el fin de determinar que posgrado debe realizar el egresado o el más adecuado para el egresado de acuerdo a su perfil.

Tabla 13.Convenciones de la relación.

Simbolo	Descripción
	**Bloque verde:** Indica los items que son independientes
	**Bloque azul:** indica los oitems que son dependientes
	**Bloque rojo:** Indica los items que brindan información de seleccione el programa postgradual que realizo y mencionar el nombre del programa postgradual que realizo.
	**Flecha negra:** Relacion unidireccional entre ítems.
	**Flecha azul:** Relación bidireccional entre ítems.
	Estrella amarilla: ítem que ofrece mayor información en una relación de dependencia mutua.

En las figuras 1, 2 se pueden apreciar un ejemplo del proceso de aprendizaje del modelo de usuario con base en las relaciones de los ítems. Inicialmente, se inicia a partir

de los parámetros del programa y en qué temas de actualización le gustaría capacitarse (figura 1), posteriormente es observado el tipo de posgrado de un egresado en el momento que Menciona el nombre del programa o curso de formación posgradual que haya realizado (figura 2) así como la conexión entre seleccionar el programa de posgrado y mencionar el programa que realizó en la plataforma con los ítems objetivos del sistema; También es posible mirar las relaciones indirectas que surgen en las líneas de flujo. Teniendo en cuenta lo anterior, este modelo permite las relaciones entre ítems de manera que se utilicen para deducir qué tipo de posgrado realizó el egresado.

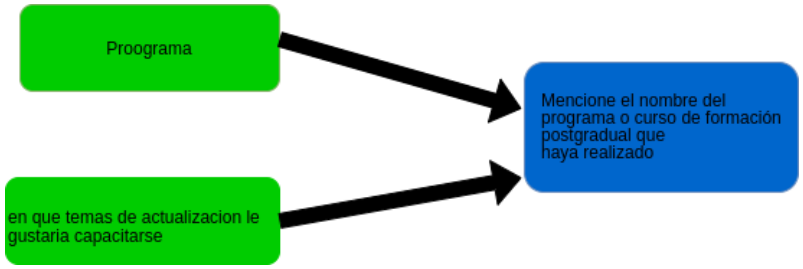


Figura 1. Relación de ítems sobre mencionar el posgrado que ha realizado el egresado.

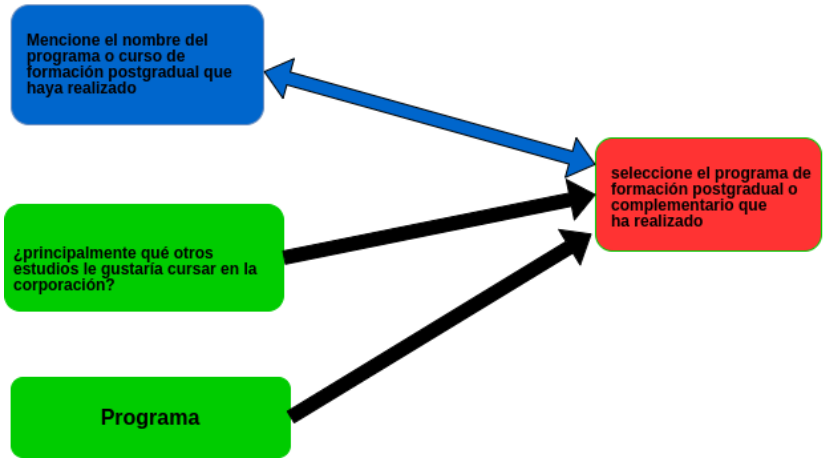


Figura 2. Relación de ítems con “seleccionar el programa de formación posgradual o complementario que ha realizado”.

El diagrama de flujo evidencia que además de tener en cuenta la información del usuario o lo que el usuario brinda mediante el sistema que contiene el modelo, el agente inteligente obtiene información adicional. Como contemplar la intervención inferida y recomendada. Dentro del modelo la información adicional permite que el agente inteligente se retroalimenta y profundizar mejor sus relaciones, teniendo como una consecuencia una recomendación más acertada en el futuro.

3.7.5 Ajuste de parámetros.

Teniendo en cuenta que se han definido dos objetivos de minería de datos, se va a dividir esta tarea en dos partes, una por cada objetivo, dada las diferentes condiciones de los parámetros del modelo a construir según el objetivo que se desea conseguir.

- **Objetivo 1**: Se pretende identificar los perfiles de los egresados con una exactitud de un 80% o más de confiabilidad.

Para cumplir con este objetivo se realizó una categorización de las entidades para cada atributo, es decir, se les da un valor numérico único a cada opción de respuesta a

la pregunta formulada que corresponde, permitiendo así la categorización de las respuestas para poder realizar un análisis del perfil del egresado a través de sus preferencias. En la imagen 34 y 35 se evidencia los resultados de las técnicas utilizadas en el entorno de desarrollo Jupyter-Notboock, para realizar primero una categorización y después determinar una similaridad con cada uno de los perfiles basándose en el principio de distancia euclidiana para determinar la similaridad, además se ponderaron los atributos con el fin de tener una medida de esta similaridad, los resultados, se basan en la prueba realizada con un perfil aleatorio.

```
#convirtiendo lo datos textuales en datos numericos, categorizacion

X=pd.get_dummies(data=X, columns=['programa', 'tipo_programa', 'quiere_estudiar'])
X
```

	programa_comunicacion social y periodismo	programa_contaduria publica	programa_ingenieria de sistemas	programa_ingenieria industrial	programa_ingenieria mecatronica	programa_tecnologia agroambiental	programa_tecnologia de la informacion y comunicacion
0	0	0	0	0	0	1	0
1	0	0	0	0	0	1	0
2	1	0	0	0	0	0	0
3	1	0	0	0	0	0	0
4	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...
70	0	0	0	1	0	0	0
71	0	1	0	0	0	0	0
72	0	1	0	0	0	0	0
73	0	0	0	0	1	0	0
74	1	0	0	0	0	0	0

75 rows × 17 columns

Imagen 34. Resultados categorización

```
#Prueba filtrado por conocimiento
er = filtrado_conocimiento(x, X, Y)
er
```

	index	sim
0	4	1
1	5	0
2	11	0
3	16	0
4	23	0
5	33	0
6	36	1
7	39	0
8	48	0
9	49	0
10	50	0
11	55	0
12	57	0
13	60	0
14	62	0
15	63	0
16	64	0
17	66	0

Imagen 35. Resultados categorización

Analizando los resultados de la función aplicada al Dataset con el fin de realizar una categorización de los datos textuales existentes a datos categóricos en representación de ceros y unos, se obtuvo la anterior matriz, en esta se evidencia como dependiendo del atributo seleccionado se rellena con 1 si ha cursado determinado programa previamente identificado con los valores únicos de los atributos como se evidencia en el Anexo A, y si no fue cursado se le da un valor de cero. Seguido de esto se presenta el resultado de la técnica de distancia euclidiana realizada y los porcentajes resultantes, teniendo en cuenta la prueba aleatoria realizada, mencionada previamente.

- **Objetivo 2**: Recomendar según el perfil del egresado previamente identificado, el programa de formación posgradual o complementaria más idóneo a cursar, por parte de un egresado. Para la consecución de este objetivo, se pretende recomendar el campo “Menciona que programa de formación posgradual o complementaria ha realizado”, es decir, se tomó como clase principal a este atributo, y se eligieron los parámetros correspondientes para el entorno de construcción de la técnica escogida. A continuación se visualizan los resultados de la técnica utilizada en el sistema de recomendación basado en conocimiento.

	programa	Majority_estudio
0	comunicacion social y periodismo	tecnologias de la informacion y comunicacion e...
1	contaduria publica	tributacion
2	ingenieria de sistemas	tecnologias de la informacion y comunicacion e...
3	ingenieria industrial	lean six sigma
4	ingenieria mecatronica	tecnologias de la informacion y comunicacion e...
5	tecnologia agroambiental	educacion ambiental
6	tecnologia de la informacion y comunicacion	tecnologias de la informacion y comunicacion e...
7	tecnologia gastronomia	gestion de innovacion emprendimiento y marketing

Imagen 36. Más votado por programa

```
#Prueba filtrado colaborativo
fc = filtrado_colaborativo(x)
fc
```

	programa	tipo_programa	Majority_estudio
7	ingenieria de sistemas	especializacion	tecnologias de la informacion y comunicacion e...

Imagen 37. Más votado por programa

En la imagen 28 se puede observar como el principio de la técnica de mayor votado realiza un análisis dentro del Dataset y recomienda el programa de formación posgradual más votado, según el programa de pregrado realizado, para el caso del programa de ingeniería de sistemas se evidencia como el programa más votado es el de tecnologías de la información y comunicación, esto se corrobora teniendo en cuenta el previo análisis de datos realizado. En la imagen 29 se muestra el resultado de la recomendación hecha, a la prueba aleatoria ejecutada al algoritmo, que según los datos ingresados, recomienda el programa y el tipo de programa más votado.

3.7.6 Ejecución de los modelos.

Se ejecutaron cinco técnicas para el cumplimiento de los objetivos de la inteligencia de datos, propuesta, sobre un Dataset de 75 instancias y 5 atributos, en dos entornos, el primero Weka donde se aplicó J48 y KNN y en Jupyter-notebook Distancia euclidiana y mayoría ponderada, continuación se visualizan la ejecución de cada técnica.

- **Modelo 1.**

  En la imagen 38, se evidencian los resultados al aplicar el algoritmo J48, seguido de esto, se presenta la matriz de confusión resultante y por último el árbol de decisión construido por Weka en nodos y relaciones.

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

Correctly Classified Instances          51               68      %
Incorrectly Classified Instances        24               32      %
Kappa statistic                          0.3689
Mean absolute error                      0.036
Root mean squared error                  0.149
Relative absolute error                 58.4626 %
Root relative squared error             89.2989 %
Total Number of Instances               75
```

Imagen 38. Resultados J48

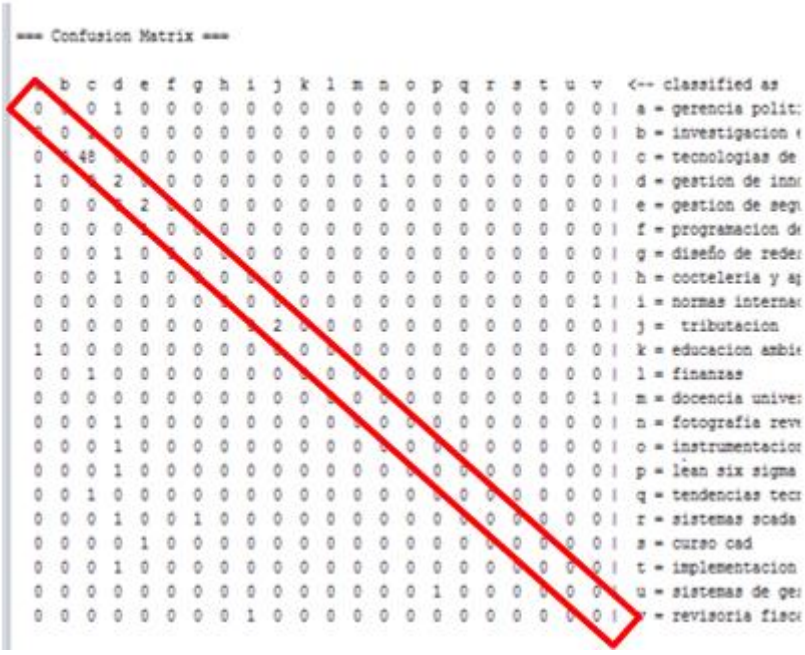


Imagen 39. Matriz de confusión J48

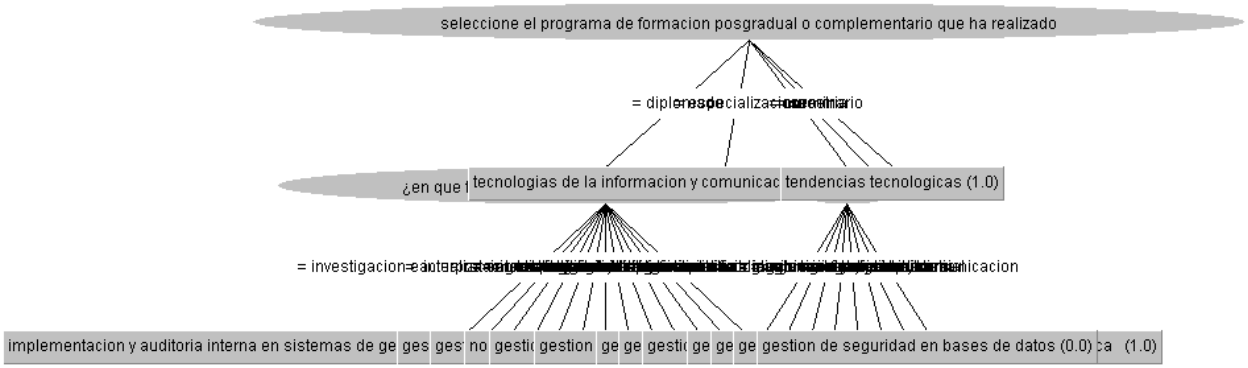


Imagen 40. Árbol de decisión J48 (imagen detallada en Anexo B)

El análisis realizado de la ejecución del algoritmo J48 presenta como resultados el porcentaje de instancias correctamente clasificadas en la imagen 38, también se presenta a matriz de confusión resultante en la imagen 39 donde se evidencia las falencias previamente encontradas en el Dataset utilizado para la construcción del modelo, en la diagonal marcada con rojo, que evidencia, como la categoría C del atributo clase es el más cursado, por último se presenta el árbol de edición resultante, donde se

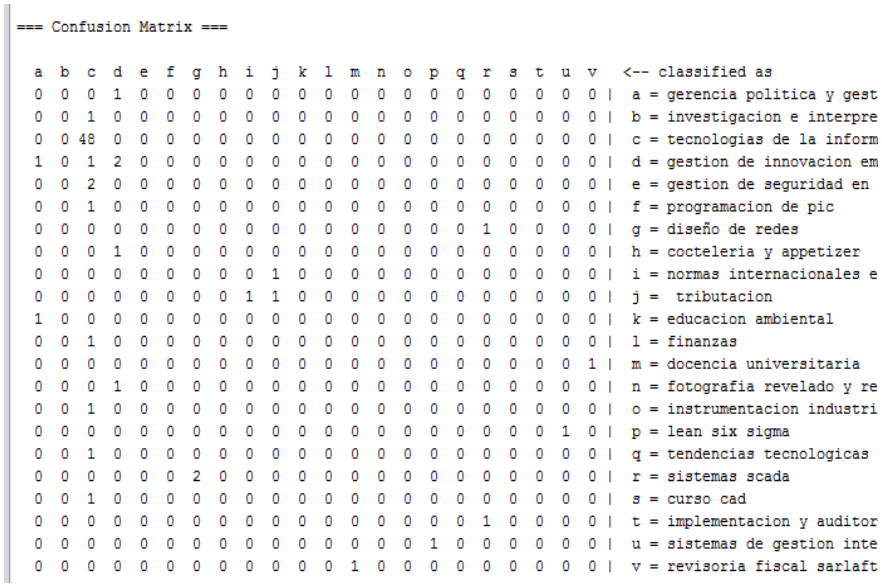
evidencia el orden de los nodos y relaciones creadas por el algoritmo, en el que se evidencia la importancia del atributo (tipo de programa) en la relevancia de sus clasificaciones, para verlo detalladamente diríjase al Anexo B.

- **Modelo 2.**

En las siguientes imágenes se muestran los detalles para la ejecución del algoritmo KNN en Weka y sus resultados obtenidos, como también la matriz de confusión resultante del algoritmo.

