INTELIGENCIA DE NEGOCIO APLICADA AL ÁREA DE PERMANENCIA DE LA CORPORACIÓN UNIVERSITARIA UNICOMFACAUCA

Luis Felipe Vivas Trujillo
Cristian Camilo Mariño Buitrón

Titulación
Proyecto investigativo

Director(a):
Phd. Gineth Magaly Cerón Ríos

Codirector:
Mg. Francisco Javier Obando

Corporación Universitaria Comfacauca – Unicomfacauca
Ingeniería de sistemas
Popayán
Noviembre
2021

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

Se realizó un tablero con la herramienta Power BI para facilitar la toma de decisiones frente a la permanencia estudiantil y gestión de datos históricos relacionados con la Corporación Universitaria Comfacauca – Unicomfacauca, este

proceso se llevó a cabo porque actualmente las universidades tanto públicas como privadas luchan constantemente por incrementar el porcentaje de permanencia de los estudiantes, para llevar a cabo la implementación de este sistema se hizo necesario una investigación para la identificación de variables de deserción, como también, el análisis del modelado de datos, la arquitectura del *datamart*, el diseño centrado en el usuario del sistema.

Para ello se usó una metodología iterativa basada en la observación, identificación del problema, desarrollo de la tecnología y pruebas de campo. Para el modelado de datos CRISP-DM que consta de comprensión de los datos, preparación de los datos, modelado, evaluación e implantación. Para el diseño del sistema diseño centrado en el usuario y para la arquitectura del DW enfoque Kimball. De ahí se desarrolló el sistema de inteligencias de negocios que ayuda al soporte de decisiones del área de permanencia académica.

De este trabajo se obtuvieron contribuciones de investigación científica como el sometimiento de un artículo a una revista tipo B según Colciencias, 3 ponencias en diferentes eventos, la apropiación social del tablero BI por la Unicomfacauca, la transferencia de conocimiento del mismo a las personas que lo requieren como coordinador de permanencia, decanos, directivos etc. Por otro lado, se lograron avances significativos en nuevo conocimiento como la limpieza de datos con CBR, predicción en Power Bi con KNN, además del manejo de una herramienta como Power BI.

De los resultados más destacados se puede concluir que la categoría socioeconómica afecta la permanencia de los estudiantes aumentando el número de desertores, al igual que su estado civil que influye en la indecisión de continuar con una carrera o escoger otra. También se puede concluir que el algoritmo que mejor predicción dio fue el KNN y CBR en limpieza. Además, que, con el modelo de datos, se deja un protocolo de relaciones que se puede implementar en otros tableros en diferentes instituciones educativas universitarias, ya que se cuenta con datos que la mayoría tiene.

ABSTRACT

A dashboard was made with the Power Bi tool to facilitate decision-making regarding student permanence and management of historical data related to the comfacauca -Unicomfacauca

university corporation, this process was carried out because currently both public and private universities are constantly fighting To increase the percentage of permanence of the students, to carry out the implementation of this system, an investigation was necessary to identify dropout variables, as well as the analysis of data modeling, the architecture of the datamart, the focused design on the system user.

For this, an iterative methodology based on observation, problem identification, technology development and field tests was used. For CRISP-DM data modeling consisting of data understanding, data preparation, modeling, evaluation and implementation. For the design of the system user-centered design and for the architecture of the DW Kimball approach. From there, the business intelligence system was developed that helps to support decisions in the area of academic permanence.

Scientific research contributions were obtained from this work, such as the submission of an article to a type B journal according to Colciencias, 3 presentations at different events, the social appropriation of the BI board by Unicomfacauca, the transfer of knowledge from it to the people who asked it. require as permanence coordinator, deans, managers etc. On the other hand, significant advances were made in new knowledge such as data cleaning with CBR, prediction in Power Bi with KNN, in addition to the use of a tool such as Power BI.

From the most outstanding results, it can be concluded that the socioeconomic category affects the permanence of the students, increasing the number of dropouts, as well as their marital status, which influences the indecision to continue with a career or choose another. It can also be concluded that the algorithm that gave the best prediction was KNN and CBR in cleaning. In addition, with the data model, a relationship protocol is left that can be implemented in other boards in different university educational institutions, since there is data that most have.

INTRODUCCIÓN

La deserción estudiantil universitaria es una problemática que afecta a las universidades, en especial las instituciones privadas, quienes necesitan la contribución económica de los estudiantes para la continua prestación de servicios, por lo tanto, este aspecto conlleva a que estas organizaciones promulguen diferentes estrategias para combatir la problemática de la deserción.

Actualmente las universidades de educación superior buscan contar con un sistema de alertas con la intención de hacer un seguimiento a sus estudiantes con riesgo de deserción, tomando como referencia variables obtenidas a través de consulta de la literatura, como lo hace [1] quien desarrolla un software de alertas para el apoyo a la permanencia del estudiante, considerando variables proporcionados en estudios previos realizados por el Ministerio de Educación Nacional para detectar riesgos de deserción.

Partiendo de lo dicho anteriormente, la Corporación Universitaria Unicomfacauca en el área de permanencia académica ha venido adelantando un sistema de seguimiento a la permanencia estudiantil denominado SISPE el cual permite tener seguimiento y control de las intervenciones propuestas para aquellos estudiantes en riesgo a desertar. No obstante, este sistema carece de un análisis profundo e histórico de los datos, en donde se correlacionen los diversos factores influyentes en la permanencia académica, todo esto presentado al usuario final mediante una visualización intuitiva y clara.

Por lo cual es indispensable la implementación de un sistema de inteligencia de negocio (en adelante SIN) que ofrezca lo mencionado anteriormente, de tal manera que supla las carencias que presenta actualmente el SISPE, logrando facilitar aún más la toma de decisiones en el área de permanencia académica de la corporación.

El presente trabajo está integrado por 5 capítulos; el primero muestra condiciones generales como objetivos y el problema abordado. El segundo describe el modelado de datos para identificar variables para el ETL. El tercer capítulo se lleva a cabo la extracción, transformación y carga de datos en el *DataMart* para realizar sus respectivos análisis con la herramienta POWER BI. El cuarto capítulo describe la arquitectura del sistema y la evaluación de expectativas del usuario final. En el último capítulo se tienen las conclusiones y recomendaciones a las que se llegó con la realización de este proyecto.

1.1. Planteamiento del problema

El mundo va evolucionando y mejorando constantemente en todos su aspectos (laboral, educación, salud entre otros) quienes adoptan tecnologías de la información para albergar gran cantidad de datos generados con el pasar del tiempo y así llevar un mejor control de las operaciones organizacionales, sin embargo, no basta solo con almacenar los datos sino que hay que profundizar en el comportamiento de los mismos, maximizando conocimientos para tomar mejores decisiones y de esta manera formar una organización más competitiva.

Muchas instituciones de educación superior cuentan con datos, pero no cuentan con herramientas avanzadas que hagan un buen proceso para proporcionar información útil mediante la visualización de los mismos, por lo tanto, esto conlleva a no tener suficientes argumentos al momento de tomar una decisión, ocasionando dificultades al momento de emprender el camino hacia el alcance de metas y objetivos organizacionales. Por ejemplo: una universidad puede tener un datos de sus estudiantes como estado civil, edad, estrato económico, notas entre otros, pero estos son solo datos, para encontrar información relevante que ayude a la toma de decisiones en el área de permanencia académica, es necesario hacer un análisis de estos datos, mostrando una visualización de los mismo para determinar en qué medida incide cada variable.[2]

Por otro lado, las universidades como la nacional de Cajamarca de Perú[3], la nacional de Trujillo [4], entre otros, también han implementado este tipo de SIN obteniendo buenos resultados, el primer estudio construye un *DataMart* para hacer un seguimiento a la parte académica creando reportes mediante el programa *Power BI*, el segundo trabajo hace un estudio para determinar el nivel de impacto de un SIN, implementan el SIN usando la herramienta Pentaho, como resultado se obtuvo que la SIN impacta significativamente.

Como lo hacen otras universidades, la corporación universitaria Comfacauca quisiera entrar en ese proceso, específicamente el área de permanencia académica, debido a que cuentan con datos, pero no poseen un sistema que permita su visualización y ayude con aporte de conocimiento para la toma de decisiones, como, por ejemplo, visualizando los estudiantes con tendencia a ser balanceados creando como medida preventiva asesoría en temas donde se requiera fortalecimiento de conocimientos.

La Corporación Universitaria Comfacauca – Unicomfacauca cuenta con un sistema de alertas llamado SISPE, el cual genera información relacionada con el

riesgo de deserción estudiantil; el sistema presenta las diferentes alertas asociadas al estudiante, como por ejemplo: alertas financieras (estado de pago matricula), alertas académicas (estado de notas, inasistencias, asesorías), condición vulnerable (repitencia), además un *dashboard* informativo acerca de las alertas generadas por categoría , entre las funciones más resaltables. Sin embargo, este sistema carece de un modelado de datos que implemente técnicas para detectar información dentro de grandes volúmenes de datos, que conlleva a que la organización no obtenga conocimientos que faciliten un análisis profundo e histórico de los datos, en donde se correlacionen los diversos factores influyentes en la permanencia académica; por ejemplo este sistema solo indica si el estudiante "y" tiene un riesgo a la deserción alta, media, baja o si requiere una intervención, esto, sin dar mayores detalles de la raíz de la problemática, como lo podría hacer la SIN, que ofrece una visualización intuitiva y clara facilitando la toma de decisiones frente a la deserción estudiantil.

Expuesto lo anterior este trabajo trata de traer a discusión la siguiente interrogante ¿Es realmente factible que la implementación de un sistema de inteligencia de negocio en el área de permanencia académica de la Corporación Universitaria Comfacauca – Unicomfacauca que permita la visualización de los datos suministrando información para la toma de decisiones al área de permanencia académica?

1.2. Objetivos del proyecto

1.2.1. Objetivo general

Implementar un Sistema de inteligencia de negocio (SIN) para la detección temprana de estudiantes propensos a abandonar sus estudios universitarios de la Corporación Universitaria Comfacauca - Unicomfacauca y de esta manera apoyar a la toma de decisiones dentro de la corporación.

1.2.2. Objetivos específicos

- Diseñar un SIN de negocios que aporte a toma de decisiones en el área de Permanencia académica de la Corporación Universitaria Comfacauca – Unicomfacauca

- Implementar un SIN que permita visualizar e identificar a los estudiantes propensos a desertar de la Corporación Universitaria Comfacauca – Unicomfacauca

- Evaluar la efectividad de un SIN realizando una prueba de conceptos (por medio un *Focus Group*) con el área de bienestar y permanencia académica de la Corporación Universitaria Comfacauca - Unicomfacauca.

1.2. Estado del arte

Con el día a día, las organizaciones generan una gran cantidad de datos, provenientes de diferentes fuentes de información como, archivos de texto, hojas de cálculo y bases de datos transaccionales, entre otros. Estas herramientas permiten guardar datos más no transforman estos datos en información de valor, que permitan ver indicadores para dar respaldo a la toma de una decisión, [5] señala que la información es un recurso de gran valor que debe ser bien tratada y administrada, entonces las organizaciones incluidas las universidades deben contar con un buen sistema como lo son los SIN, para alcanzar meta y objetivos, debido a esto, según [6] se hace necesario hacer uso de métodos más eficientes para el tratamiento de datos de las organizaciones, es importante que el sector académico como las universidades adopten estas medidas para ser más competitivas, debido a esto construyen una bodega de datos y mediante tableros de control poder observar información clave como rendimiento académico por materia, carrera, ciclo ETC. por otro lado, como se expone en [7], advierte que el panorama educativo actual cuenta con una deficiencia a la hora del manejo de información, no cuenta con indicadores que permitan medir el desempeño de un estudiante, especialmente del egresado, por tal motivo se pone en funcionamiento un SIN para construir gráficas que muestre indicadores como el trabajo realizado como opción de grado: tesis, práctica profesional etc.

En los trabajos de investigación [8][9][3][10][11][12]se evidencia un problema en común, donde diferentes universidades no contaban con una herramienta tecnológica que pudiese convertir los datos en información útil y confiable que respalde la toma de decisiones para cumplir con sus respectivos objetivos, por consiguiente, en busca de una solución, las diferentes universidades optan por implementar un SIN bajo la metodología Kimball. En [8]dice que la universidad cañete no contaba con una base de datos para ser análisis de los datos, por lo tanto, se implementa un SIN que permite hacer un resultado analítico de los datos teniendo en cuenta indicadores como asignaturas matriculadas, aprobadas, desaprobadas, número de créditos, promedio entre otros, obteniendo como resultado una mayor satisfacción por parte de los tomadores de decisiones debido a que los reportes se hacen en poco tiempo. [9]Desarrollan un *Data Mart* (en adelante DM) para disminuir el porcentaje de deserción de los estudiantes, con esta tecnología se descubre cuáles son los cursos que más se dificultan a los estudiantes entre otros. [3]Manifiesta que la creación de una DM en la

Universidad Nacional de Cajamarca ayudó a visualizar la información de mejor manera como notas del estudiante, información del mismo, docente, cursos dictados ETC, por lo tanto, un mejor seguimiento académico. [11]Hace uso de un SIN para establecer un perfil de deserción del estudiante, teniendo en cuenta indicadores como carrera, semestre cursado, materias cursadas, modalidad de ingreso entre otros para tener un apoyo a la toma de decisiones. [12]Este trabajo manifiesta que la universidad ha tenido problemas a la hora de tomar decisiones como subir la cantidad de vacantes o aumentar nuevas escuelas, debido a esto, se desarrolla una bodega de datos teniendo en cuenta las siguientes medidas destacadas: estudiantes matriculados por materia, aprobados, desaprobados, inhabilitados entre otros.

Como resultado de la implementación del SIN se obtuvo un impacto positivo, pues ofrece a los encargados de la toma de decisiones reportes de información mediante gráficas que ayuda a entender de forma clara, sencilla y rápida cómo se están comportando los diferentes aspectos del ámbito académico, cabe señalar que, estas universidades son más competitivas al tener información confiable que brinde un apoyo a la toma de decisiones.

Existen múltiples herramientas para llevar a cabo la SIN, entre ellas está *Power BI* y la suite de Pentaho (proporciona herramientas como *Pentaho reporting* para generar informes y *Pentaho Dashboard* para ver los datos, entre otras), por otro lado en [13]se realiza un estudio comparativo entre herramientas "*open source"* y de propietario, los resultados indican que la mejor herramienta para llevar a cabo los SIN es el software propietario Tableau (ofrece herramientas "Tableau pre" encargada de dar forma y limpiar los datos, "*Tableau desktop*" enfocada a la visualización de datos) , sin embargo la herramienta Pentaho fue la más destacada de la categoría *open source*, es recomendable para instituciones educativas que no cuenta con los recursos económicos suficientes para implementar un SIN.

En [4] al colocar en marcha esta tecnología obteniendo buenos resultados en la gestión académica, sin embargo existen universidades que implementan un SIN para indagar sobre algo más específico, como en [14]realizan una integración de diferentes bases de datos para la construcción de un DM que permita dar información detallada sobre los docentes como por ejemplo: ¿En qué facultad los profesores dominan más el tema? , su trabajo presencial, tutorial y actitudinal para poder tomar decisiones.

Adicional a esto [15] implementa el uso de DM que permita enfrentarse a procesos como la acreditación, el DM va enfocado a la parte docente que permita obtener indicadores para medir la productividad académica, indicadores como la distribución del personal docente en diferentes sedes e institutos, cantidad de

horas en la que el docente ejerce su labor que se puedan comparar con la cantidad de estudiantes matriculados. Con el uso del DM se observa que la Universidad está haciendo una buena labor, ya que los resultados arrojados en los reportes son buenos.

En [7] aclara que el panorama educativo tiene deficiencias a la hora del manejo de la información, por esta razón es que [16], [17] hacen uso de SIN para obtener gráficas con indicadores que permitan medir el desempeño de estudiantes, docente u otros aspectos, que hacen DM obteniendo como resultado un menor tiempo para realizar las tareas y mayor satisfacción por parte de las personas encargadas de tomar una decisión.

[18] Lleva a cabo un SIN enfocada a brindar información a la oficina de admisión de la Universidad José María Arguedas, implementando gráficas que muestran información como cantidad de postulantes por colegio, lugar de donde pertenecen entre otros. Se logra identificar entre varias cosas cuales son las carreras con mayor demanda, en base a esto, el SIN aplicado en el ámbito académico evidencia la eficiencia de la gestión académica proporcionando información rápida y confiable para la toma de decisiones.

Los procesos de globalización y el rápido desarrollo de los sistemas de información han llevado a una alta competencia dentro de la organización no solo entre empresas sino también entre universidades. Debido a esto, en [19] se concluye que es vital la correcta integración de datos debido a su capacidad de extraer conocimientos útiles de heterogeneidad de datos y anomalías. Evidenciando que al realizar la integración de datos previamente al proceso ETL redujeron significativamente el tiempo necesario para el proceso de ETL de datos de estudiantes.

En [20] se encontró que el rendimiento académico de un estudiante de pregrado es una evidencia útil para la determinación de patrones que ayudan a establecer aquellos futuros estudiantes que demuestren señales tempranas de bajo rendimiento académico para su asesoría temprana por parte de los docentes. En [21] se pudo encontrar relación significativa en el bajo rendimiento académico con las discapacidades tanto psicológicas como físicas que puede llegar a presentar un estudiante de educación superior a la hora de iniciar su etapa universitaria, gracias a un sistema de toma de decisiones se puede llegar a determinar qué estudiantes puede llegar a necesitar asesoría incluso si no llegan a reportar su situación de discapacidad.

Por otro lado [22] se evidencia gran utilidad de un SIN a la hora de ayudar a las universidades en la selección del estudiante apropiado, además de ofrecer un amplio campo de análisis en diferentes campos relacionados con la educación superior en la l época actual. En [23] se constató que el SIN sobre el sistema de

educación superior de cualquier país puede llegar a ser realmente útil en la monitorización, análisis y pronóstico de tendencias futuras en la educación superior.

Gracias a la revisión que hace [24] acerca de los hallazgos generales relacionados con proyectos de SIN en 11 universidades del reino unido se puede llegar a dilucidar que para la toma de decisiones en la educación superior es realmente importante el tiempo de decisión y el análisis comportamental, ya que los datos de los estudiantes se generan en tiempo real dando una ventaja competitiva enorme. En la referencia [25] se hizo un estudio comparativo entre 106 artículos científicos acerca de la implementación de inteligencia de negocio en universidades, analizando componentes de SIN utilizados y su combinación adecuada para un análisis de la información eficiente, arrojando como resultado que el SIN era usado orientado a la gestión administrativa , la metodología de cuadro de mando integral (BSC) se enfoca en el aseguramiento de la calidad (certificados de calidad) , el *data mining* se centró específicamente en el rendimiento académico y finalmente el DW fue de manera transversal entre los tres enfoques.