```
Number of Leaves  :     24

Size of the tree :      27


Time taken to build model: 0.07 seconds

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

Correctly Classified Instances          54               72      %
Incorrectly Classified Instances        21               28      %
Kappa statistic                          0.4947
Mean absolute error                      0.0305
Root mean squared error                  0.1477
Relative absolute error                 49.5995 %
Root relative squared error             88.5225 %
Total Number of Instances               75
```

Imagen 41. Resultados KNN



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t  u  v   <-- classified as
  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  a = gerencia politica y gest
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  b = investigacion e interpre
  0  0 48  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  c = tecnologias de la inform
  1  0  1  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  d = gestion de innovacion em
  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  e = gestion de seguridad en
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  f = programacion de pic
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0 |  g = diseño de redes
  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  h = cocteleria y appetizer
  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0 |  i = normas internacionales e
  0  0  0  0  0  0  0  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0 |  j =  tributacion
  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  k = educacion ambiental
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  l = finanzas
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1 |  m = docencia universitaria
  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  n = fotografia revelado y re
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  o = instrumentacion industri
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0 |  p = lean six sigma
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  q = tendencias tecnologicas
  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  r = sistemas scada
  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  s = curso cad
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0 |  t = implementacion y auditor
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0 |  u = sistemas de gestion inte
  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0 |  v = revisoria fiscal sarlaft
```

Imagen 42. Matriz de confusión KNN

Teniendo en cuenta los anteriores resultados de la aplicación del algoritmo KNN, se evidencia el porcentaje de instancias correctamente clasificadas en la imagen 35, seguido de esto se obtiene una matriz de confusión resultante de la aplicación del algoritmo, que ratifica lo anteriormente detallado con el anterior algoritmo clasificado, con respecto a la preferencia de la categoría C del atributo clase elegido.

- **Modelo 3.**

Este modelo híbrido trae consigo la consecución de los dos objetivos de minería de datos propuestos, al integrarse los resultados de la similaridad aplicando distancia euclidiana a la técnica de mayoría ponderada, se tiene como resultado que siendo un perfil nuevo, este se compara con todos los perfiles dentro del Dataset y pondera la similaridad, trayendo como resultado los perfiles más parecidos, para después recomendar el programa posgradual más votado dependiendo del programa del perfil con mayor similaridad, todo esto desarrollado y ejecutado bajo una prueba de un perfil

aleatorio en el entorno de desarrollo Jupyter-notebook. Teniendo en cuenta el siguiente caso de prueba:

```
#Prueba de algoritmos de recomendación
#Sea x la respuesta y entradas de los algoritmos
#x es un ejemplo de las entradas de una columna nueva, para ver que se pr
x = ["ingenieria de sistemas", "especializacion", "maestria", "seguridad
```

```
#Prueba filtrado por conocimiento
er = filtrado_conocimiento(x, X, Y)
er
```

Imagen 43. Caso de prueba

```
#Algoritmo de filtrado hibrido
#x es la respuesta para sugerir un resultado
#er es la salida obtenida por el algoritmo de filtrado por conocimiento
#fc es la salida obtenida por el algoritmo de filtrado contributivo
#X_ori es el dataset original de respuestas de la encuesta
#Y_ori son las respuestas finales originales de respuestas de la encuesta
def filtrado_hibrido(x, er_ori, fc_ori, X_ori, Y_ori):
    er = er_ori.copy()
    fc = fc_ori.copy()
    X = X_ori.copy()
    Y = Y_ori.copy()
    X_er = X.iloc[er["index"].to_list()]
    Y_er = Y.iloc[er["index"].to_list()]
    Y_er.to_list()
    data_er = X_er.copy()
    data_er["estudio"] = Y_er;
    f = lambda x: x.mode().iat[0]
    f_er_colaborativo = data_er.groupby(['programa', 'tipo_programa'])['
    cond_1 = f_er_colaborativo["programa"]==x[0]
    cond_2 = f_er_colaborativo["tipo_programa"]==x[1]
    f_er = f_er_colaborativo[cond_1 & cond_2]
    length = len(f_er)
    fc_list = fc.iloc[0].to_list()
    f_er = f_er.append({"programa":fc_list[0], "tipo_programa":fc_list[1]
    f_er = f_er.drop_duplicates(subset = ["programa", "tipo_programa", "
                      keep = "first", inplace = False)
    return f_er
```

Imagen 44. Algoritmo filtrado híbrido

```
#Prueba filtrado hibrido
#Para esta se necesita haber realizdo las demás pruebas con anterioridad
#x es la respuesta para sugerir el programa
#er es la salida del algoritmo de recomendación por conocimiento
#fc es la salida del algoritmo de recomendación contributivo por programa
#X es el conjunto de datos original
#Y es
fer= filtrado_hibrido(x, er, fc, X, Y)
```

```
fer
```

	programa	tipo_programa	Majority_er_estudio
0	ingenieria de sistemas	especializacion	tecnologias de la informacion y comunicacion e...

```
fer
```

	programa	tipo_programa	Majority_er_estudio
0	ingenieria de sistemas	especializacion	tecnologias de la informacion y comunicacion e...

Imagen 45. Resultado algoritmo híbrido

Dados los resultados de la aplicación del algoritmo híbrido se analiza como este algoritmo concluye en resultados favorables con los objetivos propuestos en la investigación y sumado a esto da una confiabilidad asertiva de la recomendación que se quiere realizar.

3.7.7 Evaluar el modelo.

Si bien en la siguiente fase se ejecuta una evaluación de los modelos generados, la evaluación que se presentará a continuación está más enfocada a los objetivos de la minería de datos propuestos previamente, mientras que en la siguiente fase se orienta más al cumplimiento de los objetivos de negocio. Como bien se establecieron las condiciones para evaluar los modelos y las funciones que estarían involucradas a este proceso, cabe mencionar la validación cruzada, siendo la principal función de validación

en términos de minería de datos, esta función tiene como resultado los indicadores de error de las instancias correctamente clasificadas y las no correctamente clasificadas.

Además de estos indicadores para los casos de los algoritmos J48 y KNN se tuvo en cuenta el análisis del resultado de la matriz de confusión de cada algoritmo; Para la evaluación de los modelos propuestos basados en un sistema de recomendación basado en casos sé, utilizó “leave one out-cross validation”. A continuación se presenta una tabla correspondiente a los resultados de las métricas que se evidencian dentro de la aplicación de las funciones de evaluación para cada modelo ejecutado. Hay que tener en cuenta que los modelos 3, 4 y 5 son evaluados en la ejecución del modelo 6, ya que este es un modelo híbrido que contiene los anteriores modelos y evalúa sus resultados.

Tabla 14. Resultados pruebas de validación.

	Instancias correctamente clasificadas	Instancias incorrectamente clasificadas
Modelo 1	72%	28%
Modelo 2	68%	32%
Modelo 3	81%	19%

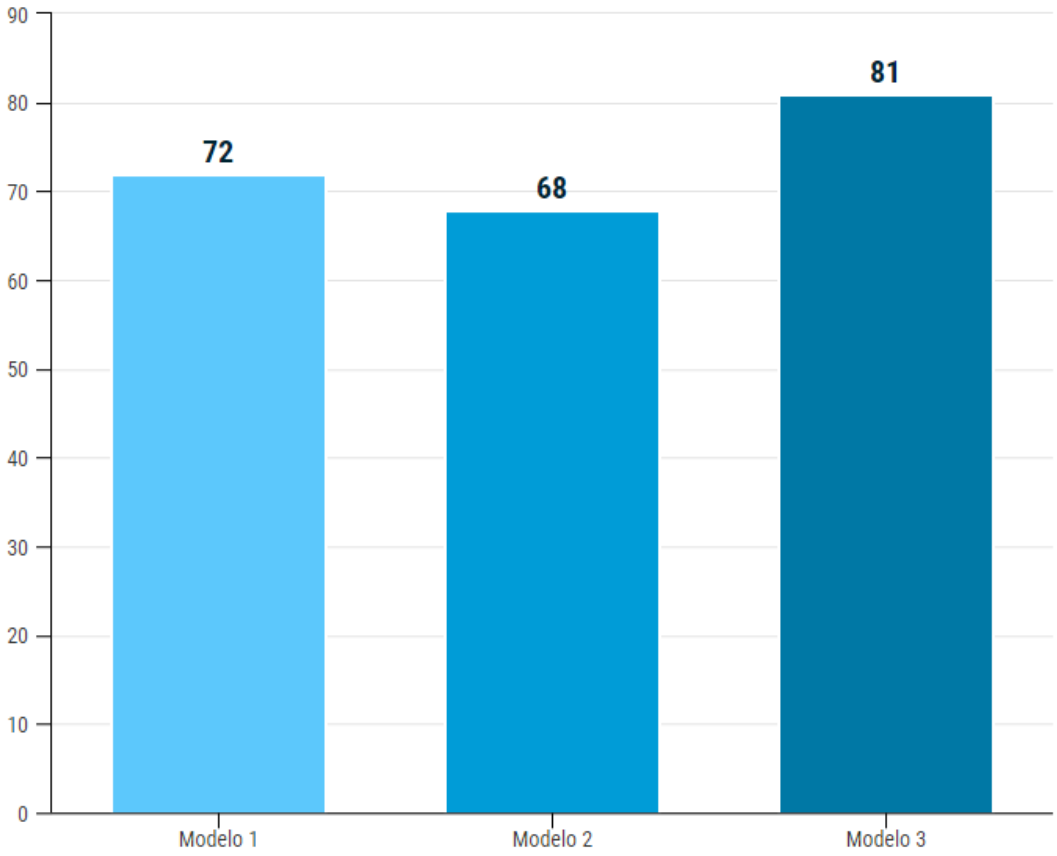


Figura 3. Barra de porcentajes resultados de la evaluación aplicada.

3.7.8 Revisar el proceso.

Durante el proceso del modelo se realizaron tareas de construcción, planeación y evaluación para el mismo, donde se evidencia en imágenes y descripción lo realizado en esta fase, como parte muy importante se presenta al final, una pequeña discusión de la evolución hecha a los modelos analizados, y el modelo escogido.

Capítulo cuatro

4. FASE 5 Evaluación:

El propósito de esta fase en el desarrollo de la metodología es evaluar los modelos creados anteriormente, sin embargo, en comparación con la evaluación del modelo realizado anteriormente, esta evaluación se enfoca desde el punto de vista del cumplimiento de los objetivos del negocio. Luego de la evaluación, se presenta el informe del modelo aprobado, este contiene como fueron cumplidos los objetivos, aquí también se tienen en cuenta los factores que sé no se hayan tenido en cuenta, y está expuesto a cualquier corrección en las tareas anteriores.

4.1 Evaluar los resultados:

Modelo 1. Para el cumplimiento del objetivo 2 de minería de datos.

Este modelo la primera instancia se tenía como un modelo viable, principalmente por el algoritmo en el que se basaba, también se tenía la facilidad con la que puede ser implementado y mejorado el algoritmo dentro de la herramienta Weka, sin embargo, como se identificó en etapas previas él Dataset se inclinaba hacia un tipo de dato dentro del atributo clase a recomendar, por eso, luego de observar los resultados de la ejecución del modelo, se tuvo un resultado de menos de 70% de instancias correctamente clasificadas, lo que evidenciaba una clasificación en la cual no se podía confiar para seguir desarrollando este modelo.

Modelo 2. Para el cumplimiento del objetivo 2 de minería de datos.

Este segundo modelo trajo consigo resultados similares al del algoritmo anterior, al aplicar KNN, sobre el Dataset, se evidenció en su matriz de confusión como existía el

desbalance ya identificado, además arrojó un resultado menor a 80% de instancias correctamente clasificadas, aunque este modelo tuvo un porcentaje mayor de instancias correctamente clasificadas al anterior, seguía sin cumplir los requisitos establecidos para dar cumplimiento los objetivos de minería de datos propuestos, por este motivo se tomó la decisión de no seguir desarrollando el modelo bajo estas herramientas y técnicas.