[26] Se establece que en el análisis de información de un estudiante universitario no solamente pueden ser útiles los datos generados por la institución como promedios o asistencia, sino que también denota la importancia de la información generada por el mismo estudiante en su tiempo libre como pueden ser búsquedas académicas en Google, Youtube y demás herramientas virtuales que pueden ser usadas actualmente como soporte del ambiente académico. Dicho esto, se plantean distintas herramientas, métodos y enfoques a la hora de realizar el análisis de la información con miras a la mejora educativa y al buen rendimiento de los estudiantes.

Como se menciona en [27], el soporte que pueda ser brindado por los interesados en un proyecto de SIN a nivel educativo, es fundamental tanto en la fase recolección de información como el análisis de la misma, dado que si el apoyo al proyecto es escaso se puede llegar a punto de que los resultados no sean los esperados por ambas partes. De esta manera en [27] lograron diseñar una propuesta de gobernanza de SIN ajustada a las necesidades de la universidad en el que se realizó el estudio, ofreciendo una comparación con sistemas similares de otras empresas; comparación a través de niveles de eficiencia, capacidad de gestión y medición.

En [28] gracias a la propuesta de un modelo de SIN en las universidades iraníes y su posterior prototipo se puedo realizar una monitorización del estado de la educación superior en dicho país además de realizar una comparación propia con países aledaños de Oriente medio dando información clara y oportuna acerca de

estudiantes y egresados del sistema educativo iraní.

Como lo describe en [29] se implemente un SIN en la universidad de Cundinamarca con fin de dar solución a una problemática creciente de deserción dentro del plantel más específicamente en la carrera de ingeniería de sistemas, suministrando información útil para el planteamiento de estrategias. Todo esto aplicando herramientas propias de este tipo de tecnología como lo son Microsoft *Power* BI en el procesamiento BI y Microsoft Excel, Microsoft Access, Sophos SSL VPN Client, Cliente Oracle (ODAC 12.2c versión 1 (12.1.0.0.2)) para el manejo, acceso, y creación de las fuentes de datos. Culminado el proceso se generó un tablero de mando (*Dashboard*) en el cual se visualizaba información relevante acerca de la zona geográfica de residencia, semestre con mayor índice de deserción. Adicional a lo anterior se pudo aclarar preguntas tales como si el género y la edad era un factor determinante en la deserción, cuál era el porcentaje de deserción precoz entre otras.

En [30] se propone desarrollar e implementar herramientas de SIN y minería de datos para generar estadísticas e informes periódicos y específicos que apoye el proceso de toma de decisiones y búsqueda de patrones en la deserción de la universidad San Gil en Santander, Colombia. Utilizaron como fuente de datos el SNIES de dicha institución, obteniendo datos personales, familiares, académicos y financieros del estudiante, datos del docente, asistencia, entrevista psicológica y datos estadísticos con los cuales se construyó una DW. Posteriormente se realizó el proceso de minería de datos con técnicas predictivas como Bayes Net y el árbol J48, obteniendo beneficios más favorables al aplicar este último, encontrando una herramienta totalmente eficiente y recomendable en la predicción de los perfiles estudiantiles más propensos a desertar.

En [31] se acude a la minería de datos empleando la técnica árboles de decisión para determinar los aspectos más influyentes en la deserción estudiantil de la universidad pública universidad de Nariño y la institución universitaria CESMAG que manejan un índice de deserción de 49% y 56%, dentro de los resultados hallados cabe señalar que el 100% de los estudiantes que desertaron son solteros y perdieron al menos una materia en el primer semestre.

Algunas universidades han desarrollado un sistema de alertas con el fin de hacer un seguimiento a sus estudiantes que presentan comportamientos asociados a la deserción.[1] crea un sistema de alarmas tempranas con el fin de crear medidas de intervención, estas alertas fueron basadas según variables proporcionadas Ministerio De Educación Nacional, quien en su estudio [32] asegura que el factor principal de deserción es el académico, seguido de los

factores socio-económicos e institucionales. De igual manera [2] desarrolla un sistema de alertas llamada SAT como innovación para el apoyo de la permanencia estudiantil, con el objetivo de detectar de manera temprana los factores de riesgo, permitiendo hacer su respectivo acompañamiento a cada estudiante.

Cabe señalar que [33] aunque no hace uso de SIN, hace un aporte importante para la permanencia académica, puesto que aquí se realiza un estudio observacional para determinar la frecuencia de deserción, este estudio afirma que la deserción se reconoce como multifactorial así también lo afirman [34] y [35] quienes también afirman que las causas que conllevan a un estudiante a no permanecer en la universidad se debe a un conjunto de factores, como académicos, económicos entre otros. En el estudio [35] realizan una investigación en una universidad privada de ecuador mediante revisión sistemática de artículos, entrevistas a estudiantes desertores, docentes y grupos focales, logrando identificar los factores que llevaron a los estudiantes a abandonar sus estudios. En [34] determinaron 5 factores principales que conllevan a la no permanencia de un estudiante dentro de sus estudios profesionales (falta de asesorías, inadecuado ambiente estudiantil, falta de seguimiento académico, deficiente calidad educativa y al servicio en general) mediante técnicas de minería de datos, proporcionando a al área de permanencia un soporte para la toma de decisiones.

1.3. Aportes investigativos

- Implementando el mapeo sistemático acompañado de la realización de encuestas a personas con experiencia en el tema de deserción, se logra tener un listado de variables que fueron categorizadas en distintos factores, estas variables son las que cobran mayor peso al momento de que un estudiante de educación superior deserte de sus estudios académicos.

- La universidad Unicomfacauca, gracias a este trabajo, cuenta con un modelado de datos y DM (*DataMart*) donde se encuentran almacenados los datos históricos de las distintas variables causantes de deserción, estos datos están almacenados de tal manera que se puede llevar a cabo un proceso analítico de cualquier índole de manera fácil, sin mayor dificultad.

- Se deja a la universidad corporación universitaria Unicomfacauca un *Dashboard BI,* se trata de una herramienta diseñada de forma metodológica y con base en diseño centrado en usuario para facilitar la toma decisiones y obtener información de los datos, el cual hace un seguimiento a los principales indicadores de deserción como (rango de edad de mayor

deserción, las carreras con más deserción, entre otros), proporcionando así reportes (gráficas con estadistas) fáciles de interpretar y que dan soporte para la toma de decisiones.

1.4. Publicaciones

- Artículo el cual se pretende indexar en una revista de categoría B que lleva por nombre Investigación e Innovación en Ingenierías, esta es una revista electrónica, editada semanalmente por el sello editorial de la Universidad Simón Bolívar, el objetivo de esta revista es publicar artículos relevantes y que generen conocimiento sobre aspectos teóricos o prácticos de las metodologías y métodos usados en los campos de la ingeniería.

- Artículo que se va publicar en la revista Investigación e Innovación en Ingenierías, lleva por título “MODELADO DE DATOS PARA EL ANÁLISIS DE LA DESERCIÓN ESTUDIANTIL” que describe el proceso de registrar el diseño de sistemas de software complejos a diagramas fáciles de entender, este artículo comprende el modelo conceptual, el modelo lógico y el modelo físico.

- Ponencia en el encuentro Mesar sur de investigación ACIET (Asociación colombiana de instituciones de educación superior), en este encuentro se socializó el proyecto “implementación de un sistema de inteligencia de negocios aplicada al área de permanencia de la corporación universitaria Comfacauca – Unicomfacauca”.

- Ponencia del proyecto “implementación de un sistema de inteligencia de negocio en el área de permanencia académica de la Corporación Universitaria Comfacauca – Unicomfacauca” fue expuesto en el XV encuentro departamental de semilleros de investigación Redcolsi 2021, tras haber realizado una excelente ponencia, el proyecto recibe una puntuación de 92 punto de 100 posibles, ganando así la oportunidad de participar en el encuentro nacional de proyectos de investigación ENSI 2021.

- Ponencia como invitado para el encuentro nacional de proyectos de investigación ENSI 2021 tras haber clasificado en Redcolsi, donde fue evaluado como un trabajo excelente y alta calidad.

1.5. Contenido

Capítulo 2: MODELADO DE DATOS

En este capítulo se describen todos los procesos utilizados para hacer un análisis más profundo en este contexto; se llevaron a cabo encuestas a la comunidad universitaria y mapeos sistemáticos de la literatura con el fin de identificar y clasificar los factores más relevantes que conducen al estudiante a desertar de sus estudios superiores. Adicionalmente se trabajó en la calidad de los datos suministrados por la Corporación mediante algoritmo CBR con distancia euclidiana, permitiendo de esta manera proceder con el modelado dimensional en el cual se estructuro un diagrama estrella las diferentes dimensiones y tabla de hechos con sus medidas pertinentes.

Capítulo 3: IMPLEMENTACIÓN DEL DATAWAREHOUSE - DW

En esta parte se procede al diseño dimensional y el desarrollo de la *Datamart*, para su posterior migración de datos, se especifican las herramientas usadas y su método de selección, como última instancia se presentan conclusiones.

Capítulo 4: ARQUITECTURA DEL SISTEMA

Resumen: aquí se explica el diseño y la arquitectura (en la parte del hardware como también del software) para el sistema propuesto, basado en los requerimientos funcionales, adicionalmente se presenta una evaluación que define si el sistema implementado cumple con las expectativas de los usuarios finales

Capítulo 5: CONCLUSIONES Y TRABAJOS FUTUROS

En este último capítulo del presente trabajo se expone las conclusiones que surgen a raíz de la elaboración de los puntos descritos anteriormente, adicional a esto también se da a conocer posibles trabajos que complementen el presente y/o abarquen otros contextos.

Capítulo 2: MODELADO DE DATOS

2.1. Introducción

En este capítulo se realiza un análisis de la deserción empleando el mapeo sistemático que es un método usado en el campo de la investigación para hacer un análisis más profundo de un determinado tema, en este caso la deserción, además de esto, se realiza una entrevista a personas relacionadas con la educación superior sobre los factores que causan la deserción estudiantil.

Los métodos descritos anteriormente fueron usados porque es estrictamente

necesario conocer las variables más influyentes en la deserción para poder extraerlos de manera correcta y posteriormente realizar un análisis de datos usando las técnicas de minería de datos, debe quedar claro que al identificar las variables de diferentes factores se puede tener una visión global de la problemática.

Este capítulo describe un marco para realizar un modelado de datos basado en factores de deserción, basado en la metodología CRISP DM, este capítulo está estructurado de la siguiente manera: la siguiente sección ilustra el estado del arte. la sección 3 expone resultados arrojados por encuestas realizadas a personas conocedoras del tema. En la sección 4 se describe la caracterización. En la sección 5 se describe las tecnologías usadas. En la sección 6 se presentan las discusiones y finalmente las conclusiones.

2.2. Características propenso a desertar

Se realiza un mapeo sistemático con el fin de priorizar los datos que aporta o contribuye a la deserción estudiantil en instituciones de educación superior.
Para ello se usa la metodología propuesta en el documento [36], quienes utilizan mapeo sistemáticos para responder preguntas de investigación basado en consultas a la literatura científica.

dentro de la literatura revisada para la realización del mapeo sistemático se encuentran múltiple trabajos que usan un SIN para abordar el tema de deserción, como lo hacen [37] [38][39]quienes mediante la minería de datos intentan identificar los estudiantes con alto riesgo de deserción, de igual manera [40] hace uso de los SIN para hacer un análisis del rendimiento académico de los estudiantes para identificar los desertores y posteriormente analizar desde distinta perspectivas(estado civil, estrato, etc.) . por otra parte [41][42][43] mediante la minería de datos hace un análisis sobre información académica para mirar aquellos factores que influyen en la deserción.

[44] expresa en su trabajo, haber implementado un SIN para integrar diferentes fuentes de información, para su posterior análisis, identificando patrones influyentes para construir un modelo de predicción, adicional a esto, [42] tras usar la minería de datos para el análisis de la deserción, sus resultados arrojan que los factores más influyentes son lo académico, familiares y económicos. Por otra parte, cabe señalar que [33] aunque no hace uso de SIN, hace un aporte importante para la permanencia académica, puesto que aquí se hace un estudio observacional para determinar la frecuencia de deserción, este estudio afirma que la deserción se reconoce como multifactorial.

Como lo dice [34], [35] la deserción estudiantil universitaria no depende únicamente de un único factor, sino que parte de la interacción y conjunción de varios factores tanto económicos, académicos entre otros. [34] determinaron 5 factores principales de un total de 27 mediante técnicas de minería de datos educativos tales como la falta de asesorías, inadecuado ambiente estudiantil, falta de seguimiento académico, deficiente calidad educativa y al servicio en general. Adicional a lo anterior, usaron técnicas de minería de datos como los árboles de decisión C45 encontrando patrones determinantes en la decisión de desertar de un estudiante, pero con variabilidad en contextos similares como la región, lugar de procedencia, nivel socio económico.

En [45] se realizó una investigación acerca de los posibles motivos que tiene los estudiantes en reprobar y desertar de los distintos cursos relacionados con matemáticas, por lo cual se recopilaron datos mediante una metodología de diseño mixto cuantitativo-cualitativo, haciendo uso de cuestionarios, entrevistas y un grupo focal de estudiantes del curso de matemáticas generales. A partir de los datos recopilados pudieron señalar que variables como sexo, área de procedencia, estrato, horario del curso matriculado, carrera que cursa entre otras son esenciales a la hora de identificar al estudiantado con riesgo de deserción.

[45] puntualizó que factores como Bajo rendimiento académico en el primer examen parcial del curso, Priorización hacia otro curso matriculado, Deficiencia en conocimientos previos, Falta de interés por el estudio, Poca dedicación al estudio del curso, son los más motivos con más incidencia en la decisión de abandonar sus estudios por parte de los estudiantes.

[35] Realizan una investigación en una universidad privada de ecuador mediante revisión sistemática de artículos, entrevistas a estudiantes desertores, docentes y grupos focales, logrando identificar los factores que llevaron a los estudiantes a abandonar sus estudios. Los resultados arrojaron que los factores predominantes en esta universidad privada fueron el factor económico, integración social, metodología docente, razones vocacionales, motivaciones y estrategias empleadas por los estudiantes para el estudio. Como se plantea en [46] la deserción se presenta con mayor frecuencia principalmente en los primeros años del programa universitario, por tanto, se ha hecho relevante un estudio profundo que permita un análisis cuantitativo y cualitativo de las causas de la deserción. Con lo cual se ha podido clasificar las causas en 7 macro factores como: Factor familiar, Factor individual, factor sector educativo, factor sector IES, factor económico, factor social y factor cultural

En [47] se implementó una inteligencia de negocio en la cual se revisó la literatura

tanto internacional como nacional para definir los atributos de los perfiles con riesgo de deserción observando el valor de las variables que componen estos factores, así definir empíricamente el tipo de atributo de cada perfil.

Gracias a la investigación de [48] mediante los resultados de minería de datos implementada en la facultad de ingeniería de sistemas de la universidad de Cundinamarca se puede estimar que las variables con más probabilidad de presentarse en los distintos tipos de deserción son el género, la edad, programa académico, materia, empleo, ubicación geográfica, entre otras, dando un patrón entre toda la literatura consultada.

[30] se utilizaron los datos de las bases de datos de UNISANGIL para poder realizar minería de datos a sus estudiantes con el fin de determinar los estudiantes más propensos a abandonar sus estudios, usaron información personal y familiar como son edad, género, lugar de procedencia, estrato socioeconómico, origen étnico, estado civil, entre otros. Un segundo conjunto de datos de la parte académica y financiera (colegio, puntaje ICFES, becas, créditos), obteniendo resultados alentadores con respecto a la utilización de estas técnicas en el sector de permanencia académica.

De igual forma en [11] se identificaron factores de riesgo que afecta al estudiante al momento de decidir dejar sus estudios, por lo tanto, los clasificaron en Factores académicos (horario, promedio, carrera, semestre), factores demográficos (edad, procedencia, sexo), factores socioeconómicos (nivel estudios padres, ingreso padres, convivencia), todo esto con el fin de filtrar la gran cantidad de datos obtenidos de la universidad.

[49] se realizó un análisis de datos de estudiantes ingresados a la facultad de ingeniería electrónica y computación de la universidad de Guadalajara tomando como intervalo de tiempo el año 2008 hasta el año 2016, todo lo anterior con el fin de determinar los factores que provocan la deserción escolar. Pudieron concluir que el nivel de deserción se veía más intensificado en los primeros 4 semestres de la carrera. Mediante entrevista a diferentes estudiantes desertores de esta ingeniería se identificaron que los principales motivos que los llevó a tomar esta decisión fueron principalmente dificultades en cursos de matemáticas, falta de adaptación al modelo académico, desagrado por la carrera y factores económicos como necesidad de trabajar

Realizó un estudio de corte transversal con el objetivo de construir un perfil de riesgo al abandono estudiantil, el cual constó 1897 estudiantes con edad promedio de 22.6 los cuales 38.8 eran hombres y 61.2 eran mujeres. La información se recolectó mediante cuestionarios a través de 4 segmentos

divididos en socioeconómico, familiares, protectores (deporte, cultura, ocio) y vivencias académicas. Considerando la prueba chi-cuadrado para establecer la hipótesis de independencia entre las variables, en la etapa inicial se tomó como principales características: edad, sexo, estrato, estado civil, estado laboral, promedio académico, carrera, adaptación, nivel educativo padres entre otras más. Dado lo anterior, se puede concluir que el nivel de adaptación jugó un lugar importante en la permanencia académica, ya que era directamente proporcionales a otros factores importantes como el funcionamiento familiar, buena elección vocacional y factores económicos.

Para lo cual se siguieron los siguientes pasos:

1. Preguntas claves
2. Palabras claves
3. Búsqueda avanzada en bases de datos indexadas como Google Académico, *Scopus* y el motor de búsqueda Mendeley.
4. Selección de artículos

Con ello se concluye que en la deserción estudiantil influyen 4 factores fundamentales como se muestran en cada columna de la tabla 1: individuales, socioeconómicos, académicos e institucionales, estos factores cuentan con diferentes variables que han sido discriminadas en cada fila.

Tabla 1:Clasificación de los factores principales causantes de la deserción

INDIVIDUAL	ACADÉMICO	SOCIOECONÓMICO	INSTITUCIONAL
Género	Nota	Estrato económico	Asistencia a asesoría
Edad	Horario	Número personas grupo familiar	Modalidad de ingreso
Lugar de residencia	Semestre cursado	Número de aportantes	Listado docente por materia
Estado civil	Promedio académico	Ingreso familiar	Cantidad de estudiantes por carrera
Procedencia	Inasistencias	Nivel educativo padres	Cantidad desertores por carrera
Salud	Programa académico	Ocupación de padres	Adición créditos
Discapacidad	Colegio de procedencia	Rango de ingresos recibidos	Formación docente
Empleo	Materias a cursar	Vocación	Titulación
	Materias cursadas	Ingreso económico	

2.3. Modelado de datos, Análisis con expertos

En este paso se procede a realizar una encuesta para constatar las variables más

incidentes en la deserción estudiantil según estudiantes y demás personas vinculadas en la educación superior que de una manera u otra han evidenciado casos de deserción estudiantil. Dentro de los encuestados el 85% pertenece a la Universidad Unicomfacauca como se muestra en la figura 1:

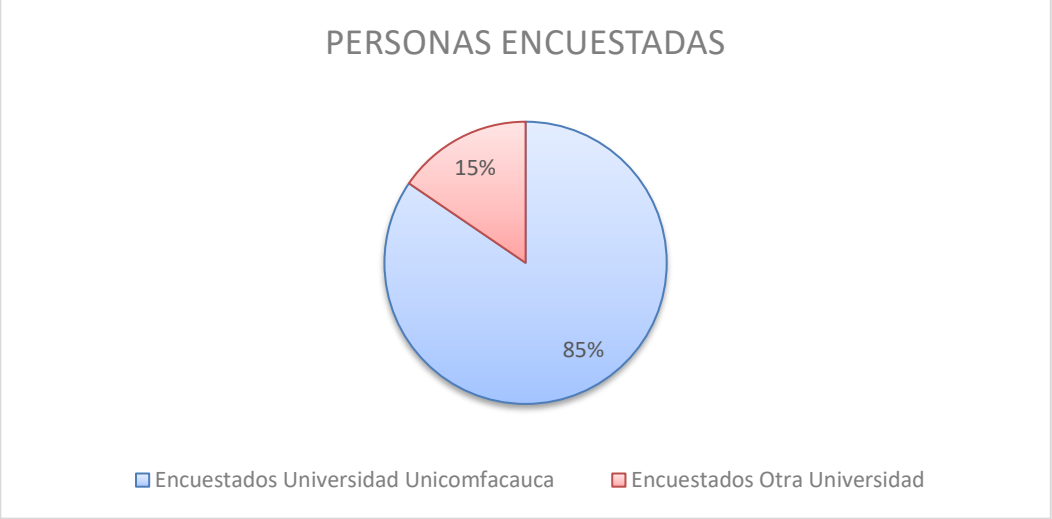

Figura 1: Personas encuestadas

La figura 2, muestra el conjunto de personas encuestadas de otras universidades se presentan las siguientes estadísticas e instituciones:

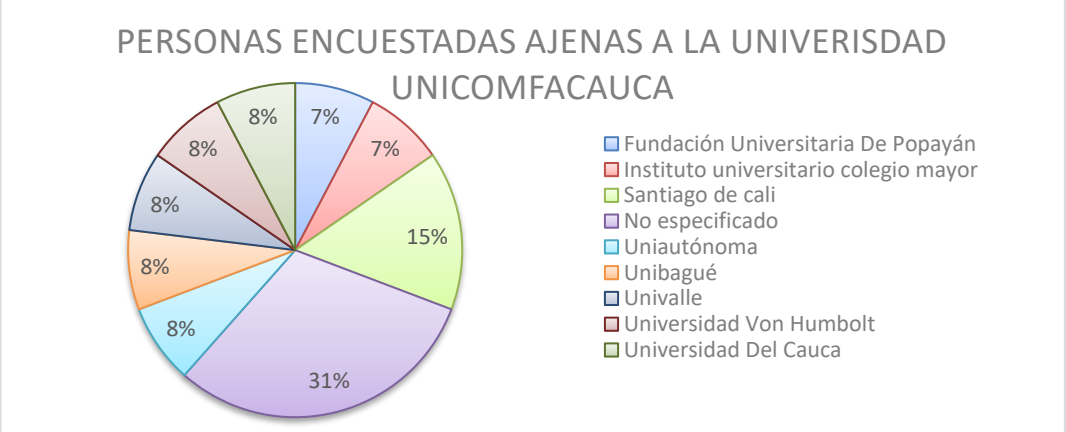

Figura 2: Personas encuestadas por institución

La Figura 3 mostrada a continuación permite apreciar el porcentaje de personas encuestadas según su edad.

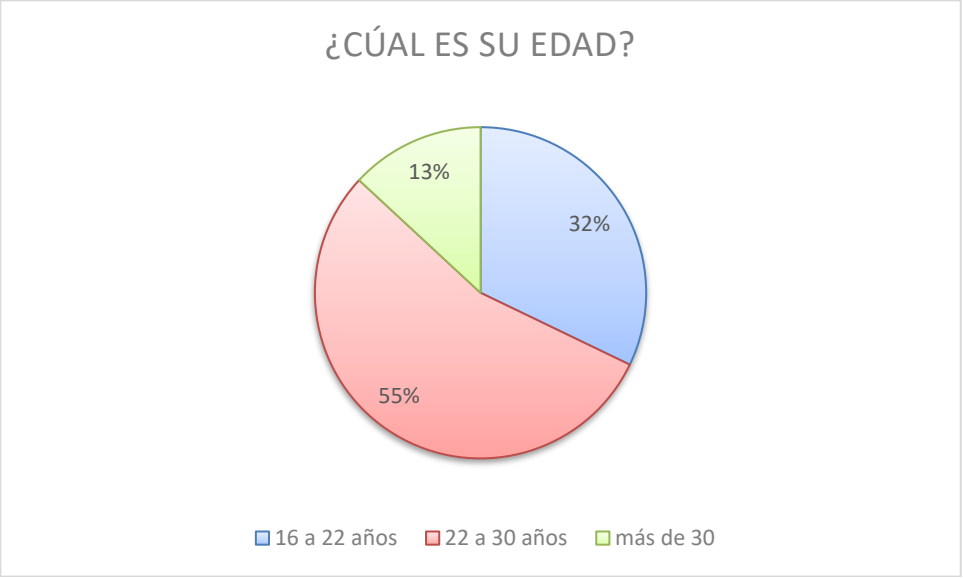

Figura 3: Muestra poblacional según rango de edad

La Figura 4 que se muestra a continuación refleja el porcentaje del sexo de persona que contestaron la encuesta.

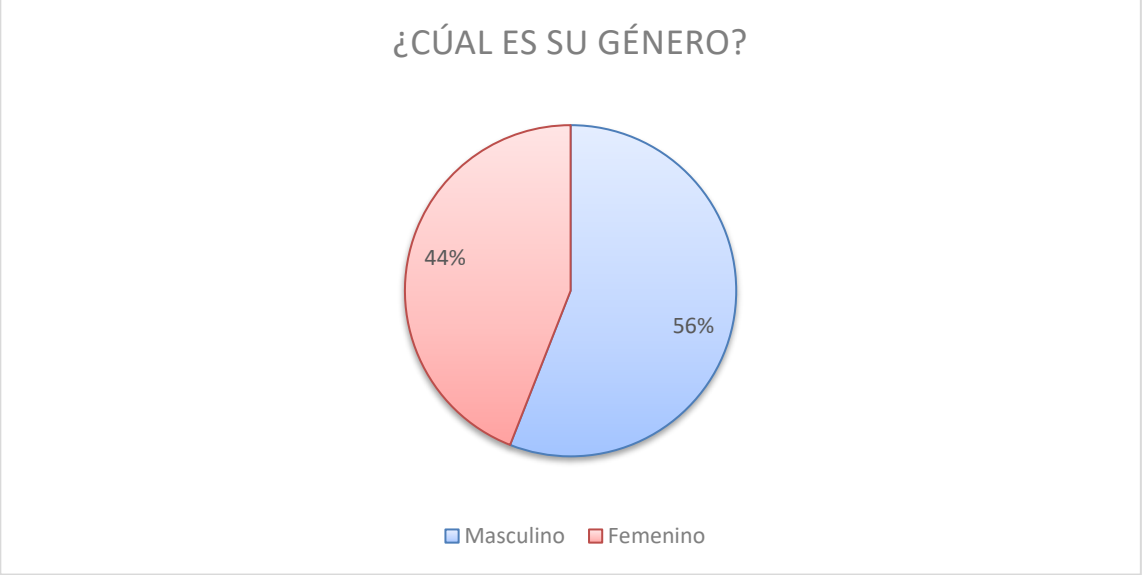

Figura 4: Muestra poblacional según género

Los encuestados arrojan que el 84,5% conoce a por lo menos una persona desertora de sus estudios superiores, por consiguiente, estas personas pueden identificar aquellas variables que incidieron en que estos individuos se desvíen

del camino hacia la finalización de sus estudios superiores. Lo descrito anteriormente se evidencia en la siguiente figura 5.

Figura 5: Muestra poblacional relacionada con deserción

En la Figura 6, como último ítem de esta encuesta investigativa se tienen las variables que según estudiantes, egresados y docentes son las más significativas dentro de este contexto de la deserción académica.

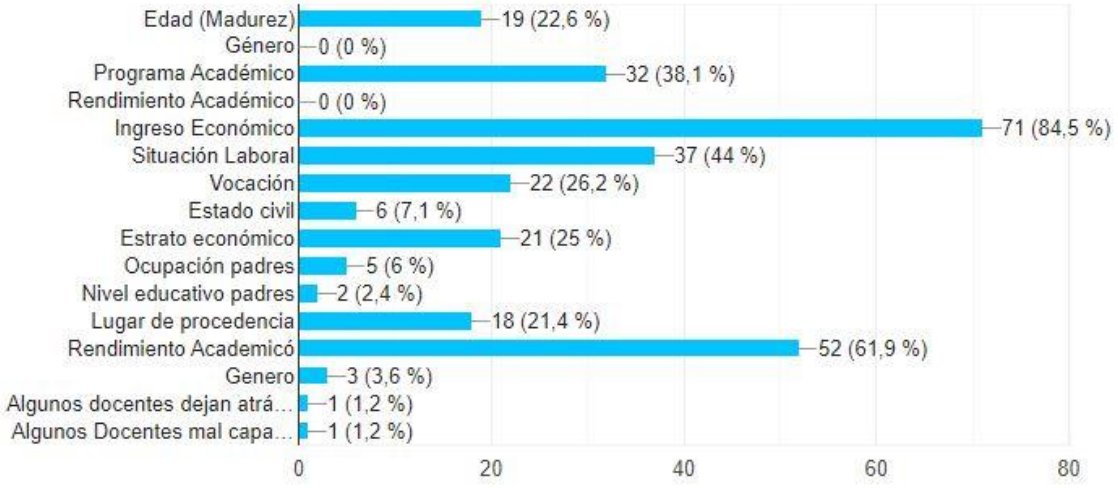

Figura 6: Principales factores causante de deserción según muestra poblacional encuestada

Con la estadística proporcionada por la gráfica anterior se puede concluir que, el

rendimiento académico al igual que el factor económico son variables que cobran mayor relevancia en la NO culminación de estudios superiores, corroborando los resultados expuestos en el mapeo sistemático quien también señala estos factores como los más significativos.

Adicional a esto se realizó una investigación de la literatura publicada en relación a los factores con mayor influencia en la decisión de un estudiante a abandonar sus estudios universitarios, tomando el año de publicación de distintos artículos en los que el factor se menciona con mayor frecuencia se graficó la Figura 7 con una muestra de 20 artículos.

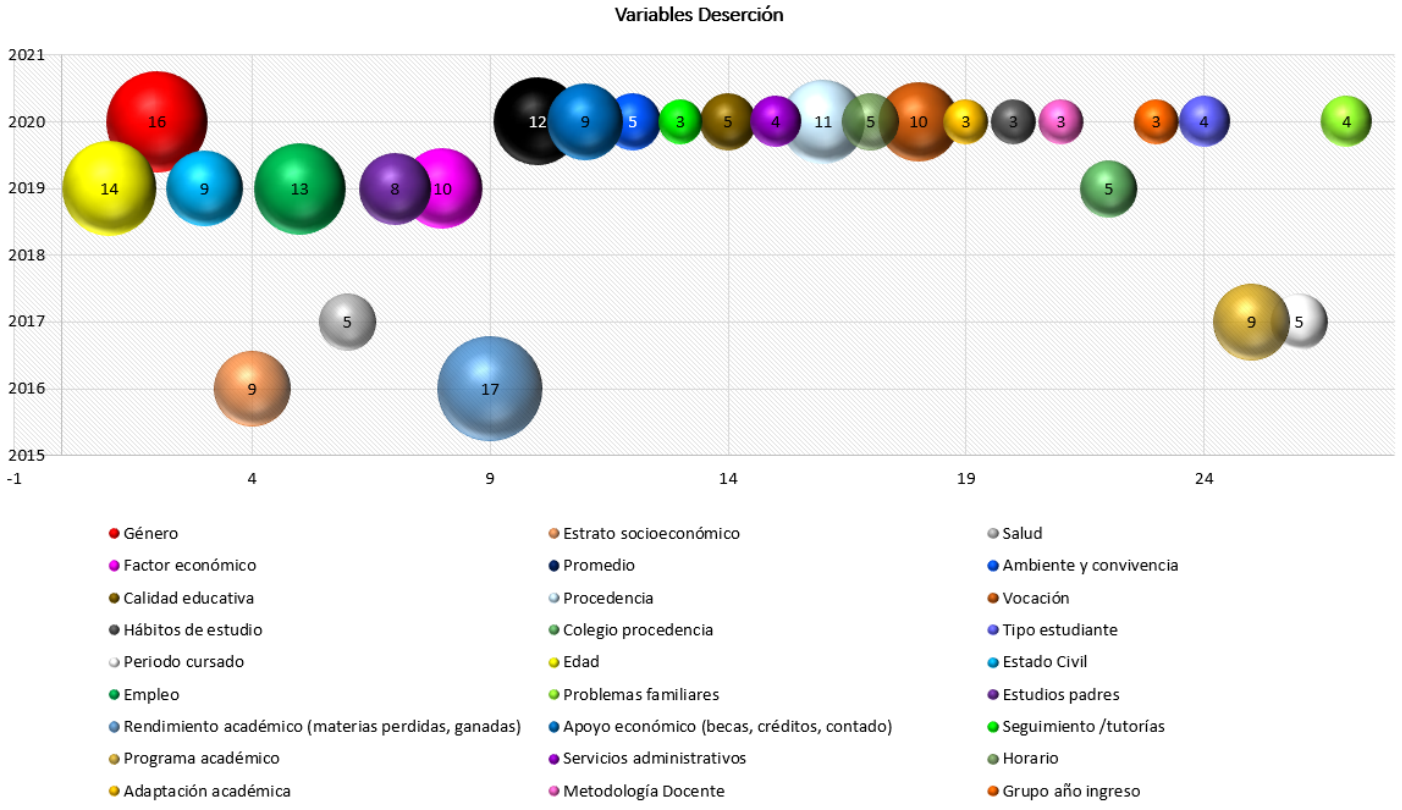

Figura 7: Factor de más frecuencia por año del articulo

De la cual se puede dilucidar cuales son los factores que pueden ser más útiles a la hora de procesar y analizar los datos en el SIN propuesto, por medio de los indicadores que presenta la gráfica, indicando el tamaño de la burbuja la frecuencia con que es usado el factor en cuestión, el año nos indica si es más usado en artículos recientes o ha decaído su uso en este tipo de investigaciones.

En la Figura 8 se muestran los distintos factores en una comparación entre años, como ejemplo se toma el factor edad, en la cual se puede concluir que ha sido indicado 4 veces en artículos del año 2019, mientras que en el año 2016 fue indicado un número de 3 veces. De esta forma se puede analizar con mayor facilidad el registro histórico de cada factor dentro de la literatura consultada.

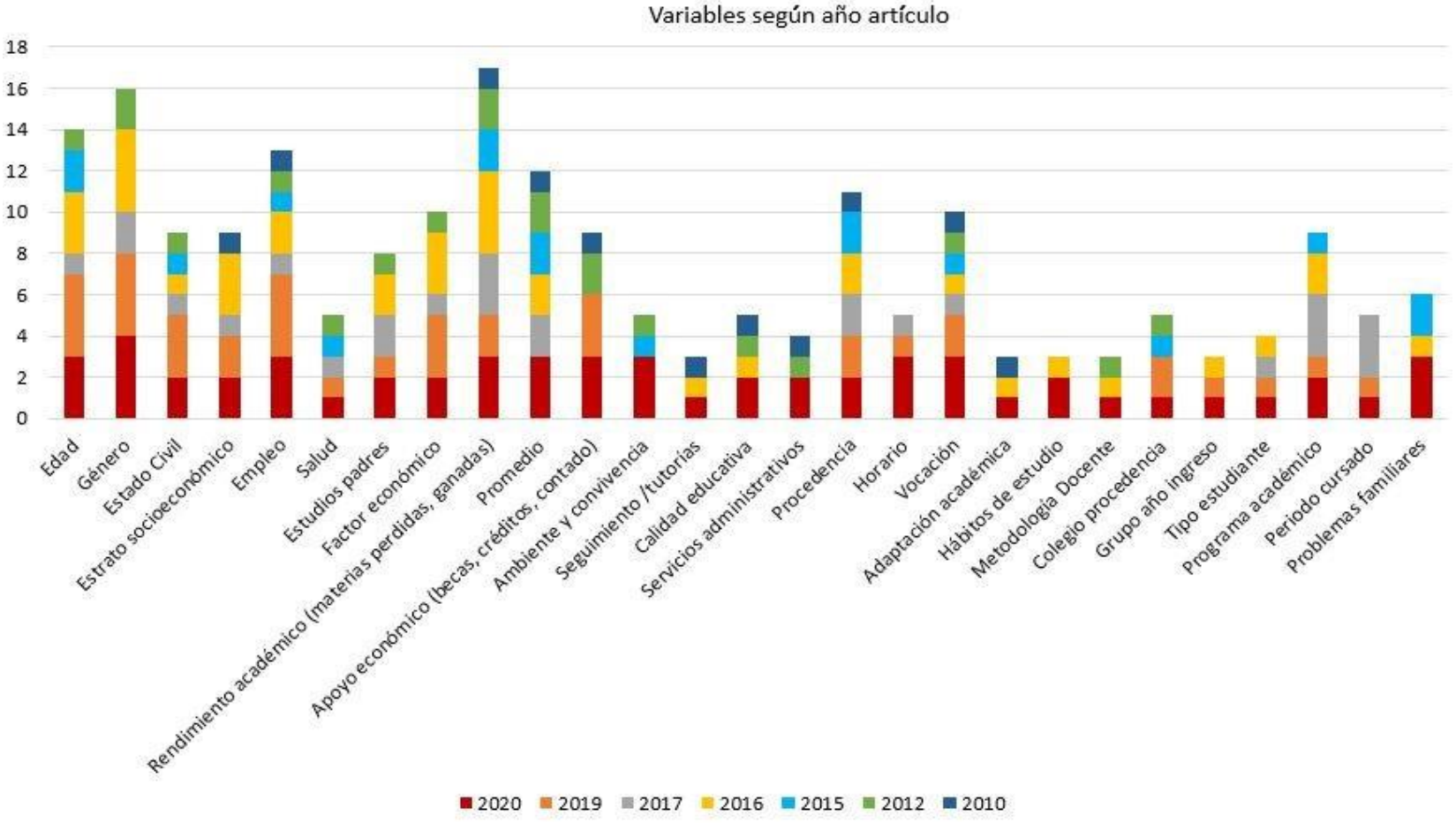

Figura 8: Factores según años de los artículos

La variable que más se vio reflejada su importancia en el sistema propuesto y dentro de lo investigado sobre los factores influyentes en la deserción fue el *rendimiento académico*, seguido del *estrato socioeconómico* del estudiante. Esto, debido a que el rendimiento académico presenta la mayor dificultad para un estudiante al momento de cursar su carrera y es el principal motivo de deserción universitaria. Como segunda variable el estrato socioeconómico nos indica indirectamente la solvencia económica y capacidad de endeudamiento de un estudiante, hecho fundamental a tomar en cuenta al ser la educación superior una inversión costosa monetariamente, independiente del plantel universitario donde se estudie. Cabe resaltar lo que se ha comentado a lo largo del documento, estas variables por si solas no son suficientes para identificar o perfilar correctamente estudiantes con riesgo a desertar, por tanto, es importante tener en cuenta un amplio conjunto de variables como el expuesto en este proyecto, el cual fue consultado de literatura referente a la deserción tanto en universidad públicas como privadas con el fin de contar con antecedentes imparciales.