Modelo 3. **Para dar cumplimiento a los dos objetivos de minería de datos.**

La construcción de este modelo involucró tres técnicas seleccionadas previamente que podrían dar cumplimiento a los objetivos. La primera técnica que se utilizó, fue la del filtrado por conocimiento utilizando el principio de la distancia euclidiana con el fin de determinar la similaridad de los perfiles, para medir esta similaridad se ponderaron los atributos y se realiza la suma de estos porcentajes, seguidamente se establece la condición si tiene un porcentaje de similaridad mayor a 80% se preseleccionan los perfiles y se visualizan en una matriz, todas con el índice de un porcentaje de más de 80% de similaridad, lo que da una confianza de la clasificación de los perfiles y la identificación de la similitud entre los perfiles existentes y uno nuevo de prueba.

La segunda técnica que se aplicó dentro de este modelo fue la mayoría ponderada donde se toma la similitud realizada anteriormente y se le recomienda el más votado entre el atributo clase, dependiendo del programa y el tipo de programa. Aquí es donde se vuelve un algoritmo híbrido, ya que se integran las dos técnicas, esta se evaluó con leave one out of cross validation, obteniendo un porcentaje mayor a 80% de confianza en el algoritmo creado.

Modelo aprobado.

Teniendo en cuenta las anteriores evaluaciones se decidió escoger el Modelo 3 como el modelo que cumplía con los objetivos de minería de datos propuestos en el proyecto, ya que los resultados de las métricas evaluativas que se habían considerado para los dos primeros modelos, evidenciaron la falta de una consolidación de los datos obtenidos en el Dataset y por consiguiente las métricas arrojaron resultados no favorables a la hora de determinar una clasificación bien hecha. Desde que se identificaron estas falencias se optó por utilizar otro tipo de técnica de inteligencia de datos que se le podía aplicar al Dataset con el fin de alcanzar los objetivos de minería de datos propuestos, es en este momento que se evalúa la técnica de un sistema de recomendación basado en casos para darle una mayor exactitud a los resultados que se pretendía obtener, como también unir en un solo modelo el cumplimiento de los dos objetivos a través del algoritmo híbrido construido.

4.2 Prueba y validación del modelo.

En esta fase se transforma el conocimiento obtenido en acciones dentro del proceso del negocio, el grupo de investigación podrá recomendar acciones basadas en la observación del modelo y sus resultados. Para la realización de este proceso se ejecutan dos tareas, la primera tarea relacionada directamente con la validación del modelo construido de manera técnica, basándose en la medida de métricas que determinan la correcta clasificación del algoritmo aplicado, la segunda tarea, en cambio, toma como referencia la presentación oral del modelo construido, por parte del grupo de investigadores al personal encargado de los egresados con el fin de obtener una validación de los resultados obtenidos, como también consideraciones,

recomendaciones y trabajos futuros que se puedan identificar, añadido a la presentación oral, se presenta una prueba de validación en un formulario online, que permitirá evaluar los resultados de esta tarea, este proceso va a aliado con la presentación del informe final de la aplicación de la metodología que se encuentra al principio de la siguiente fase.

A continuación se presenta la prueba de validación del modelo escogida, para la realización de este proceso como bien se mencionó en la finalización de la anterior fase, se tuvo en cuenta el porcentaje de similaridad mediante el cálculo de la distancia euclidiana y el porcentaje de validez del algoritmo realizado por intermedio de la aplicación de la prueba de validación (leave one out of cross-validation), además de la aplicación de métricas de correcta clasificación para obtener un porcentaje de confiabilidad en las recomendaciones encontradas.

```
#Creando el conjunto de datos para las pruebas
#Debido a que solamente hay 75 registros, se usan datos que también son de entrenamiento
msk = np.random.rand(len(data)) < 0.85
data_test = data[~msk]
X_test = data_test[["programa", "tipo_programa", "quiere_estudiar", "tema_interes"]]
Y_test = data_test[["estudio"]]
X_test
```

	programa	tipo_programa	quiere_estudiar	tema_interes
1	tecnologia agroambiental	especializacion	maestria	investigacion e interpretacion de cultivos y e...
9	ingenieria mecatronica	especializacion	maestria	tecnologias de la informacion
18	contaduria publica	especializacion	maestria	economia y finanzas
20	ingenieria mecatronica	especializacion	maestria	actualizacion tecnologica en automatizacion
22	tecnologia de la informacion y comunicacion	especializacion	maestria	seguridad informatica
38	tecnologia de la informacion y comunicacion	especializacion	maestria	programacion
41	ingenieria mecatronica	especializacion	maestria	actualizacion tecnologica en automatizacion
58	ingenieria mecatronica	especializacion	maestria	internet de las cosas
66	ingenieria de sistemas	especializacion	maestria	docencia universitaria
69	contaduria publica	especializacion	maestria	docencia universitaria
73	ingenieria mecatronica	especializacion	maestria	actualizacion tecnologica en automatizacion

Imagen 46. Conjunto de datos de prueba

```
: #Accuracy Score
  from sklearn.metrics import accuracy_score
  accuracy_score(y_true, y_pred)

: 0.8181818181818182
```

Imagen 47. Código y resultado de la prueba

A continuación se presenta el proceso de presentación oral que se realizó al personal encargado de los egresados de la Corporación, con el fin de dar a conocer el desarrollo de la investigación realizada y la metodología utilizada para dar cumplimiento a los propuestos, se presenta de manera resumida todo el proceso de investigación y desarrollo del modelo construido, esta tarea esta aliada con la presentación del informe final encontrado en el apartado (5.1 informe final), añadido a esto se analizan los resultados obtenidos de la prueba de satisfacción realizada, y se toman en cuenta las opiniones para trabajos futuros en la siguiente fase.

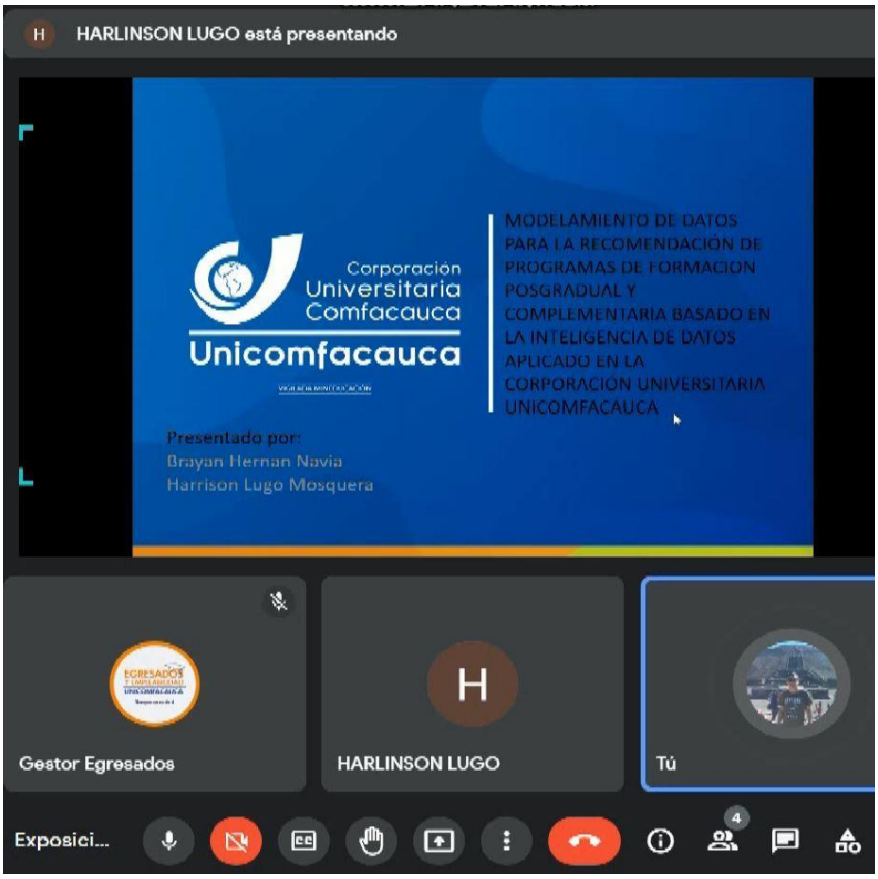


Imagen 48.Reunión virtual con el gestor de egresados

En la anterior imagen se evidencia el proceso de presentación oral realizado al gestor de egresados, para más detalles del proceso de esta presentación diríjase al Anexo C. Este proceso estuvo aliado con el informe final y presento de manera resumida el proceso de investigación del proyecto, la construcción del modelo obtenido y los resultados de los mismos, también se propuso de manera autónoma por parte de los investigadores propuestas a futuro y a considerar teniendo en cuenta el modelo obtenido y su posible utilización.

Imagen 49.Formulario de satisfacción.

En la imagen 49. Se presenta el formulario de satisfacción presentado al gestor de egresados luego de realizar la presentación oral en la cual se basa, para determinar el nivel de satisfacción de los resultados obtenidos, es decir, las recomendaciones hechas por el modelo, a partir del conjunto de datos proporcionado, para evidenciar las demás preguntas correspondientes al formulario diríjase al Anexo D.

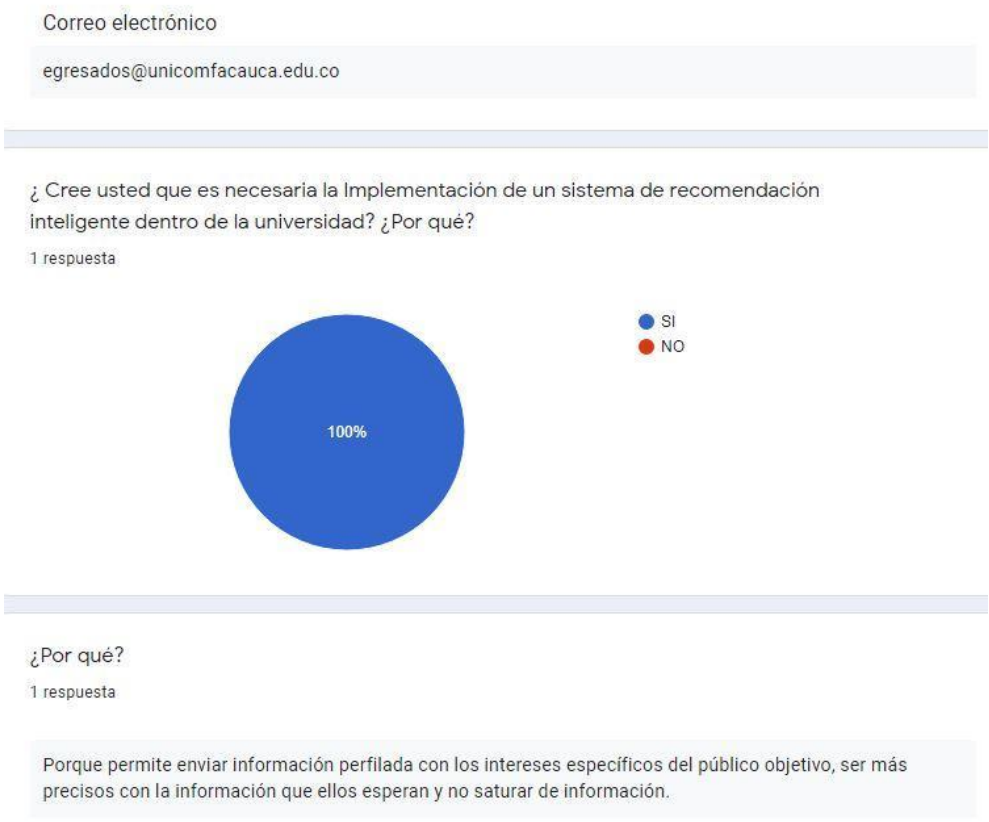


Imagen 50.

En la imagen 50 se presenta el resultado de una de las preguntas del formulario de validación presentado por el gestor de egresados, en esta se evidencia el grado de satisfacción correspondiente a los resultados y propósitos presentados por parte de los investigadores, en ella se encuentra una justificación del por qué sería factible implementar un sistema de recomendación dentro de los procesos que se involucran en la corporación universitaria Unicomfacauca, para visualizar los demás resultados de la prueba de validación, diríjase al Anexo E. Después de analizar las respuestas del formulario realizado por el gestor de egresados, se evidencia una satisfacción de los resultados del modelo presentado teniendo en cuenta métricas que determinan la

viabilidad del sistema de recomendación, como también del cumplimiento de los objetivos de inteligencia de datos propuestos.

4.3 Revisar el proceso.

Las tareas que involucran esta fase se han ejecutado sin ningún contratiempo, sin embargo, se tomaron diferentes decisiones y análisis de la ejecución de los algoritmos en Weka, donde se identificó una anomalía, esta fue corroborada volviendo a la fase anterior de análisis y construcción del Dataset, está decisión fue óptima, puesto que permitió seguir con el análisis de otro tipo de técnica de inteligencia de datos que fue favorable para los objetivos propuestos. También gracias a la evaluación realizada a los modelos se puede determinar una medida de confiabilidad en el modelo escogido y a pesar de no contar con un Dataset amplio para poder realizar pruebas, la prueba de un perfil genérico arrojó muy buenos resultados, como también la prueba de satisfacción del modelo presentado al gestor de egresados, con respecto a la presentación oral del proyecto y las sugerencias expuestas por los investigadores.

5. FASE 6 DESPLIEGUE:

En esta última fase de la metodología se pretende explicar al usuario final como poner en funcionamiento el modelo construido en las fases anteriores, para este proyecto en específico no se planteó una implantación del modelo construido dentro de algún sistema en la universidad, lo que se planteó fue un análisis de las técnicas presentes en la aplicación de la inteligencia de datos, este análisis permitió probar y elegir un modelo de sistema de recomendación basado en casos. Sin embargo, se deja como propuesta de trabajo futuro poder implementar el modelo realizado dentro de la universidad o darle un enfoque a las áreas que puede realizar un aporte el proyecto.

5.1 Informe final.

Esta tarea consiste en presentar un informe final al personal que está relacionado con los intereses de los egresados y su re vinculación a la universidad (Oficina de egresados y empleabilidad, programas de extensión, calidad institucional, mercadeo) entre otras. En este informe se resumen los procesos más importantes dentro del proyecto y el conocimiento adquirido en su ejecución, permitiendo así analizar los diferentes resultados obtenidos y los propósitos planeados a futuro, que contribuyan o no a los intereses de la universidad y sus áreas involucradas. Cabe mencionar que este informe puede ir acompañado de una presentación oral, teniendo en cuenta que aquí se toman de manera resumida los procesos.

El uso de la metodología de investigación-acción permitió generar un desarrollo iterativo que se retroalimenta a medida que se iban identificando los requerimientos para darle solución a la necesidad identificada dentro de los intereses de la universidad con los procesos de los egresados. Gracias al uso de la metodología CRISP-DM se logró encontrar predicciones a través de los datos obtenidos en el tema académico de los egresados en el programa o formación posgradual que más se asemeja a su perfil e intereses establecidos. Los lineamientos de esta metodología permitieron ejecutar un plan de identificación, extracción, integración, normalización y codificación de datos para la realización de un proceso de inteligencia de datos, con esto se lograron alcanzar los tres objetivos propuestos dentro del proyecto, además de dejar un trabajo futuro a considerar.

A continuación se destacarán las etapas más importantes y el conocimiento adquirido durante su ejecución:

En la primera fase se desarrolló una investigación con el fin de identificar el modelo del negocio de la corporación y la descripción extensa de las necesidades u oportunidades que se identificaron. Para esto la oficina de egresados y su encargada dieron su apoyo y compartieron siete tablas Excel que corresponden al formulario semestral que se realiza a los egresados de la institución desde el año 2016 al 2021, también se realizaron dos entrevistas con la encargada de los egresados. Las siguientes fases fueron las más laboriosas y se desarrollaron de forma iterativa es decir, se tomaba la decisión de volver a algunas de las etapas luego de identificar procesos a mejorar o requerimientos que iban cambiando, las fases que se involucraron en este proceso iterativo fueron de la fase dos a la cuatro, donde se procedió un extenso análisis y exploración de los datos, aquí se identificaron los filtros necesarios que debieron ejecutar para obtener un Dataset enfocado a los ítems relevantes para la consecución de los objetivos.

A continuación se procedió a construir él Dataset donde se involucran tareas en análisis de los datos, integración, normalización, categorización, y validación de los datos. Seguido de esto se procedió la elección de las técnicas de modelado y la ejecución de las mismas en la herramienta escogida para este propósito, como lo fueron (Weka y Jupyter-notebook). En esta fase es muy importante la tarea de caracterización de los datos, que permitió establecer las relaciones entre los atributos la importancia de los mismo, y la validación para escoger el atributo clase correspondiente, para cumplir con los propósitos establecidos en el proyecto, si bien no está establecido dentro de la

metodología fue un proceso que surgió de la necesidad de identificar las relaciones y la importancia de los atributos a la hora de ejecutar los modelos.

La construcción de los modelos propuestos fue ejecutada y se analizaron los resultados, denotando una falencia en el Dataset utilizado, esto obligó a tomar decisiones en cuanto a la corrección de procesos en las anteriores fases, también se tomó la decisión temprana de no seguir trabajando con los modelos propuesto en la herramienta (Weka) ya que evidenciaban una inestabilidad de los datos y una baja confiabilidad en la clasificación obtenida. Esto permitió seguir con la aplicación y análisis de otra técnica. Luego de la ejecución de los modelos, y las interacciones hechas entre estas fases, se logra evaluar los resultados, a través de un plan de evaluación previamente establecido. Los resultados de la evaluación hecha a los modelos trajeron los resultados esperados, evidenciando el modelo favorable y los no favorables para el cumplimiento de los objetivos de minería de datos, con respecto a los objetivos del proyecto, y se escogió el modelo 3, como el que cumplía con todos los lineamientos.

Seguido de la evaluación se realizó una prueba y validación del modelo en el que presentan dos tareas, la primera parte del levantamiento de un nuevo conjunto de datos, cabe mencionar que la construcción, limpieza y consolidación de este conjunto de datos fue realizada por parte de los investigadores, este conjunto de datos permite obtener una cantidad de registros más amplia, permitiendo la realización de una validación cruzada dividiendo él Dataset en dos conjuntos de datos, uno de entrenamiento y otro de prueba que permite evidenciar resultados a partir de datos reales. La segunda tarea corresponde al grado de satisfacción y opinión que se tiene por parte del gestor de los egresados y

que permite tener una medida de viabilidad del modelo construido y de las propuestas expuestas para trabajos futuros.

Por último se analizaron los resultados del modelo y se efectuaron las diferentes discusiones, conclusiones, aportes y trabajos futuros que se pudieron identificar en el transcurso del proyecto.

5.2 Análisis y resultados:

5.2.1 Conjunto de datos.

El conjunto de datos creado para la utilización en el modelo contiene cinco atributos y setenta y siete registros, este conjunto de datos es el resultado de los lineamientos con respecto a los objetivos de minería de datos y de negocio.

Tabla 15. Ejemplo conjunto de datos.

Programa	seleccione el programa de formacion posgradual o complementario que ha realizado	¿En el futuro le gustaría cursar otros estudios en la corporacion?	mencione que formacion complementaria o posgradual realizo	¿En el futuro le gustaría cursar otros estudios en la corporacion ?
tecnologia agroambiental	diplomado	especializacion	gerencia politica y gestion administrativa derechos etnicos equidad de genero cultura y paz	especializacion
tecnologia agroambiental	especializacion	especializacion	investigacion e interpretacion de cultivos y especies productoras	especializacion
			tecnologias de la	

5.2.2 Obtención del modelo.

Basado en el conjunto de datos obtenido se realizó un sistema de recomendación basado en casos, la cual utilizo diferentes técnicas de inteligencia de datos para la construcción del modelo, en las siguientes imágenes se va a detallar las reglas que se aplicaron, la descripción de las técnicas utilizadas al momento de construir el algoritmo que obtiene como resultado un modelo datos. En la imagen 51, se puede apreciar cómo se utiliza la técnica de waiting majority voting esta técnica lo que hace es identificar la categoría más votada dentro del atributo clase (nombre del programa posgradual que realizó el egresado), de esta manera recomienda, él recomienda el programa de

formación posgradual más cursado, teniendo como atributos de conocimiento (programa y el tipo de programa de formación posgradual) que haya realizado.

```
: f = lambda x: x.mode().iat[0]
  df3 = data.groupby(['programa', 'tipo_programa'])['estudio'].apply(f).reset_index(name='Majority_estudio')
  df3
```

:

	programa	tipo_programa	Majority_estudio
0	comunicacion social y periodismo	diplomado	fotografia revelado y retoque digital
1	comunicacion social y periodismo	especializacion	tecnologias de la informacion y comunicacion e...
2	contaduria publica	diplomado	docencia universitaria
3	contaduria publica	especializacion	tecnologias de la informacion y comunicacion e...
4	contaduria publica	maestria	tributacion
5	ingenieria de sistemas	curso	gestion de seguridad en bases de datos
6	ingenieria de sistemas	diplomado	diseño de redes
7	ingenieria de sistemas	especializacion	tecnologias de la informacion y comunicacion e...
8	ingenieria de sistemas	seminario	tendencias tecnologicas
9	ingenieria industrial	diplomado	lean six sigma
10	ingenieria mecatronica	curso	curso cad
11	ingenieria mecatronica	diplomado	instrumentacion industrial
12	ingenieria mecatronica	especializacion	tecnologias de la informacion y comunicacion e...
13	tecnologia agroambiental	diplomado	educacion ambiental
14	tecnologia agroambiental	especializacion	investigacion e interpretacion de cultivos y e...
15	tecnologia de la informacion y comunicacion	especializacion	tecnologias de la informacion y comunicacion e...
16	tecnologia gastronomia	diplomado	gestion de innovacion emprendimiento y marketing

Imagen 51. Weidthing Majoriting

En la imagen 52, se puede apreciar la ejecución del filtrado colaborativo y como resultado se obtiene que la especialización llamada (tecnologías de la información y comunicación) siendo el más popular o el más votado de los programas de formación posgradual realizado por los egresados que se encuentran dentro del conjunto de datos de entrenamiento.

	programa	Majority_estudio
0	comunicacion social y periodismo	tecnologias de la informacion y comunicacion e...
1	contaduria publica	tributacion
2	ingenieria de sistemas	tecnologias de la informacion y comunicacion e...
3	ingenieria industrial	lean six sigma
4	ingenieria mecatronica	tecnologias de la informacion y comunicacion e...
5	tecnologia agroambiental	educacion ambiental
6	tecnologia de la informacion y comunicacion	tecnologias de la informacion y comunicacion e...
7	tecnologia gastronomia	gestion de innovacion emprendimiento y marketing

Imagen 52. Más votado por programa

En la imagen 53, se observa como primer paso se hace una similitud entre todos los registros que hay en el Dataset a través de la aplicación de la técnica de comparación distancia euclidiana que determina el porcentaje de similaridad, después se aplica el filtrado híbrido que contiene la unión de las dos técnicas anteriormente mencionadas, para el cual se le va a pasar el nuevo perfil de prueba, a este nuevo perfil se le evalúa la similitud con todo el Dataset que se tuvo. Con esto se obtuvieron unos perfiles del Dataset el cual se le determina la similaridad que se tienen con los perfiles nuevos de ejemplo con los demás perfiles de la Dataset. Se puede analizar que hay dos perfiles que obtienen un 100% de similitud que corresponden, según su índice a los perfiles perfil “0” y el perfil “6”.

```
#Prueba filtrado por conocimiento
er = filtrado_conocimiento(x, X, Y)
er
```

	index	similitud	estudio
0	4	1.0000	tecnologias de la informacion y comunicacion e...
1	5	0.8025	gestion de seguridad en bases de datos
2	11	0.8125	tecnologias de la informacion y comunicacion e...
3	16	0.8225	tecnologias de la informacion y comunicacion e...
4	23	0.8225	tecnologias de la informacion y comunicacion e...
5	33	0.8225	tecnologias de la informacion y comunicacion e...
6	36	1.0000	tecnologias de la informacion y comunicacion e...
7	39	0.8225	tecnologias de la informacion y comunicacion e...
8	48	0.8225	tecnologias de la informacion y comunicacion e...
9	49	0.8025	tendencias tecnologicas
10	50	0.8225	tecnologias de la informacion y comunicacion e...
11	55	0.8225	tecnologias de la informacion y comunicacion e...
12	57	0.8225	tecnologias de la informacion y comunicacion e...
13	60	0.8225	tecnologias de la informacion y comunicacion e...
14	62	0.8225	tecnologias de la informacion y comunicacion e...
15	63	0.8225	tecnologias de la informacion y comunicacion e...
16	64	0.8225	tecnologias de la informacion y comunicacion e...
17	66	0.8225	tecnologias de la informacion y comunicacion e...

Imagen 53. Similitud con los perfiles del Dataset.

En la imagen 54, se aplica el filtrado colaborativo que corresponde a la recomendación del programa que más se ha realizado por los egresados en el Dataset de entrenamiento, la mayoría de los egresados han cursado la especialización en (tecnologías de la información y comunicación en la educación) que como bien se había identificado en la fase de exploración de los datos, corresponde efectivamente al programa de formación posgradual más cursado.

```
#Algoritmo de filtrado colaborativo
def filtrado_colaborativo(x):
    cond_programa = f_colaborativo["programa"]==x[0]
    cond_tipo_programa = f_colaborativo["tipo_programa"]==x[1]
    return f_colaborativo[cond_programa & cond_tipo_programa]
```

```
#Prueba filtrado colaborativo
fc = filtrado_colaborativo(x)
fc
```

	programa	tipo_programa	Majority_estudio
7	ingenieria de sistemas	especializacion	tecnologias de la informacion y comunicacion e...

Imagen 54. Caso de prueba, más votado por programa.

En las siguientes imágenes, se realiza el filtrado híbrido del algoritmo, el cual utiliza los resultados de la similitud del Dataset y el filtrado colaborativo que corresponde al más votado.

```
#Prueba de algoritmos de recomendación
#Sea x la respuesta y entradas de los algoritmos
#x es un ejemplo de las entradas de una columna nueva, para ver que se pr
x = ["ingenieria de sistemas", "especializacion", "maestria", "seguridad
```

```
#Prueba filtrado por conocimiento
er = filtrado_conocimiento(x, X, Y)
er
```

Imagen 54. Caso de prueba.

```
#Prueba filtrado hibrido
#Para esta se necesita haber realizdo las demás pruebas con anterioridad
#x es la respuesta para sugerir el programa
#er es la salida del algoritmo de recomendación por conocimiento
#fc es la salida del algoritmo de recomendación contributivo por programc
#X es el conjunto de datos original
#Y es
fer= filtrado_hibrido(x, er, fc, X, Y)
```

```
fer
```

	programa	tipo_programa	Majority_er_estudio
0	ingenieria de sistemas	especializacion	tecnologias de la informacion y comunicacion e...

```
fer
```

	programa	tipo_programa	Majority_er_estudio
0	ingenieria de sistemas	especializacion	tecnologias de la informacion y comunicacion e...