Los requerimientos del usuario fueron tomados a través de una entrevista con el encargado del área de permanencia y con el coordinador del área de sistema de información, esta entrevista se realiza en la Corporación Universitaria Comfacauca – Unicomfacauca sede Popayán, de aquí surgen los siguientes requerimientos descritos en la Tabla 2.

Tabla 2: Requerimientos de usuario

#	Descripción
001	¿Qué programa académico tiene mayor índice de deserción? Reporte alumnos que desertan por programa académico.
002	¿Qué facultad presenta mayor índice de deserción? Reporte de alumnos retirados por facultad
003	¿Qué sede presenta mayor índice de deserción? Reporte de alumnos retirados por sede
004	¿Qué materias presentan un mayor grado de dificultad para los estudiantes? Reporte alumnos reprobados por materia
005	¿Cuáles son aquellos docentes, que tienen un mayor número de estudiantes reprobados? Reporte alumnos reprobados por docente

006	¿Qué jornada de estudio presenta mayor porcentaje de deserción? Reporte cantidad de alumnos desertores por jornada de estudio (Diurna / Nocturna)
007	¿Cuáles son aquellos semestres en los cuales un estudiante está más propenso a desertar? Reporte alumnos desertores por semestre cursado
008	¿El género es un factor determinante en la deserción? Reporte alumnos desertores por género
009	¿Qué rango de edad son más vulnerables a desertar? Reporte alumnos desertores por edad
010	¿Existe un estado civil que afecte la permanencia de los estudiantes en la universidad? Reporte alumnos desertores por estado civil
011	¿Qué lugar de procedencia tiene mayor grado de deserción? Reporte cantidad de alumnos desertores por lugar de procedencia(departamento)
012	¿En qué medida afecta cada uno de los estratos económicos en la deserción? Reportes cantidad de alumnos desertores por estrato económico
013	¿Podrá el método de financiamiento afectar la permanencia del estudiante? Reporte de alumnos desertores según su método de financiamiento
014	¿De los tipos de modalidad de ingreso, existe alguno en específico que requiera mayor prioridad de atención debido a su diferencia en número de desertores presentes? Reporte alumnos desertores según la modalidad de ingreso (ComfaBecas, Becas e ingreso normal)
015	¿Los factores demográficos deben ser tenidos en cuenta al momento de combatir la deserción? Evidenciar el grado de impacto que tiene factores demográficos como la procedencia dentro de la deserción estudiantil

016	¿El sistema identifica aquellas variables que influyen en la culminación de estudios de los estudiantes? Determinar los factores beneficiosos que incidieron en la vida de los egresados de la institución

En la Tabla 3 se describen los requisitos del sistema, donde se expone lo que el Sistema de inteligencia de negocios debe hacer.

Tabla 3: Requerimientos del sistema

#	Descripción
001	¿En qué tipo de archivos se mostrarán los reportes de las consultas realizadas? Los reportes de consultas que se generarán por los usuarios podrán ser exportados a un archivo Excel y/o Pdf, permitiendo mostrar la información mediante diferentes tipos de gráficos: lineales, tortas y barras.
002	¿Qué tan bondadoso es el sistema que se pretende construir? El sistema debe brindar información de forma detallada, precisa, y de rápida respuesta, de tal modo, que el usuario final pueda obtener la información en pocos segundos.
003	¿El programa brinda seguridad ante personas no autorizadas para acceder a la información? Asegurar la confidencialidad de los datos, proporcionando una ventana donde el usuario debe ingresar sus credenciales para acceder al sistema
004	¿Cada cuánto se actualiza la base de datos? Se actualizará la información que alimenta el sistema cada corte según lo establecido por la corporación.
005	Generar informes y tableros de control muy amigables para informar a los usuarios

2.3.1. Matriz de viabilidad

Una vez entregado el conjunto de datos por parte del Coordinador del sistema de para el desarrollo de esta actividad, se procede a crear el matriz de viabilidad según los datos recibidos, aquí se tuvo en cuenta si los datos

pertenecientes a cada variable de deserción que se solicitó se recibieron de manera completa o si al menos cuentan con un 90% del total de registros. Véase la matriz de viabilidad en la Figura 9.

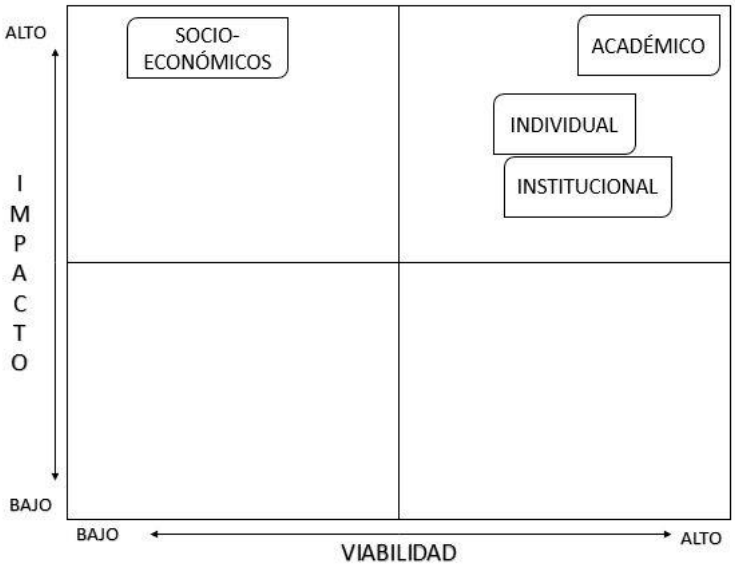

Figura 9: Matriz de viabilidad para los factores influyentes en la deserción universitaria

2.4. Caracterización mundo

Para la prueba de concepto del modelo se cuenta con tres fuentes de datos proporcionadas por la Corporación Universitaria Comfacauca - Unicomfacauca en formato CSV (Véase la Figura 10. Archivos CSV recibidos del área de sistemas de información de Unicomfacauca), en la Tabla 4 se detallan las diversas variables que contienen los datos.

Datos_Academicos_V1_BI.csv	35.076.938	3.321.083
Datos_Docente_V1_BI.csv	57.015	7.377
Datos_Horario_V1_BI.csv	3.107.884	323.055
Datos_Matricula_V1_BI.csv	2.887.854	220.973
Datos_Personal_Estudiante_V1_BI.csv	3.032.349	302.494
fecha.csv	19.193	15.472

Figura 10. Archivos CSV recibidos del área de sistemas de información de Unicomfacauca.

Cada archivo contiene información relevante de un aspecto especifico relacionada con el estudiante mediante llaves foráneas, por ejemplo, “Datos

académicos" contiene información sobre de las notas por cada materia, jornada, habilitación, homologación; "Datos docente" contiene información acerca del nivel de estudio del docente (pregrado, maestría, doctorado) junto con el nombre de la carrera universitaria estudiada; "Datos Horario" contiene información como el nombre de la materia, jornada, programa, sede, semestre, hora y día en que se imparte la materia; "Datos Matricula" contiene información acerca del tipo de la matricula, el estado de cancelación, periodo y entidad de crédito ; "Datos personal estudiante" contiene información personal del estudiante como género, estrato, periodo de ingreso y estado académico actual ; "fecha" contiene los diferentes periodos y sus respectivas fechas.

De lo anterior, estas fuentes se dividen en tres categorías de datos principales, personales, administrativos y académicos, cada una de ellas con un número total de registros de 353802, 899, 61282,14590, 27125 y 38 respectivamente. Para realizar el ETL, se procede a convertir los archivos en formato CSV a XLSX para facilitar la creación de la sabana o *data staging,* la cual almacena los datos sin procesar tal cual se recibieron para realizar su posterior tratamiento (véase Figura 11).

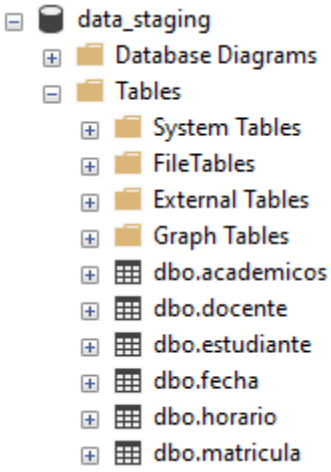

Figura 11. Sabana o *data staging* formado a partir de los datos suministrados

Una vez con los datos cargados en la sabana se hacen diferentes consultas de los datos que se requieren para el modelo dimensional propuesto (Ver **¡Error! No se encuentra el origen de la referencia.**), una vez se cuenta con la serie de datos, se categorizan según el modelado de datos de este articulo y se procede a hacer el proceso de transformación el cual consiste en realizar limpieza, depuración y des duplicación.

Revisando el data set se encontró que cuentan con inconsistencias como datos faltantes o nulos en diferentes registros y campos (véase Figura 15), mala praxis de almacenamiento, como, por ejemplo, diferentes nombres para un mismo programa(véase **¡Error! No se encuentra el origen de la referencia.**, campos con mala ortografía, errores de digitación y registros con dudosa veracidad(véase Figura 14).

nota1	nota2	nota3	nota_final
4.0	4.8	4.4	44
3	2.8	3.4	31
4.3	3.9	3.5	39
3.1	3.1	2	27
3.3	3.7	3.8	36
4.5	4	4.7	44
3.3	4.0	3.6	36
2.8	3.8	2.9	31
3.5	2.9	1.3	24
2.4	4.0	4.2	36
4.3	4.5	4.5	44
4.3	4.1	3.9	41
4.2	3.3	3.5	37

Figura 12. Inconsistencias presentadas en el *dataset* recibido

D	E
descripcion	materia
DIURNA	CONTEXTO POLÍTICO, ECONÓMICO Y SOCIAL
DIURNA	CALCULO
DIURNA	CALCULO
DIURNA	CALCULO
DIURNA	CALCULO
DIURNA	CALCULO
DIURNA	CALCULO
DIURNA	CALCULO
DIURNA	CALCULO
DIURNA	CALCULO
DIURNA	CALCULO
DIURNA	CÁLCULO DIFERENCIAL
DIURNA	CÁLCULO DIFERENCIAL
DIURNA	CÁLCULO DIFERENCIAL
DIURNA	CÁLCULO DIFERENCIAL

Figura 13. Galimatías presentados en el *dataset* recibido

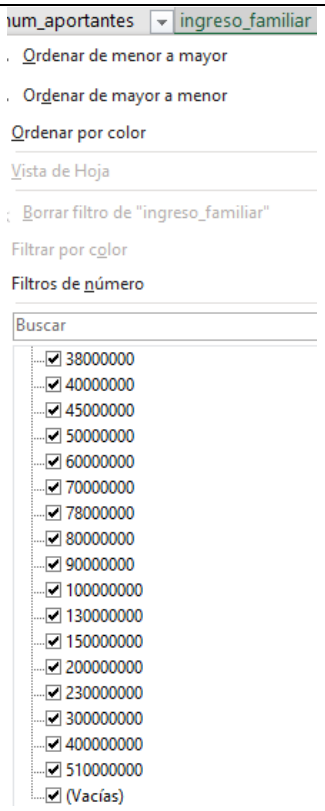

Figura 14. Datos con poca veracidad dentro del dataset

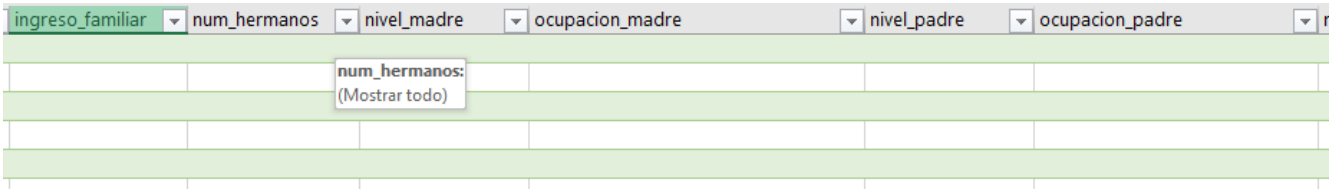

Figura 15. Datos faltantes dentro del data set

Como algunos de los registros de las diferentes fuentes aparecen nulos, se procede a completar la información haciendo uso de una técnica de minería de datos CBR (Razonamiento basado en casos, traducido del inglés), en la cual se toma como referencia aquellos registros que cuentan con información, de esta manera se compara mediante distancia euclidiana con aquellos registros que si presentan datos faltantes. Es importante aclarar que solo se comparan los campos llenos que están relacionados con el dato vacío que se quiere completar, como, por ejemplo, si se desea completar el estado civil en un registro es más preciso realizar la distancia euclidiana entre las edades del registro vacío y el registro lleno debido a que la edad está estrechamente relacionada al estado civil de una persona. Todo esto pensado con el fin de que asegurar que el *Data Mart*

tenga una mejor calidad de datos y aumente la precisión en la recomendación estadística a visualizar en el *dashboard* por el usuario. Cabe mencionar que la forma de relacionar las fuentes es con el número de identificación y código estudiantil.

Tabla 4. Fuentes de datos

Personal	Administrativos	Académico			
Número Registros: 14590	**Número Registros: 61282**	**Número Registros: 353802**	**Número Registros: 27125**	**Número Registros: 899**	**Número Registros: 38**
estado_civil	idperiodo	codigo_materia	codigo_materia Horario	codigo_docente	idperiodo
estrato	numcreditos	nota1	idperiodo	titulo	descripcion
categoria_comfa	adicion_credito	nota2	codigo_docente	nivel_academico	año
fecha_nacimiento	Deuda	nota3	descripcion	nivel_estudio	semestre
departamento_procedencia		nota_final	materia		
municipio_procedencia		homologa	programa		
departamento_reside		semestre	sede		
municipio_reside		jornada	semestre		
zona_reside		inasistencia	inicio		
personas_a_cargo		habilita	final		
estado_academico		estado	diaa		
primer_periodo		adicion_credito			
ultimo_periodo		horario			
num_personas_grupo_familiar					
num_aportantes					
ingreso_familiar					
num_hermanos					
nivel_madre					
ocupacion_madre					
nivel_padre					
ocupacion_padre					
nivel_conyugue					
ocupacion_co					

nyugue					
Discapacidad					
Fechaingreso					
fecha_grado					
Colegio					
tipo_colegio					
Tipo_rol					

2.4.1.CBR_1 Complementar datos Nulos.

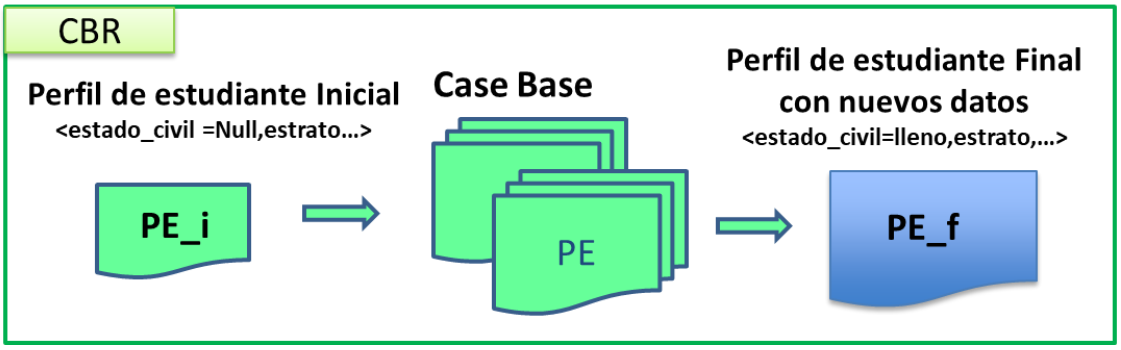

Figura 16. Proceso CBR

Este módulo tiene una base de casos que son los PE_i (Perfiles iniciales) registros totales las fuentes de datos (ver Tabla 4), cada caso es descrito por los registros de perfil de estudiante PE_i (fuente Personal). Este perfil inicial tiene registros nulos los cuales serán completados con el CBR_1, que estima el valor del registro más apropiado que complete el campo nulo de los perfiles de usuario.

Este módulo tiene una base de casos igual al número de registros de la fuente de datos (atributos), donde cada caso es descrito (PE) por varias categorías y 46 atributos, se tiene un perfil de usuario inicial PE_i que tiene al menos un registro nulo. De esta forma, la solución será un nuevo perfil de usuario PE_f, al cual se le complementará con el registro nuevo. El algoritmo compara localmente el valor de similitud para cada atributo del perfil de usuario (consulte la tabla 4) utilizando el voto por mayoría (*majority voting*) para seleccionar el registro más adecuado y así llenar ese campo nulo y dar una solución.

$$RegistroNull =< PE_i, PE > (1)$$

Ecuación 1:

$$PE_f = < EstadoCivil, Estrato, CategoriaComfa, \dots etc. > (2)$$

Ecuación 2:

Para evaluar el algoritmo, se calcula la similitud entre los atributos de PE_f y PE_q , usando una función binaria [0,1]. Se compara si los atributos del perfil recuperado PE_f son igual a los atributos del caso de prueba PE_q. Entonces se calculó el valor α (ver Ecuación 3)

$$a = sim(EstadoCivilQ, EstadoCivilR)\ \epsilon\ [0,1]$$

Ecuación 3:

$$b = sim(estratoQ, estratoR)\ \epsilon\ [0,1]$$

Ecuación 4:

$$c = sim(CategoriaComfaQ, CategoriaComfaR\ \in [0,1]$$

Ecuación 5:

$$AtributoN = sim(atributoQ, atributoR)\ \epsilon\ [0,1].$$

Ecuación 6:

El proceso de similitud entre variables se llevó a cabo mediante la fórmula de distancia euclidiana Ecuación 7, en la cual se toma como “A” el valor de referencia para la columna (A) del registro sin datos nulos sobre el cual se desea realizar la similitud y “a” el valor de referencia para la columna (a) del registro al que se desea completar el campo “X”.

$$\propto = \sqrt{((A-a)^2 + (B-b)^2 + (C-c)^2 + (D-d)^2 + (E-e)^2 + \cdots + (N-n)^2)}$$

Ecuación 7: Formula de la distancia Euclidiana

Este cálculo se realizó (x < 14590), siendo “X” el número de registros que no tienen nulos los datos que están relacionados con el campo el cual se desea completar, a su vez se realizó este proceso con cada una de los registros que presentara datos nulos.