Imagen 55. Resultado de algoritmo híbrido.

Analizando el resultado de la prueba aleatoria realizada como se muestra en la imagen 54, donde se construye un conjunto de datos de pruebas con una respuesta aleatorias a los atributos correspondientes, se aplica en primera instancia el algoritmo por conocimiento y luego el híbrido en el cual se obtiene el resultado visualizado en la imagen 55 que contiene el cuadro resaltado con rojo, en la primera línea se marca la recomendación basada en conocimiento, es decir, según su perfil cuál es el programa

de formación posgradual o complementaria que más ajusta a sus preferencias, y en la segunda línea la recomendación colaborativa que dependiente del programa de formación de pregrado puede recomendar al programa de formación posgradual o complementaria que más se haya realizado en dicho programa.

6. Discusión, conclusiones y trabajos futuros:

6.1 Discusión.

En el análisis de datos se identificaron problemas a partir de la insuficiencia de datos, se requería que la calidad de los datos fuera mejor para que así se obtuviera un buen conjunto de datos; un conjunto de datos más confiable y de buena calidad requiere de bastantes registros y de atributos, para este caso preguntas del formulario que ayuden al mejorar el modelamiento de los datos así como la relación de los atributos con los registros para que así exista la correlación y de la misma manera se obtengan buenos resultados, este inconveniente se presentó por la actualización anual de los formularios de opinión que realizan los egresados, estas actualizaciones traen diferentes características en las preguntas, no manejan un enfoque puntual en un tema que se dé importancia para aprovechar, las temáticas a tener en cuenta son, el tema académico, el tema profesional y el desarrollo personal de los egresados que no se priorizaron en la construcción del formulario, también la incoherencia y la integridad de las respuestas por parte de los egresados obligo a la homologación de diferentes registros con una respuesta acorde a la pregunta formulada.

A partir de los resultados del análisis de satisfacción de los expertos se puede construir un proyecto a futuro más grande dependiendo si el grado de satisfacción que obtuvieron del personal encargado de egresado encargado de realizar la validación era

favorable, para eso se tuvo en cuenta la prueba de satisfacción realizada al gestor de egresados y su opinión compartida con respecto a las propuestas expuestas por los investigadores y su perspectiva. Como resultado se obtuvo una opinión favorable al modelo presentado y una seria de posibles propuestas a considerar a partir del modelo construido, estos proyectos se encuentran descritos de manera más detallada en el apartado (trabajos futuros).

Para la realización del conjunto de datos de prueba se tuvieron muchas dificultades tanto en el análisis como en el procesamiento de los datos, se tuvieron que hacer muchos filtros para obtener un conjunto de datos más exacto, añadido a esto se le ejecutó limpieza y transformación a los datos. La metodología CRIPS-DM que se utilizó, contribuyo a una retroalimentación en cuanto a la construcción de los datos por muchas inconsistencias en estos. Se efectuaron dos análisis el segundo análisis se hizo un filtrado diferente que al primer análisis, se determinó un enfoque el cual es la oferta académica, se tuvieron muchas dificultades en cuanto a los campos vacíos porque había muchos campos así y se tuvieron que homologar diferentes datos esto con el fin de tener un resultado favorable entre los algoritmos aplicados, El conjunto de datos tenía muchos caracteres especiales así como mayúsculas en sus campos y tildes en las palabras, esto genera ruido al subir el archivo a Weka y por esta razón se le hizo un formateo y una mejor limpieza de datos.

Se probaron tres tipos de algoritmos los cuales fueron: CBR, KNN y J48 en el cual los algoritmos fueron eficientes, sin embargo, se hace un algoritmo híbrido por la razón que el filtrado colaborativo como el más popular de todos los cursos para las personas que no quieren dar toda su información en el sistema y el algoritmo basado en el

conocimiento para los usuarios que brindan su perfil y este se recomienda con base en sus preferencias y también sus necesidades. También se aplican unas reglas las cuales se utilizan para no recomendar y no repetir los cursos que el usuario ya ha tomado.

6.2 Conclusiones.

La limpieza de los datos se hizo de una manera en la que se filtraron los diferentes errores que no se deben cometer al momento de hacer una buena Dataset. Los errores que se tuvieron en cuenta fueron, las tildes, las comas, las mayúsculas los campos nulos, los espacios innecesarios, mirar si los campos tenían números o caracteres especiales y mejorar la escritura de las respuestas obtenidas en las encuestas. Para el análisis de datos ayudó a identificar las relaciones entre las variables, también ayudó a identificar lo que es la clase “Mencione que formación complementaria o posgradual realizo” y la relación que existe entre la clase y las variables, Las relaciones son de mucha importancia, puesto que con una buena relación los resultados son más coherentes con los objetivos planteados en este proyecto. Como tal la caracterización de los datos y el análisis de los datos, fue muy importante para el desarrollo de este proyecto, sin una buena caracterización de los datos y sin un buen análisis de estos, el proyecto tendría otros objetivos los cuales no tendrían relación con los objetivos planteados en este proyecto.

El modelo de datos presenta unas relaciones que son: sistema de recomendaciones, el más popular, con esto se hace la relación entre las dos y se desarrolla el algoritmo híbrido, para así lograr un buen modelado de datos y por ende un buen resultado de este.

Los algoritmos que se aplicaron en Weka se le hicieron un análisis y un estudio en el cual se pudo ver el procedimiento de estos algoritmos, se analizaron lo que fue el árbol de decisiones, uno de los algoritmos que es el J48 se evidencia cómo se distribuían los datos en nodos y relaciones es decir el algoritmo J48 tiene una confianza y un margen de error el cual entre más se aplica este algoritmo y más operación haga, define más la probabilidad de error, en cuanto más baja sea la probabilidad el margen de error se ve reflejado en él antes y después de la aplicación del algoritmo, así, por lo tanto, en los resultados se ve reflejado que los árboles de decisión son mucho más pequeños.

El análisis de los resultados en una primera versión no tenía una buena confiabilidad en el margen de error que fue de más del 72% este se hizo en el algoritmo KNN. Al final se hizo un filtrado híbrido al observar el proceso de la construcción de los datos del sistema basado en recomendaciones, se hizo una integración con las técnicas de los algoritmos para que de esta manera arroje un buen resultado para así cumplir con los objetivos del proyecto con un solo modelo del algoritmo híbrido. El algoritmo híbrido es el que mejor desempeño obtuvo, debido a que este soluciona la falta de datos en caso de que el usuario no brinda la información completa, el algoritmo soluciona el problema de que el usuario inicia de cero, pero ya se tiene su perfil entonces se recomienda con base a su perfil y soluciona algo importante y es que si el usuario ya hizo un posgrado el algoritmo le va a recomendar que otro curso podría tomar. Este algoritmo sirve para implementar un sistema de recomendaciones con base en los datos que tiene la universidad tomando en cuenta los formularios que se les envía a los egresados.

6.3 Trabajos futuros.

Como trabajo futuro se presentan las siguientes mejoras y aplicaciones que se le podrían realizar al proyecto y al uso del modelo construido, dependiendo de los intereses que se quieran cumplir, como también las condiciones para poder cumplir. La principal mejora que se puede establecer es la recolección de datos suficientes y adecuados para el cumplimiento de los objetivos de minería de datos que se quieran cumplir, la construcción de un Dataset fue una de las principales dificultades que se tuvieron en el desarrollo del proyecto, esto trajo consigo la toma de decisiones, en cambio, de estrategias, técnicas y herramientas. Si se hace una eficaz recolección de datos, se puede obtener un modelo confiable que arroje mejores resultados.

Teniendo en cuenta la mejora anteriormente propuesta se propone un formulario correspondiente a las variables necesarias requeridas para la construcción del conjunto de datos a utilizar en el modelo resultante del presente trabajo de investigación, se deja a disposición el formulario y la socialización del mismo a egresados que permitieron de manera consentida su realización y utilización de la información y datos obtenidos a través de este, para más información acerca del formulario realizado diríjase al Anexo F.

La aplicación de este algoritmo híbrido a la construcción de un sistema de recomendaciones adaptado a las necesidades de la Universidad Unicomfacauca, que permita visualizar los datos, ver los resultados del modelo aplicado, es decir las medidas de similitud y las recomendaciones posibles, segundo determinado conjunto de datos de prueba que se establezca para su validación. Esto también estaría sujeto a construcción y ejecución de un plan de ingeniería del software para el desarrollo del software o

aplicación web a realizar, esto basado en el algoritmo antes descrito y al sistema de recomendaciones que se vaya a realizar después.

Teniendo en cuenta las propuestas descritas anteriormente se propone ampliar el porcentaje de confiabilidad de la recomendación efectuada por el modelo, teniendo en cuenta que el porcentaje establecido en el objetivo principal del presente proyecto de investigación se escogió basándose en la no consolidación de un conjunto de datos confiable para traer resultados cien por ciento acertados con la oferta de posgrados actual y basado en la caracterización de perfiles más amplia detallada, para hacer la recomendación. A partir de esta propuesta, si bien se tiene un nivel de satisfacción aceptable se puede llegar impactar mejor y directamente en la oferta académica posgradual.

Como última propuesta a trabajo futuro, se deja dispuesto el modelo y los algoritmos aplicados con el fin de ser utilizados en diferentes problemáticas que se puedan desarrollar con la inteligencia de datos, dentro de los procesos de la universidad, como también de los futuros propósitos y oportunidades de los futuros egresados del programa de ingeniería de sistemas, siendo este un tema de gran interés en la actualidad y que tiene muchos temas de investigación y aplicación.

Referencias

Alejandro Benito Santos, R. S. (2018). *Implementación de una arquitectura.*

Altamar, D. D. (2019). Herramienta informática para apoyar los medios de participación del egresado del programa de ingeniería de sistemas de la universidad de Cartagena. *Trabajo de pregrado*. Cartagena de indias.

Ángela García Pérez, A. V. (2018). Procesos de investigación-acción en aprendizaje y servicio. *RIDAS*.

Aranda, M. G. (2008). Importancia de la metodología en los proyectos de investigación. *Universidad Nacional de La Plata*.

Asto Rodríguez, E. M. (2020). Framework basado en minería de datos para la obtención del perfil de egreso de los estudiantes del programa de ingeniería mecatrónica de la universidad nacional de Trujillo. *Tesis de maestría*. Trujillo.

Cardoso, García, Y, Arza, Valdés, L (2017) Algoritmo OneR. Su aplicación en ensayos clínicos. Revista cubana de Ciencias informáticas http://scielo.sld.cu/pdf/rcci/v11n2/rcci05217.pdf

Carlos Hernán Cardona Taborda, n. G. (2016). *Análisis de datos mediante el algoritmo de clasificación J48, sobre un clúster en la nube de AWS.* Universidad distrital Francisco José de Caldas.

Cerón, Ríos M., López, Gutiérrez, D, M., Díaz, Agudo. Belén., Recio, García J. A.(2017) Sistema de recomendación basado en CBR algoritmo para la Promoción de Hábitos más saludable, Popayán. Cauca: Universidad del Cauca.

Chaves Montero, A. (2018). *La utilización de una metodología mixta en investigación social.* Máchala-España: UTMACH.

Colombia, I. (2021). *https://www.ibm.com/*. Obtenido de https://www.ibm.com/: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview

Colombia, I. (2021). *Ibm.com*. Obtenido de ibm.com: https://www.ibm.com/co-es/analytics/data-science-ai

Curo, W. M. (2020). Aplicación Web para la elaboración de perfiles de consumidor basada en minería de datos y arquitectura cloud para el apoyo al proceso de

conversión de leads en la asociación AIESEC en Perú. *Tesis posgradual*. Chiclayo.

David Luis la Red Martínez, M. k. (2015). Perfil de rendimiento académico: Un Modelo basado en Minería de datos. *Dpto. de Ingeniería en Sistemas de Información, facultad Regional Resistencia, Universidad Tecnológica nacional, Argentina*.

Dogan, A., Birant, Derya (2019). A Weighted Majority Voting Ensemble Approach for Classification. IEEE. https://ieeexplore.ieee.org/document/8907028.

Edward Javier Girón Buitrón, C. R. (2015). Modelo de usuario conforme a la norma ISO/TR 14292 para un sistema personalizado como apoyo para la proporción de actividad física y dieta saludable. *Trabajo de pregrado*. Popayán.

Eibe Frank, M. A. (2016). The WEKA Workbench. En M. A. Eibe Frank, *Eibe Frank, Mark A. Hall, and Ian H. Witte.*

Espinoza Mina, M. A. (2018). Weka, áreas de aplicación y sus algoritmos: una revisión sistemática de literatura. *REVISTA CIENTÍFICA ECOCIENCIA*, 5, 1–26.

Francisco Javier Ariza-López, J. R.-A.-F. (2017). Spatial analysis of remote sensing image classification accuracy". Remote Sensing of Environment. *Jaen*.

Francisco Javier Obando Vidal, J. R. (s.f.). Metaheuristica hibrida para la identificación de modelos de autómata celular en trayectorias de simulación de plegamiento de proteína. *Trabajo de pregrado*. Popayán.

García, Jiménez, M., Álvarez, Sierra, A. (S. F.). Análisis de datos en WEKA -Pruebas de selectividad. Universidad Carlos III . http://www.it.uc3m.es/~jvillena/irc/practicas/06-07/28.pdf

Gelman, A. V. (2018). Limitations of "Limitations of Bayesian leave-one-out cross-validation for. *Stat.ME*.

Girón, Buitrón, J, E., Rico, Olarte, C., Cerón, Ríos, M, G., López, Gutiérrez, D, M. (1016). Framework for Data Model to Personalized Health Systems. Cartagena, Bolívar, Colombia.