A continuación, el sistema utiliza validación cruzada (Ver Tabla 55) y recupera una solución exitosa solo si la medida de similitud α ≥ 0,7. Este proceso es la comparación entre atributos PE_q y PE_r . Por ejemplo, la tabla 3 muestra un caso de prueba, se obtuvo una puntuación de α=0,86 mayor que 0,7, por lo que se agregó PE_r a la base de casos CB (PE).

Tabla 5. Prueba recuperación CBR_1

Test de caso	estado civil	estrato	Categoría comfa	departamento procedencia	municipio procedencia	atributos restantes de 46
PE_q	Soltero	3	2	Cauca	Santander	..
PE_r	Soltero	3	2	Cauca	Popayán	..
Test de caso	a	b	..		n	..
Test de caso	1	1	1	1	1	35

Con este sistema CBR_1 con pocos conocimientos de entrada se puede obtener una solución adecuada. Eso ofrece una alternativa más sencilla para completar los registros nulos y que se ajuste a las necesidades del sistema.

A continuación, en la Figura 17. Ejemplo de algoritmo usado para el completado de datos mediante CBR

```
n = 14590
Dim distancia(14590) As Variant
Dim cont As Variant

For i = 2 To n
    If Range("E" & i).Value = "" And Not Range("D" & i).Value = "" And Not Range("K" & i).Value = "" And Not Range("L" & i).Value = "" And Not Range
     "" And Not Range("N" & i).Value = "" And Not Range("S" & i).Value = "" Then
    'vacios
        Var1 = Range("D" & i) 'Genero
        Var2 = Range("K" & i) 'Departamento_reside
        Var3 = Range("L" & i) 'Municipio_reside
        Var4 = Range("M" & i) 'Zona_reside
        Var5 = Range("N" & i) 'Persona_cargo
        Var6 = Range("S" & i) 'Edad
    For j = 2 To n
        If Not Range("E" & j).Value = "" And Not Range("D" & j).Value = "" And Not Range("K" & j).Value = "" And Not Range("L" & j).Value = "" And
        ).Value = "" And Not Range("N" & j).Value = "" And Not Range("S" & j).Value = "" Then
        'llenos
          VarA = Range("D" & j)
          VarB = Range("K" & j)
          VarC = Range("L" & j)
          VarD = Range("M" & j)
          VarE = Range("N" & j)
          VarF = Range("S" & j)
          distancia(j) = Sqr((Var1 - VarA) ^ 2 + (Var2 - VarB) ^ 2 + (Var3 - VarC) ^ 2 + (Var4 - VarD) ^ 2 + (Var5 - VarE) ^ 2 +(Var6 - VarF) ^ 2)
          Else
           distancia(j) = 9999
        End If
    Next j
    cont = 100
    For k = 2 To n
        If distancia(k) <= cont Then
            cont = distancia(k)
            fila = Range("E" & k)
        End If
    Next k
    Range("E" & i) = fila
    End If
Next i
```

Figura 17. Ejemplo de algoritmo usado para el completado de datos mediante CBR

Finalizado el proceso de transformación se puede apreciar un incremento sustancial en la calidad de los datos necesarios para la implementación del proyecto, como se muestra en Figura 18 se ve la cantidad de datos nulos y errores dentro de algunos de los datos originales.

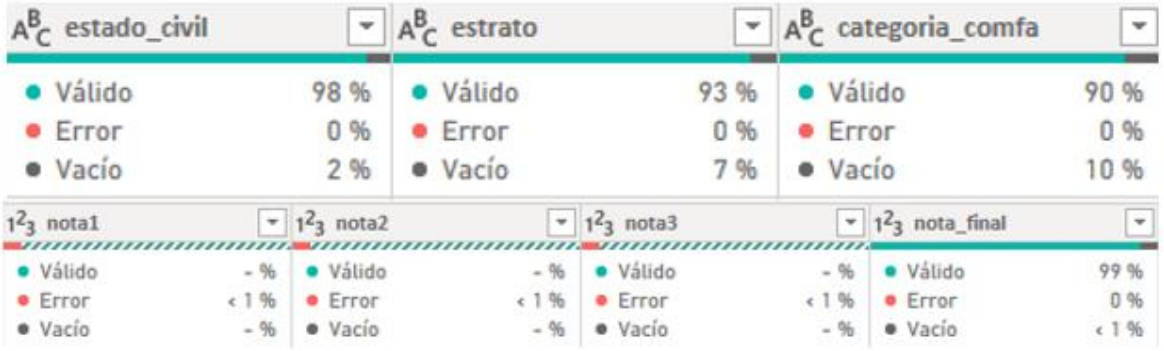

Figura 18. Calidad de algunos datos al momento de recibir el data set

Como se muestra en Figura 19, los datos ahora son aptos para su posterior carga y análisis.

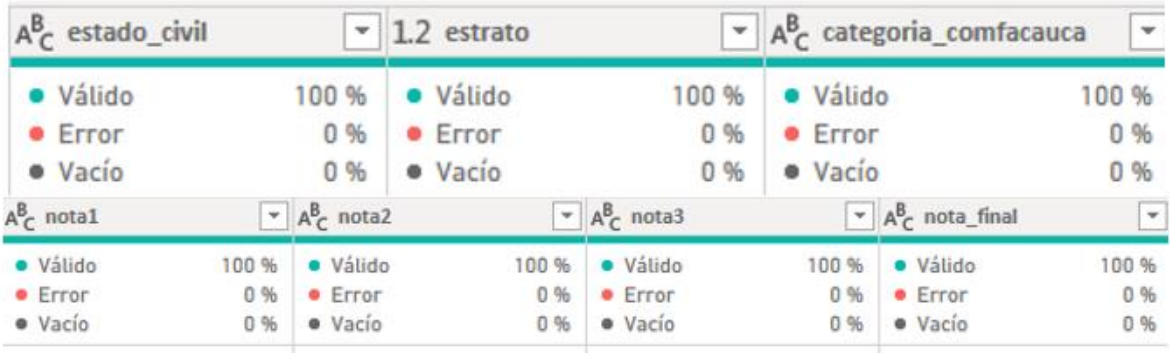

Figura 19. Calidad de algunos datos al momento de finalizar la transformación

¿Qué tanto influye en el trabajo remover registros con características, en comparación con el enfoque actual, que es crear data artificial?

El propósito de este trabajo SIN es ver como diferentes variables afectan la deserción, del mismo modo, realizar una probabilidad de deserción teniendo en cuenta las variables seleccionadas pertenecientes a diversos factores (económico, institucional, individual entre otros), bridando de esta manera una visión global del problema.

Si se presenta el caso, donde se encuentre registros vacíos que no son posibles de llenar, esto conlleva a no tener en cuenta la variable para el proceso analítico, teniendo como efecto un menor número de indicadores de deserción(menos información para la toma de decisiones), ahora bien, si las variables no tenidas en cuenta son demasiadas, pueden ocasionar la poca viabilidad del proyecto de inteligencia de negocios y no genere gran impacto, generando de esta forma la poca confianza para la toma de decisiones.
Por otra parte, lo concerniente al algoritmo que apoya a determinar la probabilidad de deserción, si contamos con pocas variables, tendremos menos recursos para mostrar información referente a la probabilidad de deserción de manera global.

2.4.2. Modelo Conceptual de Datos

Dada la naturaleza variada de los datos a tratar, se consideró adecuado agrupar los distintos factores o variables en 4 categorías cuya categorización se realizó dependiendo del nexo de las variables con un sector en concreto de la vida del estudiante. Las categorías se establecieron de acuerdo a 2 fuentes, como lo son: la investigación de literatura afina al tema y la estructura de los datos manejada por la entidad que suministró los datos. De tal manera se permite evidenciar e identificar de forma más clara las relaciones que se establecen entre las distintas categorías que afectan la deserción estudiantil. Dentro de las 4 categorías propuestas se tiene (Ver Figura 20), factores académicos: datos relacionados con el rendimiento académico; factores sociales: datos relacionados con adaptación social, familiar; factores institucionales: datos relacionados con la institución de educación superior como calidad académica, entre otras; y factores personales: engloba datos genéricos como edad, género, estrato económico entre otros.

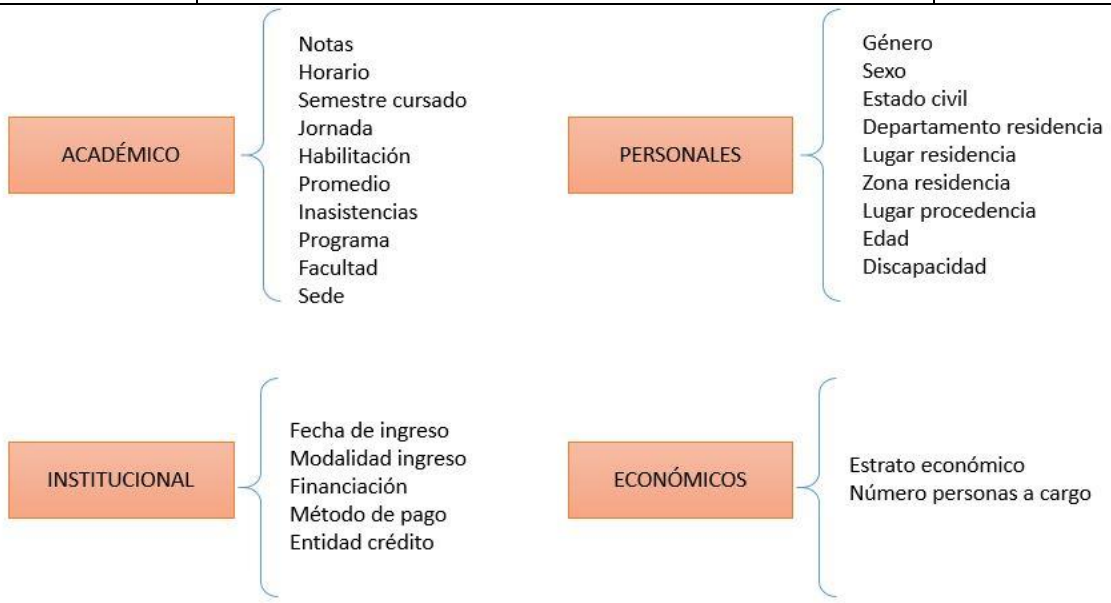

Figura 20. Categorización de las variables respecto a su influencia

2.4.3. Modelo Lógico de Datos

En todo sistema es de vital importancia tener una lógica robusta y congruente entre los aspectos del negocio, ya que se puede considerar que es la base de la correcta y efectiva implementación de este. Debido a la gran cantidad de factores influyentes en el área de la deserción estudiantil universitaria y a la compleja relación que mantienen entre sí, se opta por seleccionar los elementos más sobresalientes indicados por el mapeo sistemático implementado sobre la literatura consultada y a los cuestionarios realizados a la comunidad estudiantil.

La deserción es causada por un conjunto de factores de distinta índole como lo son académicos, sociales, económicos e inclusive institucionales; de tal forma se entra en la necesidad de construir un modelo lógico que refleje de la manera acertada las complejas relaciones que constituyen este fenómeno. Como se puede observar en la figura el modelo lógico en la deserción debe ser construido no solamente teniendo en cuenta los factores de influencia directa, sino que es necesario aludir también a los factores claves que causan indirectamente la deserción. Cada factor puede estar relacionado tanto a una causa principal como a una causa indirecta o inclusive pueden estar relacionadas de las dos formas, el sentido de la flecha en la representación de la relación indica el sentido de causa a efecto como muestra la Figura 21.

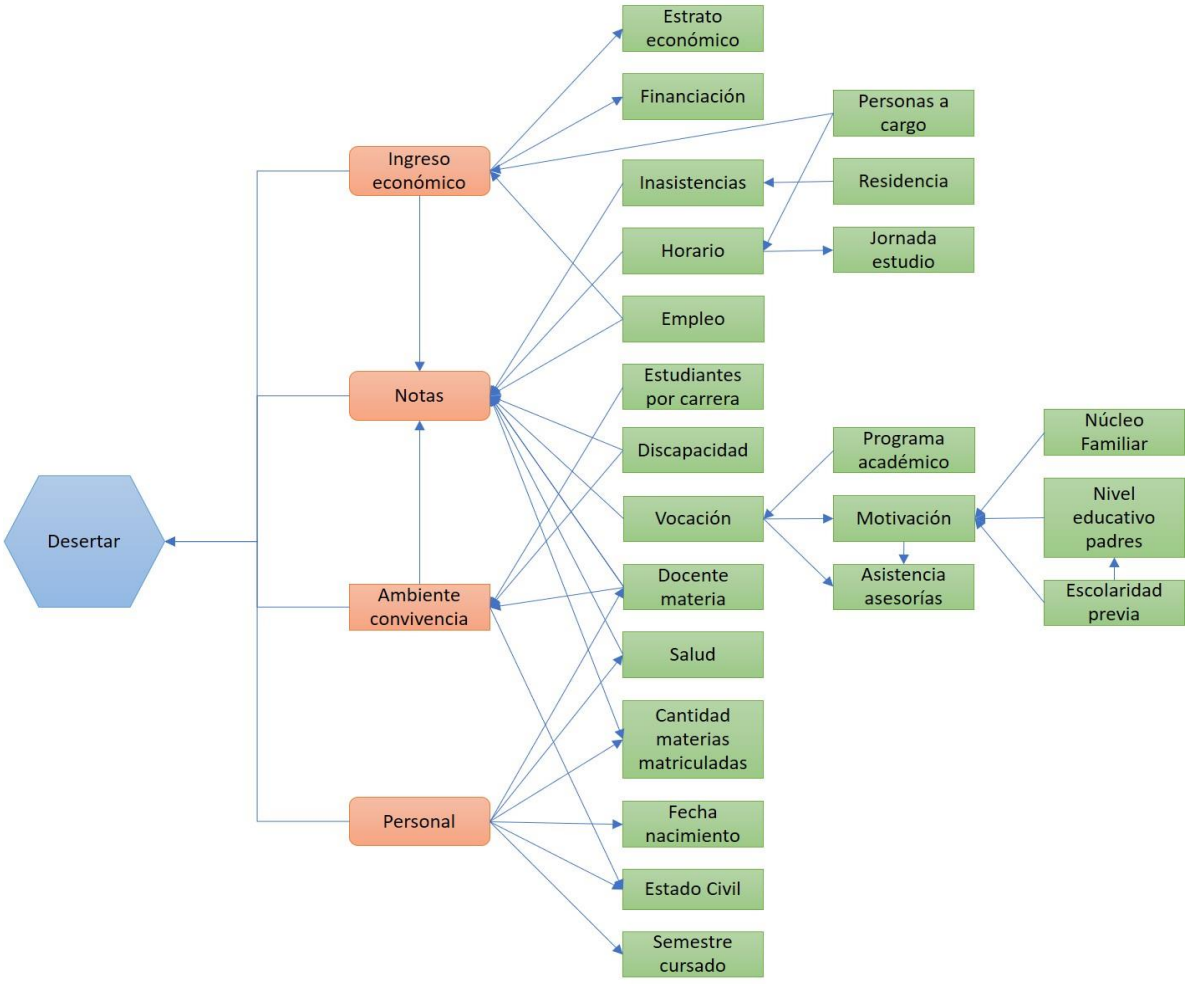

Figura 21. Relación entre los principales factores causantes de la deserción universitaria

2.4.4. Modelo Físico de Datos

Al realizar el modelado de datos, como última instancia se desarrolla el modelo físico, como lo expresa [4], el modelo físico consiste en convertir el modelo lógico en físico a través del uso de alguna herramienta y teniendo en cuenta elementos de la base de datos como tablas, atributos entre otros. Para este caso se implementa la herramienta de *Workbench* para bases de datos *MySQL*, a quien se le suministran los datos en formato *CSV*.

A continuación, en la Figura 22 se aprecia el modelo dimensional con el cual se implementó el *DM* para el departamento de permanencia académica de Unicomfacauca, el cual cuenta con 1 tabla de hechos y 6 dimensiones.