Gladys Patricia Guevara Alban, A. E. (2020). Metodologías de investigación educativa (descriptivas, experimentales, participativas, y de investigación-acción). *RECIMUNDO*, 4(3), 163-173.

Guillen, V. J. (2013). Aplicación de minería de datos como una metodología para estimar las principales causas de desempleo y subempleo profesional de los egresados de las universidades. *Tesis de maestría*. Honduras.
Guillen, V. J. (2013). Aplicación de minería de datos como una metodología para estimar las principales causas de desempleo y subempleo profesional de los egresados de las universidades. *Tesis de maestría*. Honduras.
Guzmán, J. A. (2016). Plataforma web con integración de minería de datos y redes sociales para el seguimiento a graduados del programa de ingeniería de sistemas de la universidad de Cundinamarca extensión Facatativa. *Tesis de pregrado*. Bogotá.

Herrera, M., Ruiz, S., Romagnano, M., Ganga., Lund, M., y Torres, E.(2019). Aplicando métodos y técnicas de la ciencia de los datos a datos universitarios. http://sedici.unlp.edu.ar/bitstream/handle/10915/76997/Documento_completo .pdf-PDFA.pdf?sequence=1&isAllowed=y

Hernán Joaquín Suárez Rodríguez, R. E. (2017). Sistema que analiza los hábitos de consumo de los clientes del sector de bares y restauran-. *Tesis de posgrado*. Bogotá.

J. Zico Kolter, M. A. (2007). Dynamic Weighted Majority: An Ensemble Method for Drifting Concept. *Journal of Machine Learning Research*

Javier Díaz, L. L. (2016). Personalización de la Educación a través de la creación de Perfiles dinámicos de los alumnos. *Laboratorio de Investigación en Nuevas Tecnologías Informáticas. Facultad de Informática. Universidad Nacional de La Plata*.

Jose Manuel Molina López, J. G. (2006). Técnicas de análisis de datos, aplicaciones prácticas utilizando Microsoft Excel y Weka. *Título de pregrado*. Madrid.

Joseph Alexander Rubio Tapias, C. A. (2018). Desarrollo de componentes web parametrizable de clustering para análisis de datos. *Trabajo de pregrado*. Bogotá.

Juan Miguel Moine, S. G. (2011). Análisis comparativo de metodologías para la gestión de proyecto de minería de datos. *Repositorio institucional de la universidad nacional de la plata*, 3,4.

León lipe, J. N. (2013). Propuesta de gestión de información de agrupamiento (clustering), utilizando técnicas de minería de datos para el egresado del programa profesional en ingeniería de sistemas de la universidad católica de Santa María. *Tesis para optar el título profesional de ingeniero de sistemas*.

Lorente-Leyva, I. D.-G. (2018). *Optimización del transporte universitario mediante el método de jerarquías de contracción y algoritmos de agrupación.* Springer, Cham.

Ospina, C. A. (2013). *Uso de minería de datos para el análisis y seguimiento de estudiantes de acuerdo a su perfil de visitas en página web.* (Tesis de pregrado). Universidad de los Andes, Bogotá.

PAUTSCH, J. G. (2009). Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en sistemas de computación. *Tesis de pregrado*. Misiones.

Piza Burgos, N. D. (2019). Métodos y técnicas en la investigación cualitativa. Algunas precisiones necesarias. *Revista Conrado*.

Quiroz Martínez, M. Á. (2020). Modelo de recomendación basado en conocimiento para el desarrollo del pensamiento del trabajo con objetos de aprendizaje. *Revista Conrado*.

Riquelme., e. a. (2013). Riquelme., et al (2020) aseguran que el algoritmo de KNN. *Universidad de Sevilla*.

Roberto Hernández Sampieri, C. F. (2020). *METODOLOGÍA.* Montreal-Canadá.

Rodolfo Mosquera, L. P.-O. (2016). Metodología para la Predicción del Grado de Riesgo.*) Universidad Nacional de Colombia, Facultad de Ingeniería y Arquitectura*.

Rodrigo, J. A. (2021). Validación de modelos predictivos: Crossvalidation, OneLeaveOut, Bootstraping.

Rodríguez, E. M. (2017). Lineamientos teóricos y metodológicos de la investigación. *In Crescendo. Institucional*.

Russo, C. (2019). Minería de datos aplicada a estrategias para minimizar el rezago académico y la deserción universitaria en carreras de informática de la UNNOBA. *Tesis doctoral*. La Plata.

Sangrario, R. G. (2011). Minería de datos en encuestas de profesores al fin de semestre de la facultad de ingeniería UNAM. *Tesis de pregrado*. Ciudad de México.

Schiaffino, S. (2018). *Clustering.*

Silva, M. A. (2015). Diseño y Modelado de la Base de Datos del Prototipo de Formulario. *Universidad de Guayaquil*.

Timarán-Pereira, S. R.-A.-Z.-T. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas. *Ediciones Universidad Cooperativa de Colombia*.

Torres, K. R. (6 de febrero de 2019). *https://labscn-unalmed.github.io/*. Obtenido de https://labscn-unalmed.github.io/: https://labscn-unalmed.github.io/ecologia-numerica/guiones/distancias_disimilitudes_matriz_discrepancia.html

Ubillús, E. M. (2016). Evaluación de la pertinencia de maestrías en ingeniería: aplicación en la universidad de Piura, Perú. *Escuela técnica superior de ingenieros agrónomos*.

Unicomfacauca, C. u. (2021). *https://www.unicomfacauca.edu.co/*. Obtenido de https://www.unicomfacauca.edu.co/: https://www.unicomfacauca.edu.co/

Vallejo., et al (2018). Minería de datos. Revista científica mundo de la investigación y el conocimiento. https://dialnet.unirioja.es/servlet/articulo?codigo=6732870

Vences-Nava Rodrigo, M.-P. S. (2019). Evaluación de un sistema de recomendación híbrido de trabajos de titulación. *Ingeniería Investigación y Tecnología*.

Vera, C. V. (2011). Creación de perfiles de deudores de crédito universitario, para el mejoramiento de campañas de cobranza, usando minería de datos. *Universidad Austral Chile*.

VITE CEVALLOS, H., & CARVAJAL ROMERO, H. y. (2020). Aplicación de algoritmos de aprendizaje automático para clasificar la fertilidad de un suelo bananero. *Conrado [online]*.

Anexos

Anexo A. Identificación de categorías para cada atributo**.**

Proceso de identificación de los valores únicos por cada atributo, para esto se utilizó la función (unique) de la librería panda, con el fin de utilizar esta identificación en el proceso de categorización de los perfiles de los egresados realizado en el apartado 7.3.2 ajuste de parámetros.

```
data["programa"].unique()
```

```
array(['tecnologia agroambiental', 'comunicacion social y periodismo',
       'ingenieria de sistemas', 'ingenieria mecatronica',
       'tecnologia de la informacion y comunicacion',
       'tecnologia gastronomia', 'contaduria publica',
       'ingenieria industrial'], dtype=object)
```

```
data["tipo_programa"].unique()
```

```
array(['diplomado', 'especializacion', 'curso', 'maestria', 'seminario'],
      dtype=object)
```

```
data["quiere_estudiar"].unique()
```

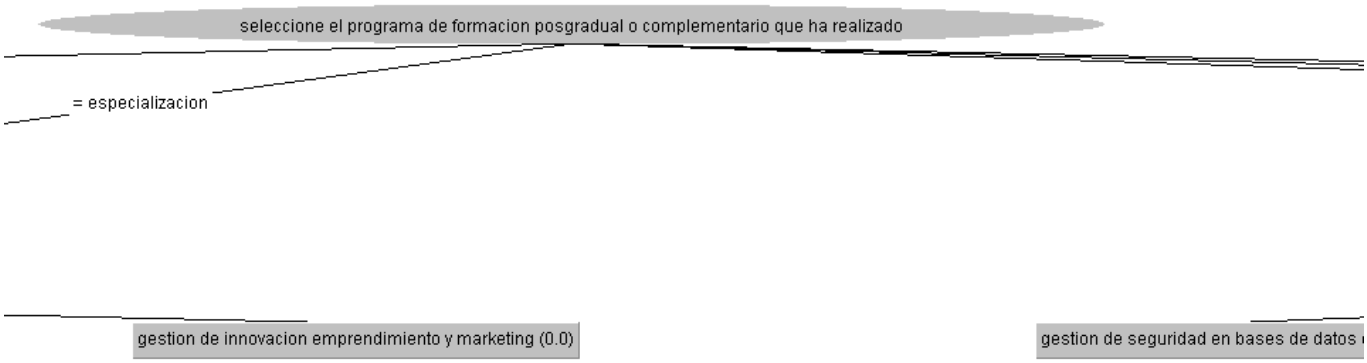
```
array(['especializacion', 'maestria', 'seminario', 'diplomados'],
      dtype=object)
```

```
data["quiere_estudiar"] = data["quiere_estudiar"].replace("diplomados", "diplomado")
data["quiere_estudiar"].unique()
```

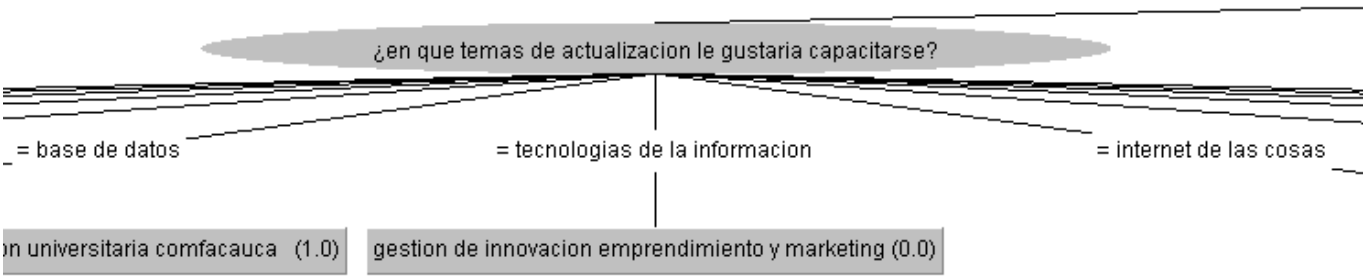
```
array(['especializacion', 'maestria', 'seminario', 'diplomado'],
      dtype=object)
```

Anexo B. Árbol de decisión resultante del algoritmo J48 aplicado en Weka.

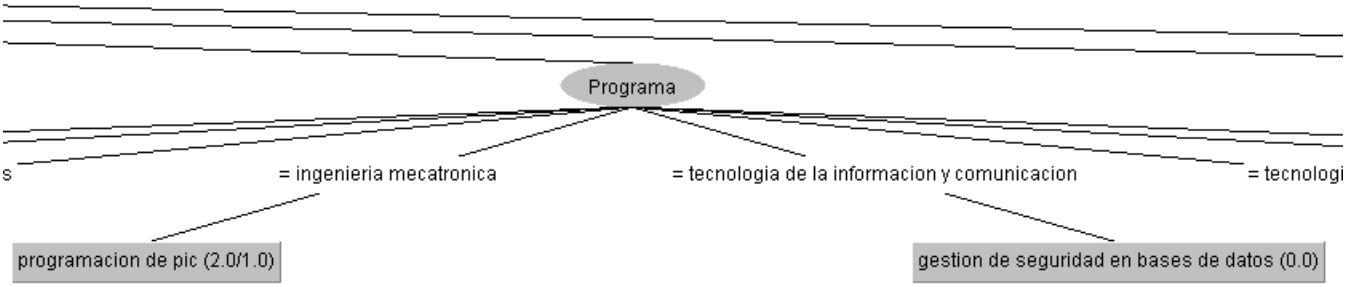
Evidencia del árbol de decisión generado por el algoritmo J48 aplicado en Weka, en partes visibles teniendo en cuenta la amplitud de este. Partiendo de su nodo principal a sus secundarios.



Nodo principal

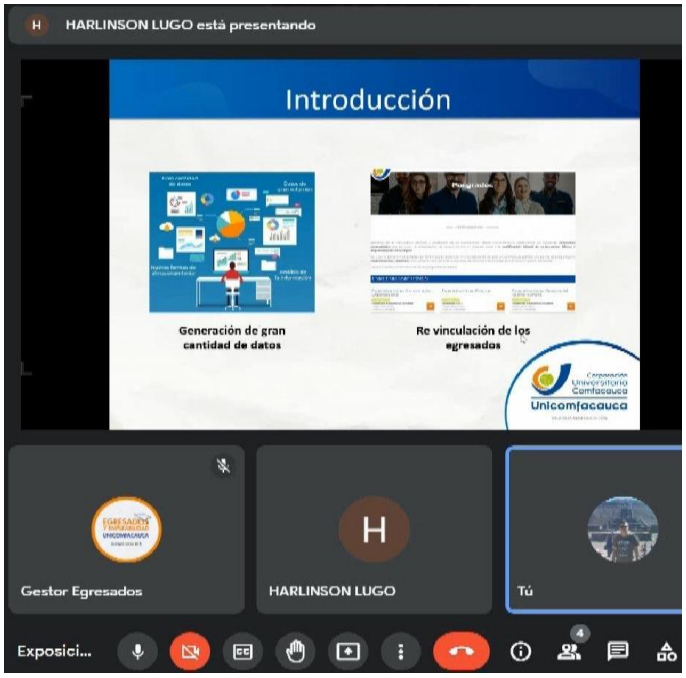


Nodo a la izquierda

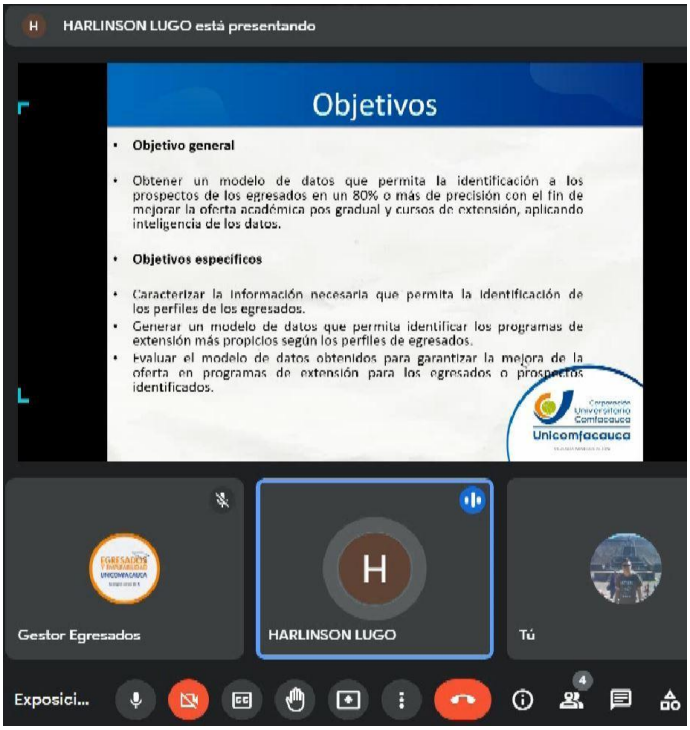


Nodo a la derecha

Anexo C. Evidencia de la presentación oral realizada al personal de egresados de la corporación.

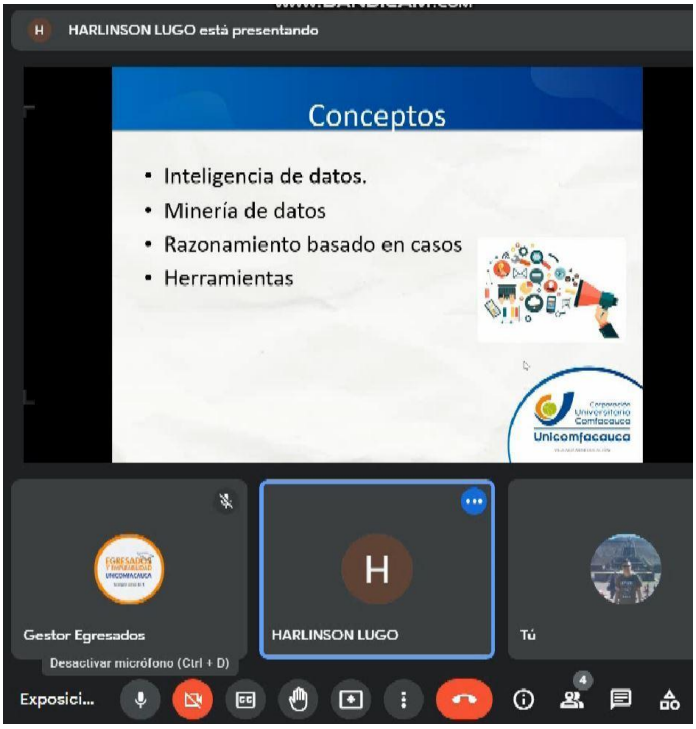


Presentación oral del contexto, justificación y formulación del problema

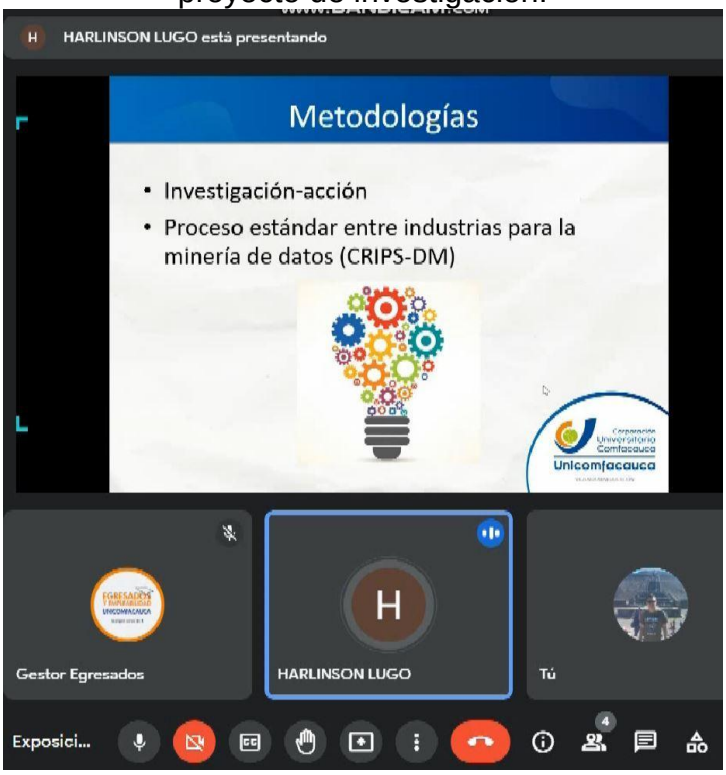


Presentación escrita y oral de los objetivos del proyecto

Anexo C1

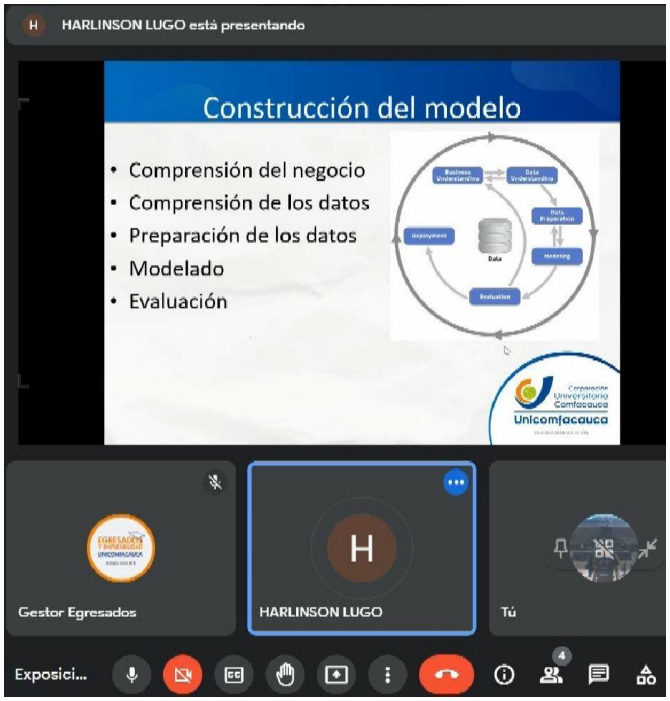


Presentación de los conceptos, términos, y herramientas utilizadas en el desarrollo del proyecto de investigación.

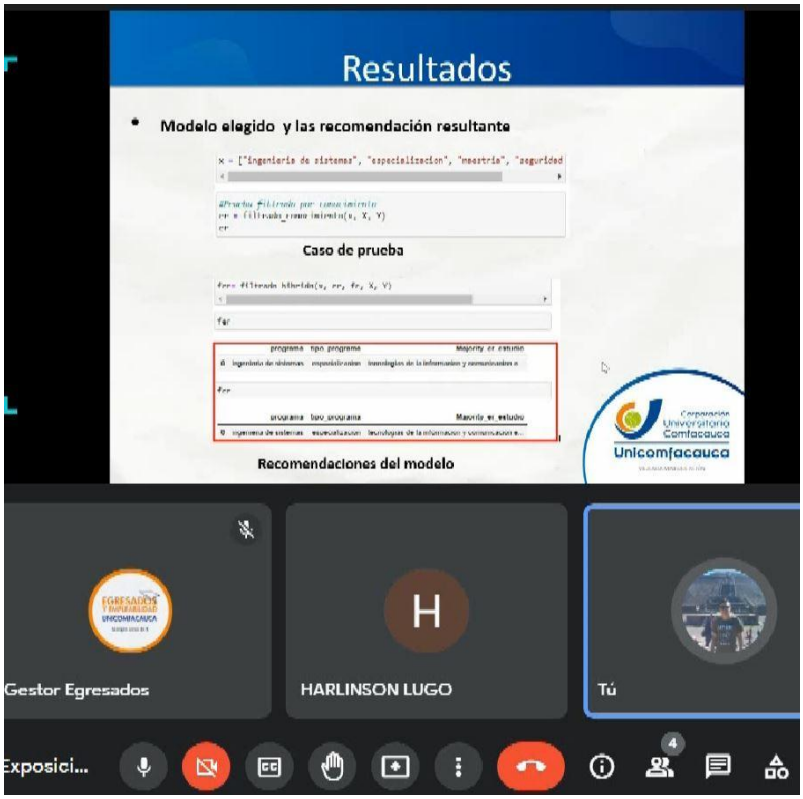


Descripción de las metodologías empleadas en el proceso de investigación y desarrollo del modelo.

Anexo C2

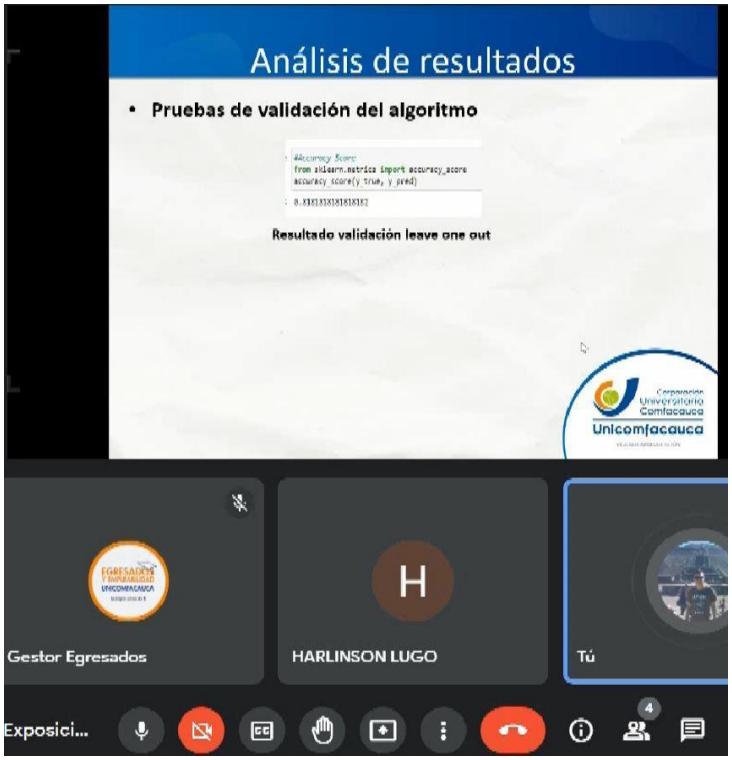


Descripción oral y gráfica de la construcción del modelo utilizando la metodología CRIPS-DM.

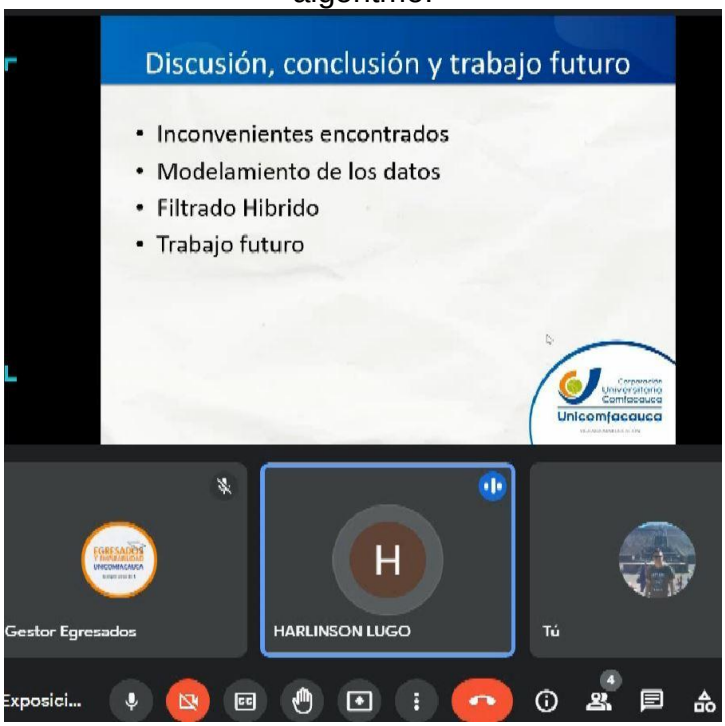


Presentación de los resultados obtenidos al aplicar el modelo escogido, teniendo en cuenta un caso de prueba.