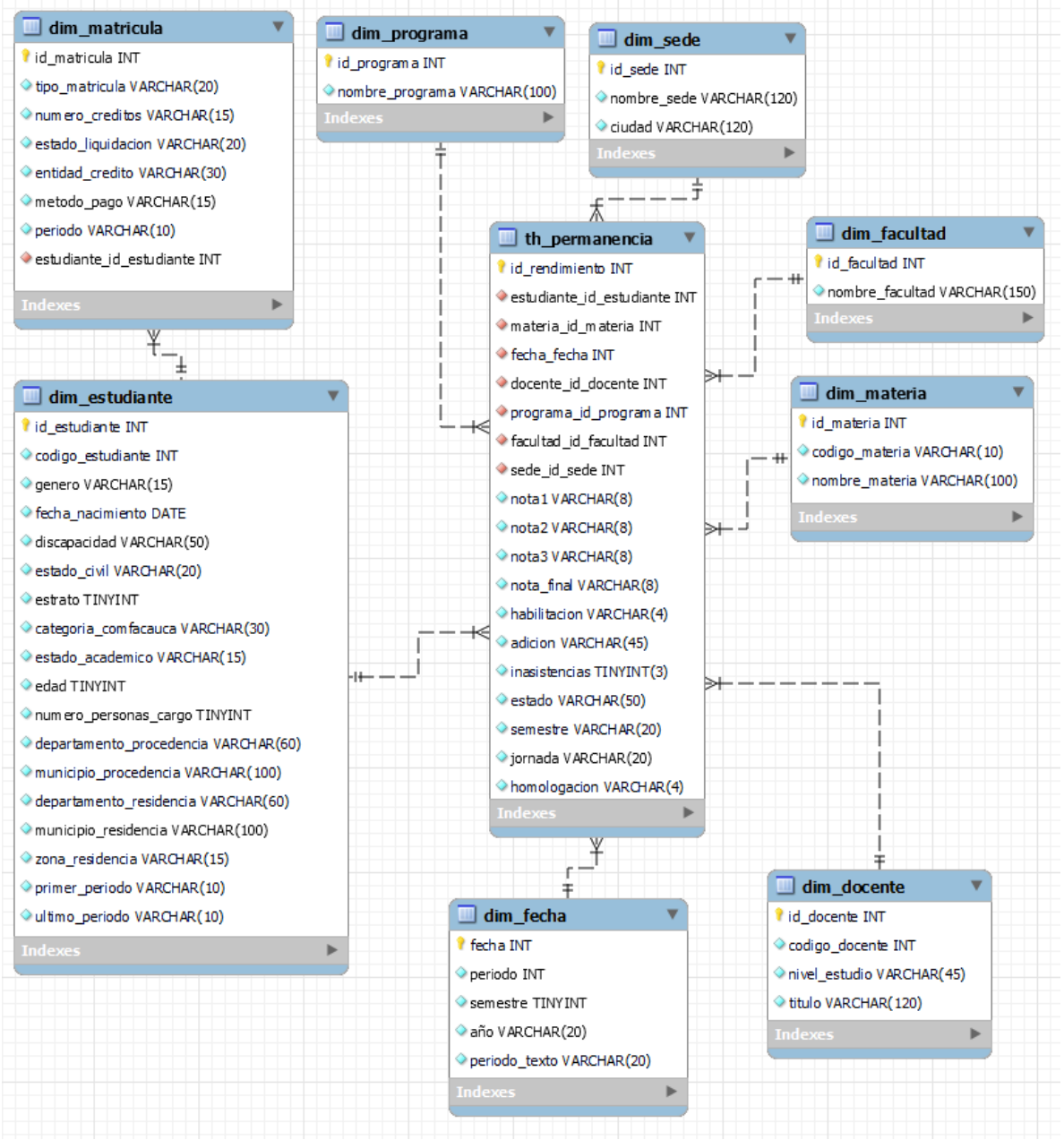

Figura 22: Modelo físico como esquema estrella del *DM* propuesto para el área de permanecía académica de Unicomfacauca

2.4.5. Descripción de campos por tablas

La Tabla 6 muestra las variables relacionadas con el docente, esta entidad se

debe tener presente para determinar que profesores tiene una mayor cantidad de alumnos reprobados afectando la permanencia estudiantil.

Tabla 6.Dimensión del docente dentro del DM

Dimensión docente		
Nombre	**Tipo de dato**	**Descripción**
id_docente	Int	Llave sustituta del docente
codigo_docente	Int	Llave de negocio del docente
nivel_estudio	varchar (45)	Nivel de estudio del docente (maestría, especialización)
Titulo	varchar (120)	Titulo obtenido en la carrera universitaria del docente

La Tabla 7 describe las variables relacionadas con la matricula del estudiante, esta entidad va relacionada con la permanencia académica debido a que contiene las variables que componen el factor Institucional, quien según lo indica la literatura consultada, cobra relevancia en la deserción estudiantil.

Tabla 7.Dimensión de la matricula dentro del DM

Dimensión matricula		
Nombre	**Tipo de dato**	**Descripción**
id_matricula	Int	Llave sustituta de matricula
tipo_matricula	varchar (20)	Registra si el estudiante es becado, normal o paga ciertos créditos
numero_creditos	varchar (15)	Número de créditos, si la matrícula es de tipo crédito, de lo contrario, no registra
estado_liquidacion	varchar (20)	Indica si es financiado, pago o iniciado
entidad_credito	varchar (30)	Entidad que financia la matricula del estudiante
metodo_pago	varchar (15)	Pago de matrícula de contado o es financiada
Periodo	varchar (10)	Intervalo de tiempo en que el estudiante a cancelado sus estudios
Estudiante_id_estudiante	INT	Llave foránea (llave de negocio de la tabla estudiante)

La Tabla 8 permite dilucidar las variables del estudiante, esta tabla contiene todos los aspectos individuales y socioeconómicos del estudiante, estos

factores son relevantes para la permanencia académica, en especial el socioeconómico.

Tabla 8.Dimensión del estudiante dentro del DM

Dimensión estudiante		
Nombre	**Tipo de dato**	**Descripción**
id_estudiante	int	Llave sustituta de estudiante
codigo_estudiante	int	Llave de negocio de estudiante
genero	varchar (15)	Sexo de estudiante
fecha_nacimiento	date	Fecha de nacimiento de estudiante
discapacidad	varchar (50)	Indica si el estudiante presenta o no presenta discapacidad
estado_civil	varchar (20)	Estado civil de estudiante (soltero, caso etc.)
estrato	int	Indica a que estrato pertenece el estudiante
estado_academico	varchar (15)	Indica si el estudiante es graduado, está en curos retirado
Edad	int	Edad del estudiante
numero_personas_cargo	int	Número de personal de la cual el estudiante se hace cargo económicamente
departamento_procedencia	varchar (60)	Departamento de donde procede el estudiante
municipio_procedencia	varchar (100)	Municipio de residencia de estudiante
departamento_residencia	varchar (60)	Departamento donde reside actualmente el estudiante
municipio_residencia	varchar (100)	Municipio donde reside el estudiante
zona_residencia	varchar (15)	Indica si reside en zona rural o urbana
primer_perido	varchar (10)	Periodo en donde el estudiante inicia sus estudios
ultimo_periodo	varchar (10)	Último periodo cursado por el estudiante

La Tabla 9, describe las variables relacionadas con la materia, aquí se ofrece información como: si la materia es registrada en diurna o nocturna, a que

programa y sede pertenece, entre otros. De este modo, determinar en qué medida estos aspectos afectan la permanencia académica.

Tabla 9.Dimensión de la materia dentro del DM

Dimensión materia		
Nombre	**Tipo de dato**	**Descripción**
id_materia	int	Llave sustituta de materia
codigo_materia	varchar (10)	Llave de negocio de materia
nombre_materia	varchar (80)	Nombre de la materia

La Tabla 10 describe las variables relacionadas con el rendimiento académico del estudiante, en esta tabla se guarda los datos (notas, inasistencias entre otros) del factor que tiene mayor envergadura en la deserción estudiantil.

Tabla 10.Tabla de hechos dentro del DM

Tabla de hechos permanencia		
Nombre	**Tipo de dato**	**Descripción**
id_rendimiento	int	Llave sustituta de rendimiento
estudiante_id_estudiante	int	Llave foránea de la tabla estudiante
materia_id_materia	int	Llave foránea de la tabla materia
nota1	varchar (8)	Nota de la materia obtenida en el primer corte
nota2	varchar (8)	Nota de la materia obtenida en el segundo corte
nota3	varchar (8)	Nota de la materia obtenida en el tercer corte
nota_final	varchar (8)	Nota final de la materia
Habilitación	varchar (4)	Indica si el estudiante presenta o no una habilitación
homologacion	varchar (4)	Indica si el estudiante homologo la materia
adicion	varchar (45)	Registra si el estudiante a pagado créditos adicionales para matricular la materia
inasistencias	int	Número de faltas a clase obtenidas por el estudiante
estado	varchar (20)	Indica si la materia ha sido aprobada o reprobada
semestre	varchar (20)	Indica el semestre al cual pertenece la materia

Tabla de hechos permanencia		
Nombre	**Tipo de dato**	**Descripción**
jornada	varchar (20)	Indica la jornada a la cual pertenece la jornada

La Tabla 11 describe las variables relacionadas con el programa académico del estudiante, en esta tabla se guarda el nombre y el identificador único de los programas académicos.

Tabla 11. Dimensión del programa dentro del DM

Dimensión programa		
Nombre	**Tipo de dato**	**Descripción**
id_programa	int	Llave sustituta del programa
nombre_programa	varchar (100)	Nombre del programa académico

La Tabla 12 describe las variables relacionadas con la sede de la corporación, se guardan el nombre de la sede, la ciudad en que se encuentra la sede de la universidad y el identificador único de la sede.

Tabla 12. Dimensión del rendimiento dentro del DM

Dimensión sede		
Nombre	**Tipo de dato**	**Descripción**
id_sede	int	Llave sustituta de la sede
nombre_sede	varchar (120)	Nombre de la sede
ciudad	varchar(120)	Ciudad de la sede

La Tabla 13 describe las variables relacionadas con las diferentes facultades de la corporación, se guardan el nombre de la facultad y el identificador único de la facultad.

Tabla 13. Dimensión del rendimiento dentro del DM

Dimensión facultad		
Nombre	**Tipo de dato**	**Descripción**
id_facultad	int	Llave sustituta de la facultad
nombre_facultad	varchar (150)	Nombre de la facultad

La Tabla 14 describe las variables relacionadas con los distintos formatos de fecha, se guardan la fecha como identificador único, el id del periodo

académico, el semestre, el año del periodo y el periodo escrito como texto.

Tabla 14. Dimensión del rendimiento dentro del DM

Dimensión fecha		
Nombre	**Tipo de dato**	**Descripción**
Fecha	int	Llave sustituta de la fecha
Periodo	varchar (100)	periodo académico
Semestre	int	semestre
Año	varchar (20)	Año
periodo_texto	varchar (20)	periodo académico escrito con letras

2.5. Data set filtrado

Basado en el estudio [13], donde los autores realizan un análisis de herramientas *open source* y de propietario buscando la mejor opción, señalan que la herramienta *Power BI* ha estado entre las mejores herramientas de propietario de los SIN en los últimos 5 años consecutivos, teniendo el 3er mejor promedio en los 5 años, según el cuadrante mágico de Gartner (aquí se clasifica las herramientas para hacer un SIN más importantes del mercado y es seguido por medios digitales y profesionales en el tema).

Adicional a esto se usa encuestas para solicitar el juicio de expertos en el tema, se realiza la consulta a tres personas:

- Gineth Magaly Cerón - Ingeniera en electrónica y telecomunicaciones, PH. D en telemática
- Francisco Javier Obando - Ingeniero en sistemas, Maestría en Computación
- Guillermo Cifuentes - Ingeniero en sistemas, consultor en *Power BI*

Estas personas especialistas en SIN sugieren la herramienta *Power BI* como una buena elección, por otra parte, la universidad Unicomfacauca nos proporciona los equipos necesarios con la herramienta instalada para la implementación del sistema, por dichas razones, se selecciona la herramienta *Power BI* para el desarrollo del sistema.

Usando el editor *PowerQuery* (véase Figura 23) para la transformación de los datos se llevó a cabo una normalización, aumentando significativamente su calidad,

todo esto, mediante procesos de corrección de ortografía, error de escritura y significado ambiguo, detalles documentados anteriormente.

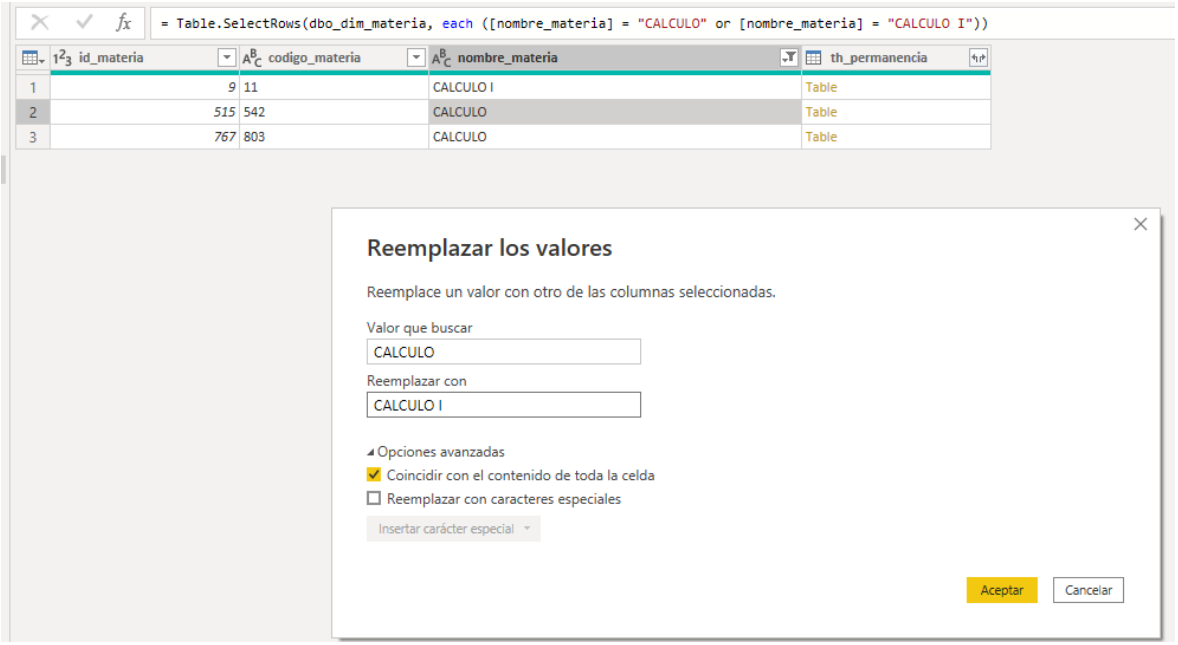

Figura 23. Editor PowerQuery llevando a cabo corrección de nombres ambiguos

2.6. Discusión

Se hizo una encuesta, que trata de tomar información de persona que tiene gran experiencia con un tema específico, en este caso la deserción, las personas encuestadas fueron profesores y estudiantes que de cierta manera están directamente involucrado con la deserción, como resultado se obtuvo que el factor académico junto con el económico son los que cobran mayor relevancia al momento de desertar.

En este sentido, mediante uso de mapeo sistemático se hace una revisión de la literatura, se consultó trabajos de investigación respecto a la deserción, de aquí se tomaron las variables de mayor peso, como resultado de este trabajo se tiene un Excel y gráficos de frecuencia que identifican las variables (estas variables están categorizadas en diferentes factores) que se requerían para hacer su respectivo proceso analítico.

Por último, se tiene un *data lake* con los datos que fueron facilitados por la Corporación Universitaria Comfacauca de las bases de datos administrativas como lo son el SIGA y el SISPE, estos datos que fueron proporcionados son

aquellos a los cuales se les va aplicar su debido proceso analítico.

2.7. Conclusiones

Tras la implementación de las encuestas a docentes y estudiantes de educación superior según su experiencia, se puede concluir que el factor académico junto con el factor económico, son dos aspectos que cobran gran relevancia en la deserción de un estudiante, por lo tanto, estos requieren una mayor atención con las autoridades encargadas de la toma de decisiones.

Por otra parte, y tras haber usado un mapeo sistemático se logró identificar las variables causantes de la deserción, clasificadas en 3 categorías más importantes como lo individual (incluye lo económico y demográfico), académico e institucional, siendo estos aspectos los tenidos en cuenta para su respectivo proceso analítico.

2.8. Metodología

Para el desarrollo de este proyecto, la Investigación de Ingeniería utilizó la metodología [13]. Siguiendo esta metodología a través de sus etapas: 1) Se obtiene una base conceptual a través de una revisión de la literatura. 2) Para desarrollo de sistemas, el diseño centrado en el usuario (UCD) se utiliza la metodología [3] y la metodología CRISP-DM [4] describe las tareas de minería de datos, con el fin de distribuir gestión de la información y concretar la minería de datos adecuada. 3) Para realizar la evaluación, se realizó una prueba de concepto de consultas SQL con base en las decisiones que desean tomar en Unicomfacauca la institución educativa donde se implementará el sistema. 4) existen muchas metodologías para el desarrollo de almacenes de datos, pero se imponen dos que son muy importantes, la metodología Kimball y la de Immon, para la construcción del DM de este proyecto se implementa la metodología Kimball [50] debido a que se construye el almacén de datos de menor hacia mayor escala, lo contrario a Immon que manifiesta construir todo un almacén a la vez.

¿Análisis con expertos para identificar variables influyentes en la deserción?

Se lleva a cabo una serie de encuestas a un grupo de expertos para obtener información e identificar las variables más influyentes dentro de la deserción estudiantil, esto se hizo siguiendo las características como. Heterogeneidad: aquí participan expertos de diferentes ramas, en este caso profesores y estudiantes

de educación superior de diversas universidades, Anonimato: durante este proceso de recolección de información, ninguno de los expertos encuestado conoce la identidad de los otros miembros del grupo. Los resultados obtenidos arrojan que las variables que más influyen dentro de la deserción son el factor académico y el factor económico, estos resultados coindicen con los obtenidos del mapeo sistemático que consistió en seleccionar las variables más relevantes según la literatura de proyectos que estudian la deserción.

En la Figura 24**¡Error! No se encuentra el origen de la referencia.** se observa el modelo de referencia de la metodología CRISP-DM. Esta metodología se siguió de la siguiente manera: las fases Comprensión del negocio y Comprensión de datos se utilizan para conocer los objetivos y requisitos del proyecto desde una perspectiva del negocio, y luego convertir este conocimiento en una definición de problema de toma de decisiones y minería de datos, así diseñar un plan preliminar para lograr el objetivo; y comenzar a recopilar datos, luego familiarizarse con los datos, identificar problemas de calidad de los datos, descubrir la información relevante en la vida útil sobre los datos o detectar subconjuntos interesantes para formar hipótesis sobre información oculta. El nivel del modelo de datos conceptual del marco propuesto se crea utilizando estas dos fases.

El nivel de modelo del marco propuesto se crea utilizando: el nivel de modelo de datos lógicos se basa en la preparación de datos fase que incluye todas las actividades necesarias para construir el conjunto de datos a partir de los datos brutos iniciales. TasNs incluyen tabla, caso y atributos elección, así como transformación y limpieza de datos para herramientas de modelado. Siguiendo la metodología, la fase de Modelado que selecciona y aplica una variedad de técnicas de modelado, y calibrar los parámetros de la herramienta a valores óptimos, se utiliza para crear el nivel de modelo de datos físicos del *framework* que va a describirse en la Sección 3. Por último, es importante concluir, socializar y presentar los resultados obtenidos

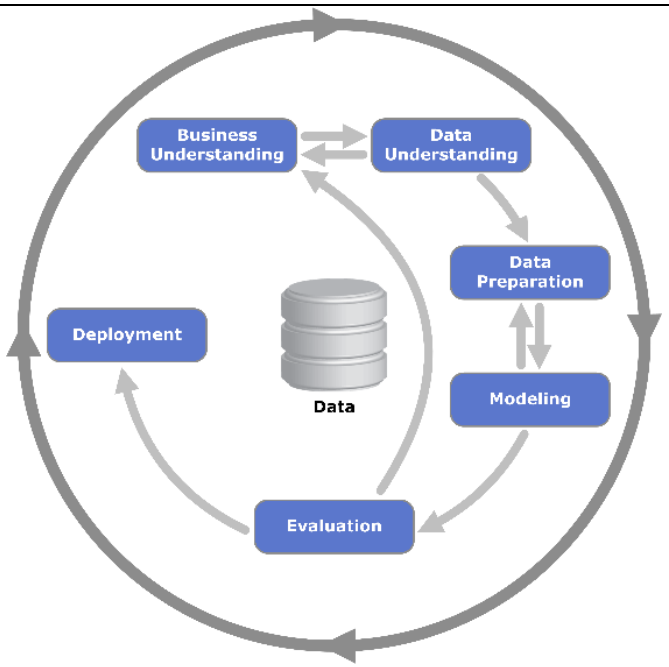

Figura 24: Modelo de referencia de la metodología CRISP-DM

Para llevar a cabo el desarrollo y la implementación de un SIN que comprende aspectos como la creación del DM junto con los algoritmos de minería de datos, se escoge la metodología Iteración investigativa que es una metodología de tipo cualitativa, que se presenta como un método que se puede aplicar con éxito para impulsar preguntas de investigación en contextos como la inteligencia artificial, robótica y otros contextos donde se presente la relación hombre-máquina. La iteración consiste en 4 pasos: la observación, identificación del problema, el desarrollo de la tecnología y pruebas de campo e inmediatamente una nueva iteración.