Anexo C3

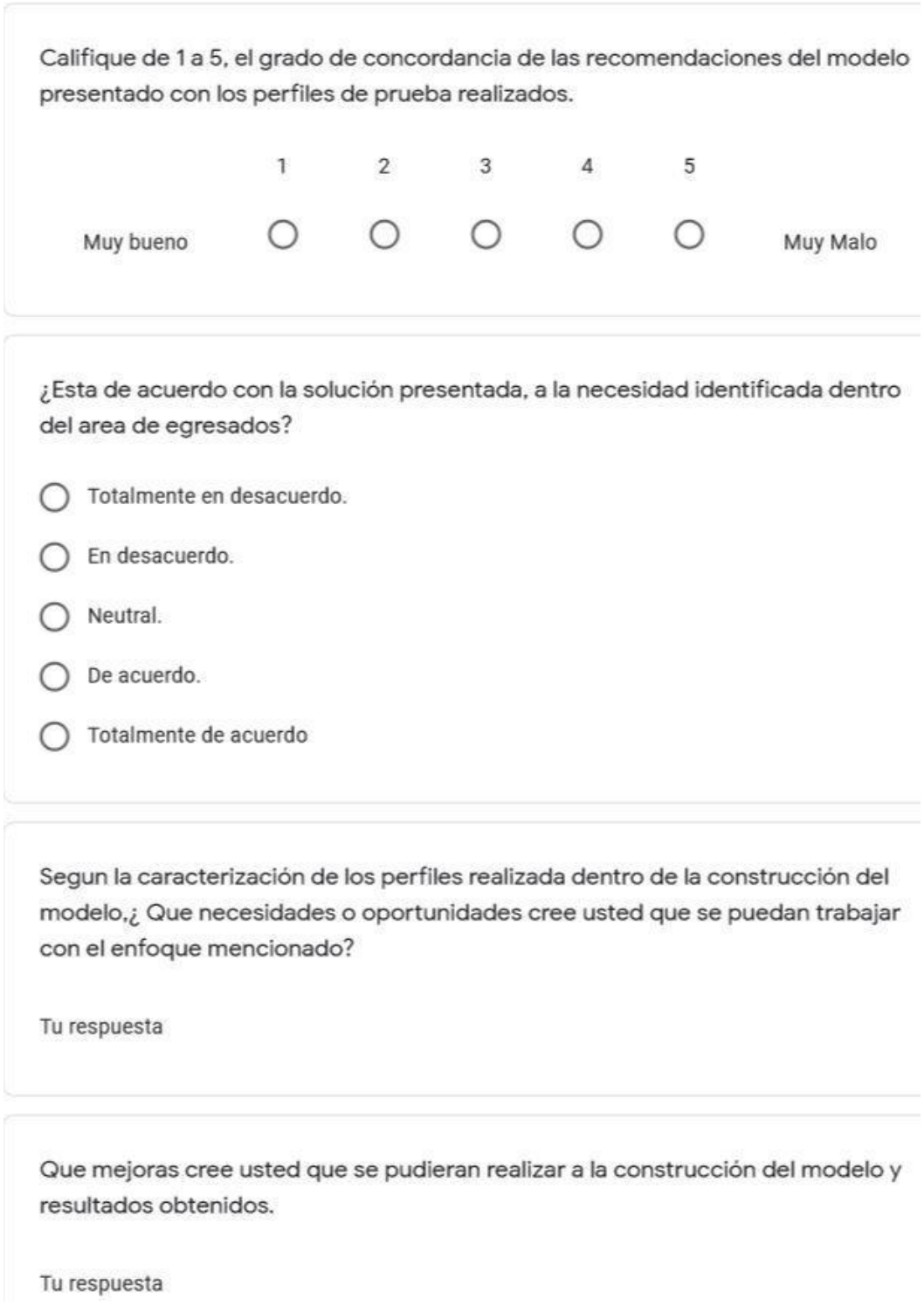


Descripción del análisis de los resultados y presentación de la validación realizada al algoritmo.



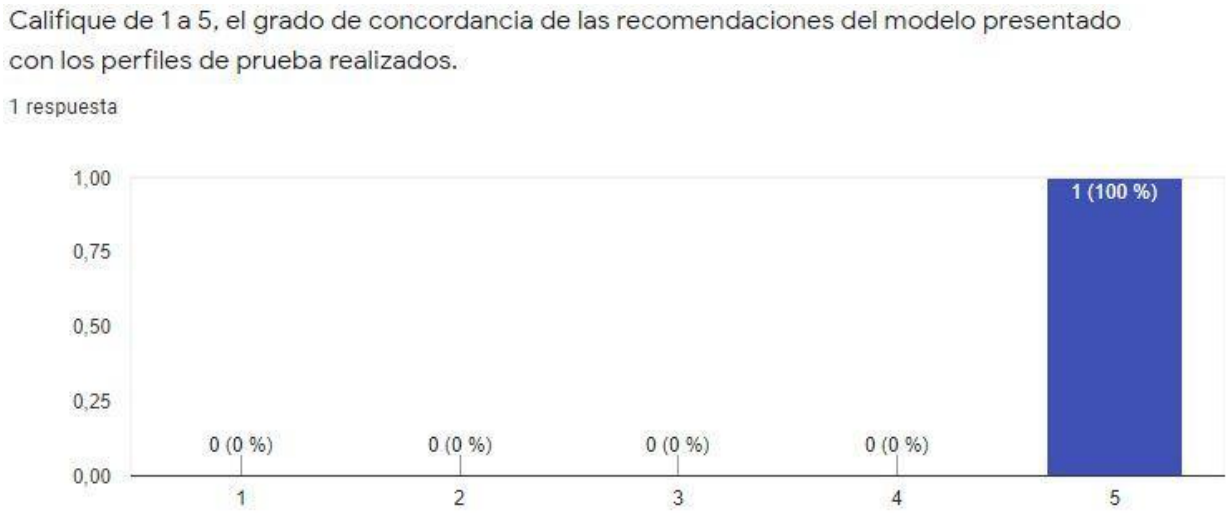
Presentación de las conclusiones, y propuesta de trabajos futuros.

Anexo D. Formulario de satisfacción presentado en la prueba de validación realizada por el gestor de egresados.

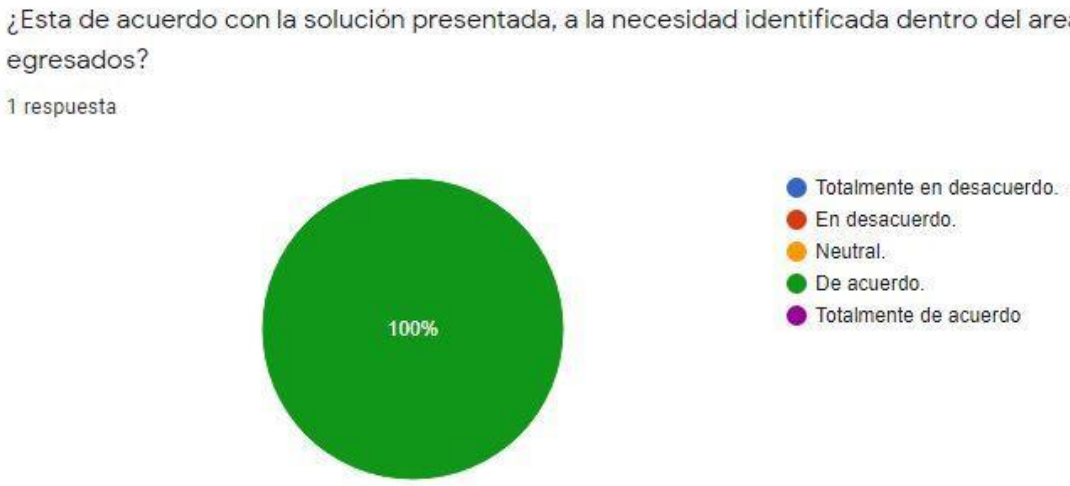


Visualización de la segunda página del formulario de validación presentado al gestor de egresados.

Anexo E. Resultados del formulario de satisfacción realizado por el gestor de egresados.

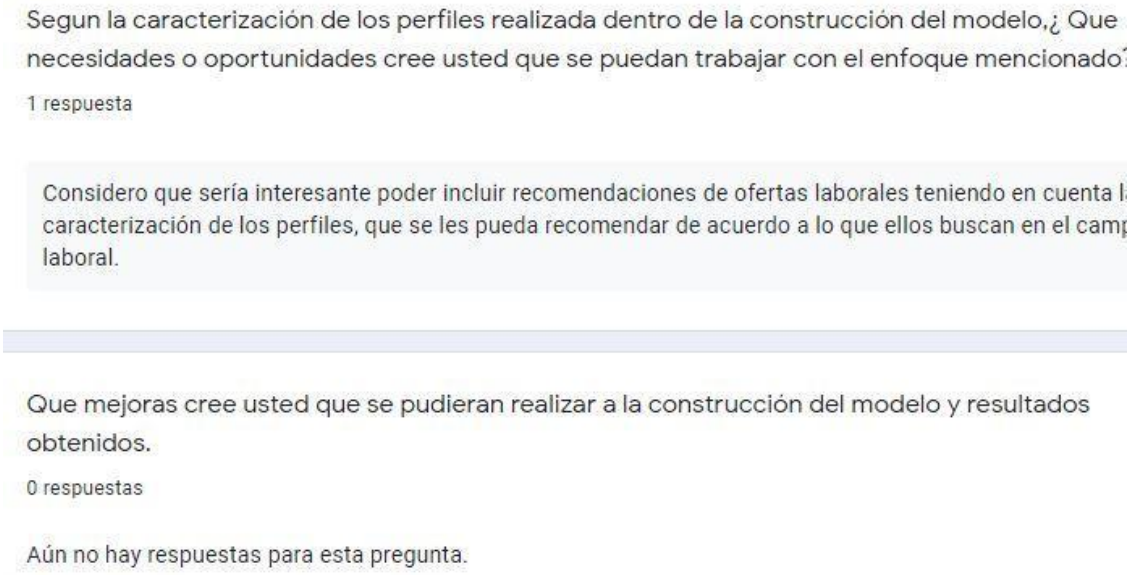


Resultado de pregunta de satisfacción pregunta dos



Resultado de pregunta de satisfacción pregunta tres

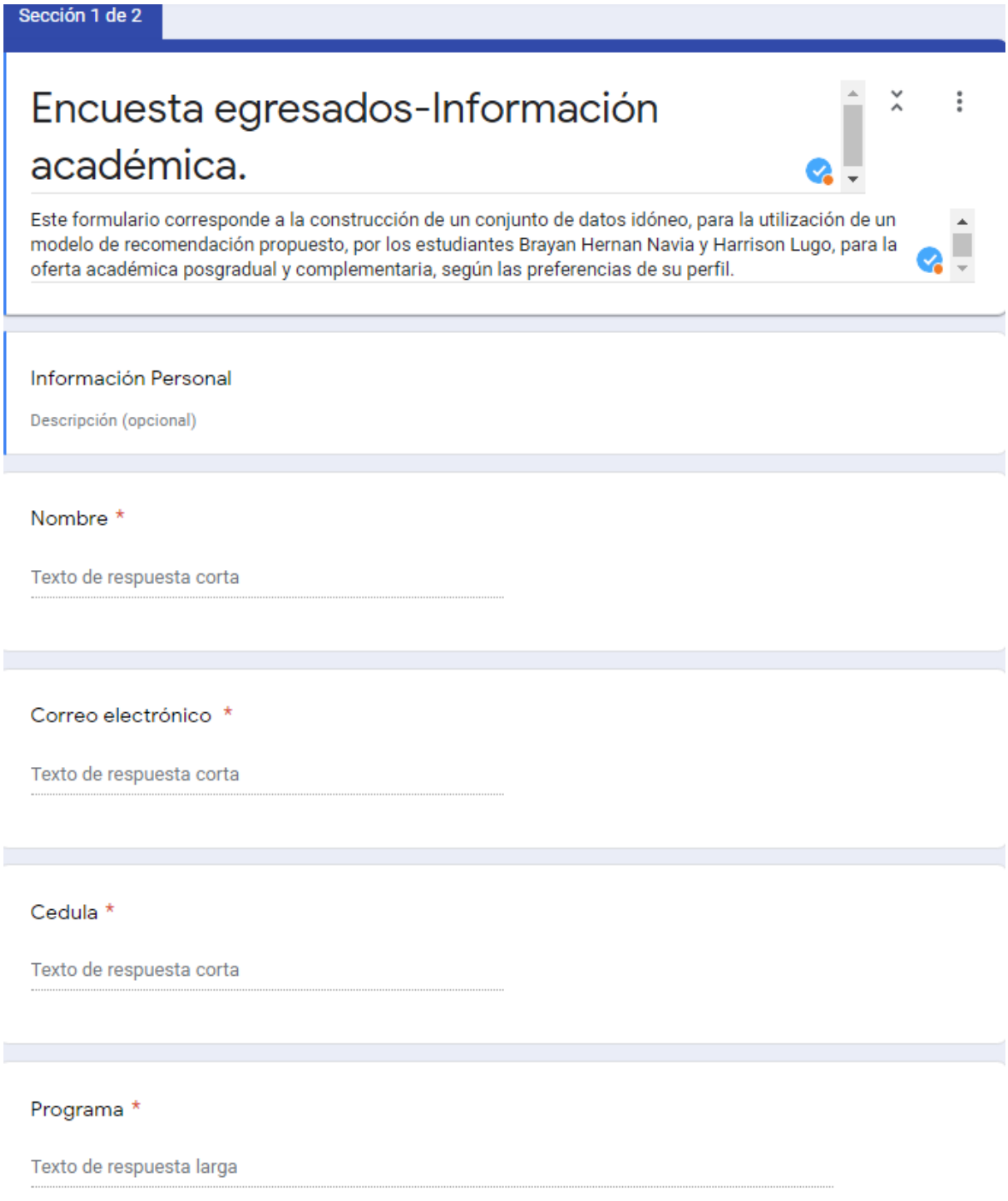
Anexo E1.



Resultado de las preguntas de opinión.

Anexo F

A continuación se presenta el formulario propuesto para la futura mejora de levantamiento de datos, con el fin de utilizar el modelo propuesto.



Sección 1 de 2

Encuesta egresados-Información académica.

Este formulario corresponde a la construcción de un conjunto de datos idóneo, para la utilización de un modelo de recomendación propuesto, por los estudiantes Brayan Hernan Navia y Harrison Lugo, para la oferta académica posgradual y complementaria, según las preferencias de su perfil.

Información Personal

Descripción (opcional)

Nombre *

Texto de respuesta corta

Correo electrónico *

Texto de respuesta corta

Cedula *

Texto de respuesta corta

Programa *

Texto de respuesta larga

Anexo F1

Segunda sección del formulario propuesto.