1. **Observación**: este es el primer paso, y consiste en hacer una observación del campo donde se va a trabajar con el fin de identificar necesidades y falencias, a las que se hará seguimiento ya que irán cambiando a medida que avanza el proyecto. En las primeras iteraciones la observación se puede centrar exclusivamente en ver las iteraciones o actividades del entorno estudiado con sus responsables inmediatos.

2. **Identificación del problema**: en el segundo paso se usan las observaciones para identificar el problema o pregunta de investigación, como en los demás pasos fluctúa según su iteración. En la primera iteración se centra en revisar la documentación acerca de la observación como lo pueden ser informes,

anotaciones y/o valoración de expertos. En futuras iteraciones se centra en las pruebas y experimentos del ciclo anterior con el fin de determinar si el problema fue solucionado o no; si fue solucionada la siguiente iteración tratará otro problema, de lo contrario seguirá con el problema inicial.

3. **Desarrollo de la tecnología**: una vez se haya plenamente identificado el problema dentro del campo, se procede a desarrollar la tecnología requerida, en este caso un SIN. Para apoyar el proceso de desarrollo del SIN se usará la metodología propuesta en [51] que permite el desarrollo ágil de bodegas de datos apoyado de la metodología Kimball.

4. **Pruebas de campo**: consiste en argumentar la nueva tecnología desarrollada con el fin de demostrar que está aborda toda la problemática encontrada, de no serlo así esta fase ofrece la capacidad de perfeccionar los aspectos inconclusos en la solución presentada o cambiar el abordaje tecnológico del problema; en este último caso iniciando una nueva iteración a partir de lo aprendido en la iteración anterior.

En la Figura 25 se presenta el diagrama del patrón de la metodología iterativa, que facilitó el proceso de gestión y desarrollo del sistema. Ya que sigue los pasos para generar prototipos funcionales en corto tiempo y permitió probar de forma ágil el SIN propuesto.

Figura 25: Diagrama del patrón de investigación iterativa, los cuatro segmentos del patrón: observar, identificar, desarrollar, y Test se muestran en su relación cíclica.

[52] "Como su nombre lo indica, la intención es que el patrón se aplique varias veces seguidas, utilizando múltiples iteraciones del ciclo para abordar un problema en lugar de un solo conjunto de Observar, Identificar, Desarrollar, Probar. De hecho, sin múltiples iteraciones, esto degenera en un modelo monolítico de desarrollo en cascada.".

Adicional a esto se desea implementar la metodología FOCUS GROUP, esta última es una metodología de tipo cualitativa que ayuda a determinar el impacto del desarrollo de sistemas dentro de la organización, en este caso, un SIN, la metodología consiste en realizar entrevistas y/o encuesta a un grupo de personas que se relacionan directamente con el sistema para verificar el impacto y determinar si la tecnología cumple con las expectativas de los usuarios.

La metodología CRISP-DM como complemento a la metodología investigación iteración, debido a que CRISP-DM fue diseñada exclusivamente para llevar a cabo proyectos donde esté implicado la minería de datos y aunque existen otras

metodologías similares, esta es una de las más usadas en el contexto académico e industrial según encuestas realizadas en los últimos años donde se mide el grado de usabilidad de las principales metodologías para minería de datos. La metodología CRISP-DM está dividida en 6 fases descritas a continuación:

1. **Comprensión del negocio**: en esta fase se busca entender cómo funciona la empresa y lo que se pretende solucionar, entendiendo sus necesidades como también verificar que se cuenta con los datos suficientes para el desarrollo de esta tecnología.

2. **Comprensión de los datos**: una fase donde se demanda un mayor esfuerzo, aquí se recolectan los datos que serán procesados, descritos y verificados

3. **Preparación de los datos**: aquí se procede a fomentar la limpieza de datos y la integración de los mismo (copiar datos de diferentes fuentes a un mismo sitio).

4. **Modelado**: tras finalizar la fase anterior, se debe estructurar los datos y escoger la técnica que más apropiada para resolver el problema.

5. **Evaluación**: en esta instancia se evalúa el modelado, si cumple con ciertos criterios de éxito, se hace la explotación del mismo.

6. **Implantación**: como último paso se transforma el conocimiento adquirido en acciones para combatir la problemática y se hace un informe final donde se explique los resultados obtenidos.

Como tercera y última metodología se desea implementar *FOCUS GROUP*, está última es una metodología de tipo cualitativa que ayuda a determinar el impacto del desarrollo de sistemas dentro de la organización, en este caso, un SIN, la metodología consiste en realizar entrevistas y/o encuesta a un grupo de personas que se relacionan directamente con el sistema para verificar el impacto y determinar si la tecnología cumple con las expectativas de los usuarios.

Capitulo 3: IMPLEMENTACIÓN DEL DATAWAREHOUSE - DW

Bodega de datos (*Data Warehouse* - DW)

Se puede definir como un gran almacén de datos recolectados previamente acerca de un modelo de negocio en específico con motivo de tener la información más compacta para su análisis y con un registro histórico de los eventos relacionados con la idea central del negocio. Se construye gracias al proceso de ETL y con finalidad de implementar herramientas de *data mining*.

[53] Dice que “Su función esencial es ser la base de un sistema de información gerencial, es decir, debe cumplir el rol de integrador de información proveniente de fuentes funcionalmente distintas (Bases Corporativas, Bases propias, de Sistemas Externos, etc.) y brindar una visión integrada de dicha información, especialmente enfocada hacia la toma de decisiones por parte del personal jerárquico de la organización.”

Data Mart (DM)

Es un bloque que conforma una bodega de datos en los cuales se almacena información específica acerca de uno de los pilares de la idea central del negocio, entendiéndose como un filtro en el cual se pueda acceder y explorar a determinados datos con más facilidad. En síntesis, se puede decir que los *DM* son pequeños *data warehouse* centrados en un tema o un área de negocio específico dentro de una organización.

Se puede decir que según [54] “se trata de una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. “

Diferencias entre DW y DM:

[55] La diferencia más marcada entre el DM y el *Data Warehouse* es el alcance que tienen, un DW está planteado para ser un almacén de datos central para toda la empresa, donde los datos de distintas áreas convergen. El DM es visto más como un subconjunto centrado a un área o tema específico de negocio que hace parte de una DW, como lo dice Kimball “cada DM debe estar orientado a un proceso determinado dentro de la organización, por ejemplo, a pedidos de clientes, a compras, a inventario de almacén, a envío de materiales, etc.”

En la**Tabla** 15 se establecen algunas de las diferencias más marcadas entra una Bodega de datos y un DM, ayudándonos a definir el límite entre estas dos entidades que puede ser en ocasiones difícil de identificar.

Tabla 15. Diferenciación entre una DW y un DM (Adaptado de [56]**)**

	Data Warehouse	Data Mart
Alcance	Creado con el fin de solucionar las necesidades de información de toda la empresa	Creado para proporcionar información a un área específica
Objetivo	Optimizar la integración de fuentes	Optimiza la entrega de reportes
Fuentes de integración	Es integrado por muchas fuentes de información	integrada solo por pocas fuentes de información
Tamaño	El tamaño puede variar de 100GB a más de un TB	El tamaño es inferior a 100GB
Pertenencia	Pertenece a toda la organización	Pertenece al área la cual va dirigida

3.1. Diseño del DW

3.1.1. Selección de herramientas

Gestor de base de datos

[57] De acuerdo con lo investigado acerca de los sistemas gestores de base de datos se evidencia que Microsoft SQL Server es una de las herramientas con mayor auge en la actualidad debido a que brinda características como el soporte exclusivo por parte de Microsoft, estabilidad, escalabilidad, seguridad, posibilidad de cancelar consultas, simplicidad en su lenguaje de consultas, multiplataforma y soporte de transacciones. La gran demanda que Microsoft presenta por parte del mercado lo hace una herramienta favorable al momento de generar experiencia para el futuro laboral, pese a ser de licencia comercial brinda gratuitamente una versión mínima con funcionalidades bastante útiles.

En la Figura 26: Cuadrante mágico de Gartner de los SGBD para el año 2019, muestra el cuadrante mágico de Gartner el cual expone los sistemas de gestión de base de datos más sobresalientes del mercado, basándose en diferentes aspectos de innovación que brinda la herramienta.

Figura 26: Cuadrante mágico de Gartner de los SGBD para el año 2019 (Tomado de [58] **)**

Como lo detalla [59] en cuanto a precios MSS (Microsoft SQL Server) tiene un costo moderado con respecto a herramientas como Oracle que pese a tener gran acogida tiene un costo aproximadamente 20% mayor en el costo de licencia y un costo de capacitación superior al presentado por MS server.

[60] Se realizó una comparativa de todos los sistemas gestores de base de datos del mercado tomando en cuenta las experiencias y opiniones de los usuarios, posicionando a Microsoft como uno de los favoritos con mejor valoración positiva junto con MongoDB, Oracle e IBM entre otros.

Adicional a lo anterior, la Corporación Universitaria Unicomfacauca cuenta con

mayor predisposición a la obtención de licencias de Microsoft por medio de paquetes de licencias que posee en la actualidad como PowerBI, lo que a su vez disminuye los problemas de compatibilidad que se puedan presentar entre las diversas plataformas implementadas en la inteligencia de negocio.

Considerando los argumentos expuestos se decidió utilizar Microsoft SqlServer Express Edición (versión gratuita) para la administración de del DM de permanencia propuesto, dejando abierta la posibilidad de la compra de la licencia al momento de la puesta en marcha de la inteligencia de negocio.

3.2. Implementación

Una vez se tiene el modelo dimensional del *DataMart* definido se genera el código SQL para crear el esquema (véase Figura 27Figura 28)

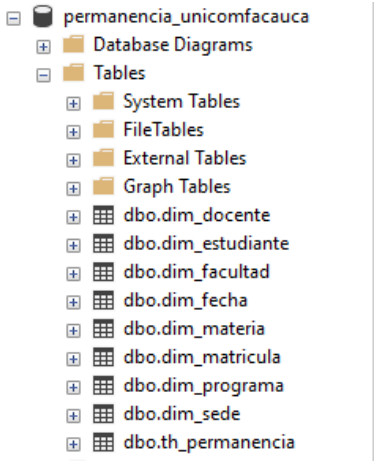

Figura 27. Esquema del DataMart propuesto

Teniendo preparado el *DataMart* se concluye el proceso ETL poblando las dimensiones y la tabla de hechos usando como insumo los datos del *data staging* mediante consultas SQL (véase Figura 28).

```
 --- poblar estudiate ---
INSERT INTO [permanencia_unicomfacauca].[dbo].[dim_estudiante]("codigo_estudiante","genero","fecha_nacimiento","discapacidad","estado_civil","estrato",
"categoria_comfacauca","estado_academico","edad","numero_personas_cargo","departamento_procedencia","municipio_procedencia","departamento_residencia",
"municipio_residencia","zona_residencia","primer_periodo","ultimo_periodo")
SELECT "codigo_estudiante","genero","fecha_nacimiento","discapacidad","estado_civil","estrato","categoria_comfa","estado_academico","edad","personas_a_cargo",
"departamento_procedencia","municipio_procedencia","departamento_reside","municipio_reside","zona_reside","primer_periodo","ultimo_periodo"
FROM [data_staging].[dbo].[estudiante]

 --- poblar fecha ---

INSERT INTO [permanencia_unicomfacauca].[dbo].[dim_fecha]("fecha","año","semestre","periodo","periodo_texto")
SELECT "idfecha","año","semestre","idperiodo","periodo"
FROM [data_staging].[dbo].[fecha]

------ poblar materia --------

insert into dim_materia ("codigo_materia","nombre_materia",)
select data_staging.dbo.materias.codigo_materiaHorario,"materia"
from data_staging.dbo.materias


--------poblar docente---------
INSERT INTO [permanencia_unicomfacauca].[dbo].[dim_docente]("codigo_docente","nivel_estudio","titulo")
SELECT "codigo_docente","nivel_estudio","titulo"
FROM [data_staging].[dbo].[docente]

--------poblar matricula-------------
```

Figura 28. Algunas consultas SQL utilizadas para realizar el poblado del DataMart

La tabla de hecho debe poblarse teniendo en cuenta las llaves foráneas de las dimensiones a las cual se hacen referencia dentro de la tabla de hechos, en el caso del datamart propuesto la tabla de datos académicos la cual presentaba datos transaccionales de las notas se convirtió en la tabla de hechos, incluyendo así las llaves foráneas mencionadas (véase Figura 29).

Results | Messages

	estudiante_id_estudiante	materia_id_materia	fecha_fecha	docente_id_docente	programa_id_programa	facultad_id_facultad	sede_id_sede	nota1	nota2	nota3	nota_final	homologacion	habilitacion	adicion	inasistencias	estado	jornada	semestre
1	1	64	200701	3	24	3	6	4.2	4.3	4.5	4.35	NO	NO	NO	0	APROBADO	DIURNA	V
2	1	25	200701	32	24	3	6	3.8	4	3.8	3.86	NO	NO	NO	0	APROBADO	DIURNA	V
3	1	34	200501	30	24	3	6	3.1	3.1	2.8	3	NO	NO	NO	0	APROBADO	DIURNA	I
4	1	42	200707	42	24	3	6	3.2	3.9	3.7	3.61	NO	NO	NO	0	APROBADO	DIURNA	V
5	1	74	200807	24	24	3	6	0	0	0	3.9	SI	NO	NO	0	APROBADO	DIURNA	III
6	1	94	200707	14	24	3	6	4.5	4	4	4.15	NO	NO	NO	0	APROBADO	NOCTURNA	VI
7	1	72	200707	92	24	3	6	4.2	3.3	4	3.85	NO	NO	NO	0	APROBADO	DIURNA	VI
8	1	32	200707	36	24	3	6	3.4	4.5	4.6	4.21	NO	NO	NO	0	APROBADO	DIURNA	VI
9	1	87	200707	25	24	3	6	3.3	3.3	4.7	3.86	NO	NO	NO	0	APROBADO	DIURNA	VI
10	1	88	200707	21	24	3	6	4.5	4	4.2	4.23	NO	NO	NO	0	APROBADO	DIURNA	VI
11	1	86	200701	21	24	3	6	3.3	4.2	5	4.25	NO	NO	NO	0	APROBADO	DIURNA	V
12	1	71	200701	21	24	3	6	4.4	4.2	5	4.58	NO	NO	NO	0	APROBADO	DIURNA	V
13	1	2	200701	22	24	3	6	3.5	4.2	5	4.31	NO	NO	NO	0	APROBADO	DIURNA	V
14	1	34	200501	30	24	3	6	0	0	0	3	SI	NO	NO	0	APROBADO	DIURNA	I
15	1	9	200501	889	24	3	6	0	0	0	3.5	SI	NO	NO	0	APROBADO	DIURNA	I
16	1	4	200501	44	24	3	6	0	0	0	4.2	SI	NO	NO	0	APROBADO	DIURNA	I
17	1	37	200801	13	24	3	6	0	0	0	3.9	SI	NO	NO	0	APROBADO	DIURNA	I
18	1	6	200607	102	24	3	6	0	0	0	3.7	SI	NO	NO	0	APROBADO	DIURNA	IV
19	1	5	200607	846	24	3	6	0	0	0	4.1	SI	NO	NO	0	APROBADO	DIURNA	IV
20	1	26	200801	114	24	3	6	0	0	0	5	SI	NO	NO	0	APROBADO	DIURNA	IV
21	1	3	200607	57	24	3	6	0	0	0	3.7	SI	NO	NO	0	APROBADO	DIURNA	IV
22	1	79	200601	797	24	3	6	0	0	0	4.1	SI	NO	NO	0	APROBADO	DIURNA	III
23	1	59	200807	107	24	3	6	0	0	0	4.2	SI	NO	NO	0	APROBADO	DIURNA	II
24	1	70	200601	395	24	3	6	0	0	0	3.7	SI	NO	NO	0	APROBADO	DIURNA	III

Figura 29. Tabla de Hechos una vez terminado el proceso de carga

3.2.1.Selección de técnicas de minería

Basado en el libro “introducción a la minería de datos” [61], quien afirma que, para la elección de una técnica de minería, se debe, como primer ítem identificar si el problema definido en este trabajo necesita una solución de un modelo de descripción (clasificación) o de predicción, dependiendo la elección se procede a analizar las diferentes técnicas enfocadas en los diferentes estilos.

El presente trabajo se basa en determinar en qué medida diferentes factores inciden en la deserción estudiantil y sobre estas estadísticas poder predecir que estudiantes tiene alta probabilidad de deserción, como se describió anteriormente, se necesitó predecir que estudiantes van desertar, por consiguiente, según [61] las técnicas adecuadas para procesos de predicción son las siguientes:

- K- Vecino cercano
- Árboles de decisión
- Análisis discriminante
- Redes Neuronales

Análisis discriminante: esta técnica permite clasificar objetos a diferentes clases de manera fácil, a través de P variables llamadas variables discriminantes, y colocándolos en k grupos el cual permite su clasificación, la metodología de esta técnica discriminante consiste en el análisis de un submuestreo de datos, pero para que esta técnica arroje resultados de manera correcta, se deben de contar con 2 supuestos, el primero que los datos deben de estar con una distribución normal, la segunda es que los datos no deben ser atípicos, si no se cuenta con estos 2 supuestos los resultados arrojados no serían los más acertados, generando clasificaciones falsas, pues así lo señala [62].

Redes neuronales: esta técnica a diferencia de la discriminante no necesita el supuesto de tener datos en forma de distribución normal, esta técnica se utiliza para reconocer patrones, y con el paso del tiempo la red se va fortaleciendo aprendiendo de forma más rápida, lo que quiere decir que tiene la capacidad de aprender y mejorar su funcionamiento.

Una red neuronal no necesita de un algoritmo para resolver un problema, ya que ella genera su propia distribución de pesos (sinapsis, impulsos nerviosos transmitidos entre ellas) en los enlaces mediante aprendizaje, como las redes neuronales aprenden a diferenciar patrones, no es necesario crear modelos a priori ni necesidad de especificar funciones de distribución de probabilidad. La red neuronal está conformada por tres capas, la primera es la entrada que es la capa que recibe la información con su respectivo peso, la segunda es la capa oculta que contiene neurona conectadas para determinar las distintas topologías de la red, la última es la de salida quien ofrece información de la red hacia el exterior [63].

En la literatura revisada se presentan varios trabajos donde se ha implementado

en su mayoría árboles de decisión y redes neuronales, por lo que este aspecto, nos conlleva a crear 2 modelos, uno en basado en árboles de decisión y otro en redes neuronales, para seleccionar el modelo que mejor se acople a la problemática. Lo descrito anteriormente se hace porque según [61],no existe el modelo perfecto, de ahí que, se debe probar varios modelos y escoger el mejor.

K - Vecino cercano (K-NN):

Algoritmo de clasificación, regresión y de aprendizaje supervisado (requiere datos de entrenamiento), se caracteriza por no aprender de un modelo sino de "memorizar" las instancias de formación que luego utilizará como base de futuras predicciones, generalmente tomado como base para redes neuronales entre otros. Por lo anterior mencionado presenta un coste de memoria elevado y amplio tiempo de respuesta debido a la cantidad de datos almacenados necesarios para la clasificación de datos futuros. En el reconocimiento de patrones de este algoritmo usa las instancias de formación para determinar los valores más cercanos o próximos (distancia) al valor de referencia (vecinos cercanos). El valor K de este algoritmo se toma como la cantidad de vecinos a los cuales se quiere tomar como muestra, el cual debe ser definido al momento de construcción del modelo. Una pequeña cantidad de vecinos tendrá un bajo sesgo, pero una alta varianza, y un gran número de vecinos tendrá una varianza más baja pero un sesgo más alto por lo que el valor asignado de K representa una variante importante en la precisión de los datos siendo este el principal problema presentado por este algoritmo.

Según [64] se obtuvo un porcentaje de acierto del 67.07% para el algoritmo de K-NN en una muestra de 723 instancias con un valor k de 10, a comparación del 66.6% del algoritmo de árboles de decisión J48; se realizó una prueba con 6500 instancias obteniendo el algoritmo K-NN un 70% frente a un 98.8% del algoritmo de árboles de decisión J48, por lo cual se infiere que este último es mucho más preciso a mayor cantidad de instancias.

Árboles de decisión:

[65] Es un método de clasificación para la minería de datos el cual clasifica los datos en categorías ya establecidas con el fin de predecir la probabilidad de ocurrencia de un resultado en función de la relación con distintas variables, cabe añadir que es una técnica de aprendizaje supervisado por lo cual necesitan de un conjunto de datos de entrenamiento para su correcta ejecución. Es ideal usarla cuando se cuentan con gran cantidad de dimensiones y posibilidades ya que se asemeja a diagrama de flujo, en donde se evalúa cual es el camino o

decisión más adecuado en cada situación mediante reglas establecidas.

Como lo dice [37][66][30], gran parte de la literatura relacionada con la deserción estudiantil indica que en especial el algoritmo J48 arroja buenos resultados en la tarea de clasificación de estudiantes desertores. [30] En concreto compara la técnica de árboles de decisión con la técnica de *BayesNet*, encontrando un 96% de acierto por parte del algoritmo J48 contra un 88% de *BayesNet*.

En contrate a lo anterior [47] menciona que debido a su forma representativa simple al restringirlo a una representación de árbol o regla puede restringir significativamente la forma funcional del modelo y su poder de aproximación.[30]Tiene un porcentaje de acierto similar al de otros métodos de predicción, pero pueden variar en las dimensiones clasificadas como importantes como establecimiento educativo previo.

3.3. Evaluación

Con base a lo investigado sobre algoritmos de minería de datos y la recomendación de expertos se tomó la decisión de implementar el algoritmo de KNN (K Nearest Neighbors) debido a que algoritmos como redes neuronales presentan poca documentación en esta área y gran limitación por parte de la herramienta Power BI ,como sucede también respecto a otros algoritmos complejos, adicional a esto, dentro de los algoritmos candidatos a utilizar, KNN se acerca más a las necesidades del proyecto, permitiendo realizar un perfilamiento de los estudiantes propensos a desertar, tomando las variables más influyentes en la deserción que se consultaron y evaluando que estudiantes en curso de sus estudios tienen mayor similitud con las características de un estudiante desertor; de esta manera el algoritmo KNN marca la diferencia frente a algoritmos como el de árboles de decisión ,específicamente el J48, que ofrecen una jerarquía de dichas variables, imposibilitando un resultado imparcial y confiable.

La evaluación del algoritmo de minería de datos escogido, se realizó mediante el cambio de valores en columnas ya conocidas y ejecutando el algoritmo sobre estas, de tal manera que, si el algoritmo establece en estas columnas el valor que fue cambiando, nos indica que el porcentaje de acierto y predicción es el esperado (véase **Anexo D**). Por tales pruebas se confirmó la idoneidad de esta técnica para las necesidades del proyecto, no obstante, se hace necesario un estudio más a profundidad con respecto a esta área.

3.4. Discusión

3.4.1. Arquitectura

[67]Como lo define Ralph Kimball una bodega de datos es la unión de pequeños componentes departamentales dentro de una organización llamados *DM* los cuales van enfocados a un tema en concreto. Por esto, la DW debe estructurarse empezando desde lo pequeño y simple para ir aumentando su tamaño y complejidad, debido a esto es importante estructurar adecuadamente dichos *DM* para que una vez los demás departamentos de la organización conformen estos almacenes de datos la integración entre estos *DM* se verá reflejada por las dimensiones que puedan llegar a tener en común, conformando así una Bodega de datos robusta.

A corto plazo, en el término de la implementación del *DM*, este podría tanto alimentar como alimentarse de la información brindada por el departamento de sistemas, por medio de las dimensiones y tabla de hechos que se establezcan en el diagrama estrella del *DM* de dicho departamento. De esta manera la información depurada que es suministrada por el SIN previsto para el área de permanencia, no solo podría mejorar el seguimiento a los estudiantes desde este departamento, si no, también usado en otros aspectos por demás departamentos.

Con respecto a todo lo anteriormente expuesto, la propuesta del DM para el departamento de permanencia académica de Unicomfacauca debe ceñirse a estas pautas con el fin de lograr en un futuro la implementación de una bodega de datos robusta en su totalidad, la cual permita la adecuada explotación del conjunto de datos de los departamentos de la Corporación.

3.4.2. Selección de las técnicas

A pesar de que los diferentes tipos de técnicas de minería son muy buenos, la técnica de redes neuronales se descarta porque tiene un grado de complejidad más elevado, y como el proyecto hace énfasis en la visualización de datos, entonces se decide no emplear demasiado tiempo y esfuerzo en la implementación de esta técnica, sin embargo, entre la técnica de KNN vecinos cercanos y árboles de decisión, se selecciona KNN debido a que este tiene mayor precisión que arboles de decisión a de acuerdo a la cantidad de registros recibidos en el *dataset.*

3.4.3. Bodega de datos

Se presenta un desacuerdo al momento de asignar un nombre a almacén de datos donde se ingresaron los datos a estudiar, parte del grupo de investigación opto por llamarle a este almacén DW, pero la otra parte sugería que era un DM, por tal motivo se llevó a una socialización con argumentos por ambas partes del grupo para determinar la terminología del almacén de datos, tras no llegar a un acuerdo se toma como medida la consulta a un profesional del tema con mucha experiencia, quien sugirió que el almacén es un DW por su alcance.

Por otra parte y una vez construido el modelo dimensional copo de nieve, el paso a seguir es la creación del DW y la posterior carga de datos, aquí surge un problema, no se tenía bien en claro que programa usar para realizar estos pasos, y aunque la herramienta de inteligencia POWER BI permite hacer estos procesos, se decide crear y poblar el DM con la herramienta SQL SERVER debido a que se cuenta con mayor conocimiento en el uso de este programa, una vez realizado el proceso de creación y carga del DM, se exporta toda la información (DM) a POWER BI donde se realizar todo el proceso analítico

3.4.3.1. Análisis de datos

Terminado el proceso de extracción y transformación de los datos descritos anteriormente, se lleva a cabo el proceso de carga, donde los datos son cargados desde el gestor de base de datos SQL SERVER a DM con la herramienta de inteligencia de negocios Power BI, que nos facilita la forma para hacer el proceso analítico, transformando los datos en información que respalde la toma de decisiones de las autoridades encargados

Para conseguir las transformación de los datos en información y conseguir cumplir con todos los requisitos nombrados en el capítulo 2, se hizo necesario crear una serie de columnas calculadas (columnas creadas adicionalmente, para crear datos necesarios a partir de otros mediante una fórmula única) y medidas (se usan para calcular valores que requieren las visualizaciones) usadas para identificar de manera rápida en qué medida afectan las diferentes variables de deserción mediante la creación de reportes (gráficas).

A continuación, se describen las columnas calculadas y medidas creadas para el cumplimiento de requisitos:

los datos proporcionados por permanencia académica, específicamente en el factor individual del estudiante, no contiene el atributo “Edad” que represente la edad del estudiante, requerida para satisfacer uno de los requisitos, por lo

tanto, se crea la columna calculada llamada Edad a partir del atributo “Fecha de nacimiento” como se muestra en la siguiente ilustración:

$$= (HOY(\quad) - \$2862))/\,365$$

Figura 30. Medida para calcular la edad según la fecha de nacimiento

la imagen anterior refleja la creación de la columna calculada “Edad”, que se creó en el programa Excel y posteriormente se carga junto con los demás datos en *Poder BI*, esta columna fue creada porque fue una variable faltante en los datos que se identificó de forma inmediata, por lo tanto, se crea dentro del proceso de transformación.

Por otra parte, se cuenta con las medidas, en la tabla de hechos “Permanencia” se crea la medida llamada “Desertores”, esta medida cuenta la cantidad de desertores existente dentro de la universidad para su posterior correlación con las diferentes variables y mostrar sus estadísticas en una visualización. la siguiente imagen (Figura 31) muestra la fórmula usada para la creación de la medida “Desertores”:

$$desercion = calculate(distinctcount(th\_permanencia[estudiante\_id\_estudiante]),\ \dim\ \_estudiante[estado_{academico}] = "retirado")$$

Figura 31. Medida para calcular los estudiantes desertores

También se crea la medida llamada “Reprobados”, esta medida hace un conteo de los estudiantes que han perdido cierta cantidad de materias con el fin de identificar cuáles son aquellas materias más pérdidas, los profesores que más reprueban estudiantes, entre otros. la forma en que se crea la medida se refleja en la siguiente figura:

$$reprobados = calculate(count(th\_permanecia[materia\_id\_materia]),\ th\_permanencia\ [estado]\ = "REPROBADO")$$

Figura 32. Medida para calcular los estudiantes reprobados

Y por último se crea la medida que lleva por nombre “Deserción matrícula”, medida usada para contar los desertores y asociarlos según variables del factor institucional como por ejemplo el método pago, como se observa en la Figura 33.

$$desercion\_matricula = calculate\ (count(\dim\ \_matricula[estudiante\_id\_estudiante]),\ \dim\ \_estudiante[estado\_acadaemico] = "retirado")$$

Figura 33. Medida para calcular desertores según matricula

3.4.3.2. Inconvenientes

Ya creada la columna calculada “Edad” y la medida “Desertores”, se crea el reporte para identificar cual son los rangos de edad donde los estudiantes están más propensos a abandonar sus estudios, con el reporte creado, se dio un supuesto cumplimiento al requerimiento de desertores por rango de edad. pero, el rango de desertores por edad encontrado no era el correcto, porque, en el proceso analítico aplicado para encontrar el rango de edad más donde se presenta más desertores se tuvo en cuenta la edad actual de la persona, en lugar de usar la edad en que la persona abandona sus estudios.

Para solucionar el reporte de desertores por edad, se crea otra columna calculada a partir de la columna “Edad” con el objetivo de encontrar la edad en que el estudiante deserta de sus estudios superiores (véase Figura 34)

$$edad\_desercion \;=\; [edad] \;-((\;38\;-[ultimo\_periodo])\;/2)$$

Figura 34. Medida para calcular la edad de deserción

en la imagen anterior se muestra la forma que fue creada la columna calculada llamada “Edad deserción”, donde a la edad actual del estudiante se le resta la cantidad de tiempo que ha transcurrido desde que desertó de sus estudios, obteniendo de este modo la edad de deserción, proporcionando información correcta que satisface el requisito de rango de edad de mayor número de estudiantes que desertan.

3.4.3.3. Predicción de los estudiantes con probabilidades desertar

Aquí se realizó un proceso de análisis en los datos mediante el algoritmo KNN (*K- Nearest Neighbors*) con distancia euclidiana, con un K vecinos igual a 9, en el cual se comparó los datos de los estudiantes que desertaron en el pasado con los estudiantes que actualmente están en curso sus estudios de esta manera se pudo perfilar los posibles alumnos qué en sus características tengan similitud con estudiantes que desertaron. El algoritmo comparó diversas características del estudiante, llamados a partir de ahora “predictores”, entre las que están: el género, la edad, el estrato socioeconómico, el programa, la zona donde reside, discapacidad, personas a su cargo, departamento y municipio de procedencia, departamento y municipio de residencia.

El algoritmo fue implementado en PowerBI con lenguaje Phyton (véase Figura 35) tomando dos diferentes *dataset* para su ejecución, uno de entrenamiento y

otro de prueba, la versatilidad de PowerBI facilita la ejecución de diferentes lenguajes de programación como R y Phyton permitiendo realizar tanto el ETL como minería de datos para el sistema propuesto en estos lenguajes, todo sin necesidad de instalar librerías o extensiones.

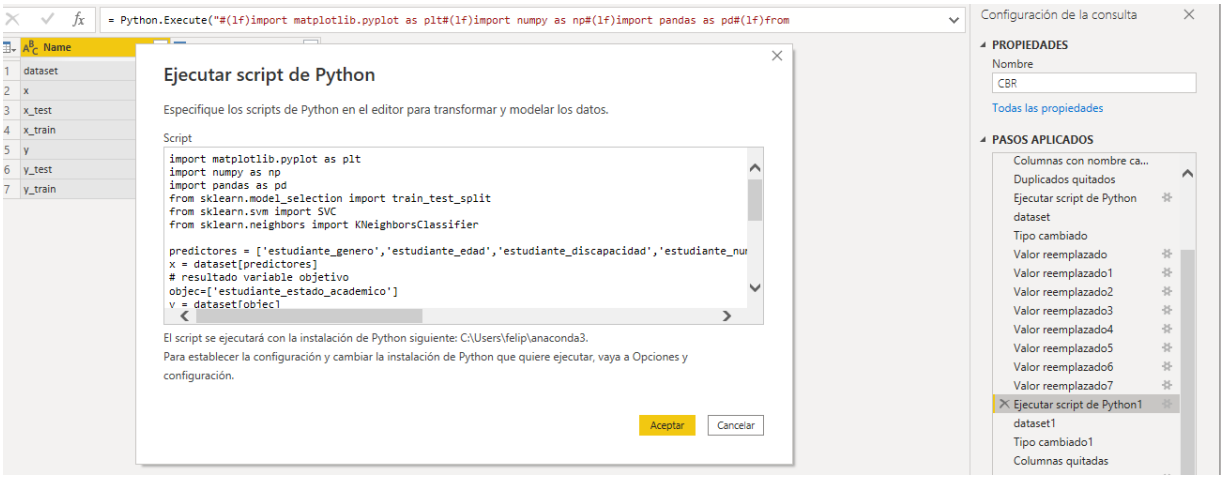

Figura 35. Algoritmo KNN en lenguaje Phyton

Además, permite exportar los resultados del algoritmo ejecutado en columnas calculadas (véase Figura 36) de fácil implementación y visualización a la hora de construir el prototipo.

Prediccion
MEDIO
BAJO
MEDIO
BAJO
BAJO
BAJO
MEDIO
BAJO
BAJO
ALTO
MEDIO
MEDIO
MEDIO
ALTO
MEDIO

Figura 36. Columna calculada con los resultados del algoritmo de predicción

Como se puede apreciar, la predicción asigna un nivel de probabilidad de deserción al estudiante entre Bajo, Medio y Alto; de acuerdo a la similitud con las variables definidas como "predictores" en estudiantes con calidad de desertores, graduados y egresados. De tal manera, si un registro presenta similitud con sus K vecinos más cercanos los cuales son desertores, a este registro se le asignara un nivel de probabilidad de deserción Alto. Si de lo contrario, el registro presenta similitud con sus K vecinos más cercanos los cuales son graduados, se le asignara un nivel de probabilidad de deserción Bajo.

El algoritmo de predicción como se planteó, genera un importante soporte a la toma de decisiones, debido a la dificultad que presenta el departamento de permanencia académica al identificar o perfilar estudiantes desertores de manera convencional o análoga. Lo anterior en conjunto con los reportes estadísticos presentados brindan una visión de la problemática más amplia de lo que se venía trabajando anteriormente.

3.5. Conclusiones

La herramienta POWER BI tiene una alta curva de aprendizaje, porque nos permite crear medidas, columnas calculadas, transformación de datos en otros, de una manera fácil y sencilla porque a pesar de no tener mucha familiaridad con esta herramienta de inteligencia de negocios, se cumplió con los objetivos sin mayor problema.

Puede presentarse el caso, de no saber si se está implementando una bodega de datos o un DM como se dio en el presente trabajo, debido a que estos dos términos tienen gran similitud, aunque esto es solo cuestión de terminología, es recomendable llamarlo por el término que más se asemeje, en este caso se decide llamar DM al sitio donde se alojan los datos debido a su alcance.

Capítulo 4: ARQUITECTURA DEL SISTEMA

La importancia de conocer a profundidad los distintos aspectos que pueden afectar un modelo de negocio parte del hecho la necesidad de integrar las diferentes áreas y fuentes de información de una organización. La centralización de la información en un proceso de inteligencia de negocio no solo facilita el acceso, sino también el análisis histórico y estratégico de la gran cantidad de datos generados continuamente por los procesos de una empresa.

Por tal motivo se plantea implementar una inteligencia de negocio para el área de

permanencia académica con el fin de integrar información de distintas áreas las cuales estén relacionadas directa o indirectamente con la deserción universitaria de tal modo predecir los posibles perfiles estudiantiles con riesgo a abandonar sus estudios.

En este orden de ideas se establece una arquitectura que facilite la puesta en marcha de lo anteriormente mencionada, la cual está compuesta por diferentes módulos que se ejecutarán en el orden establecido Figura 37: Diagrama extendido de la arquitectura de un SIN en el área de permanencia académica de Unicomfacauca.

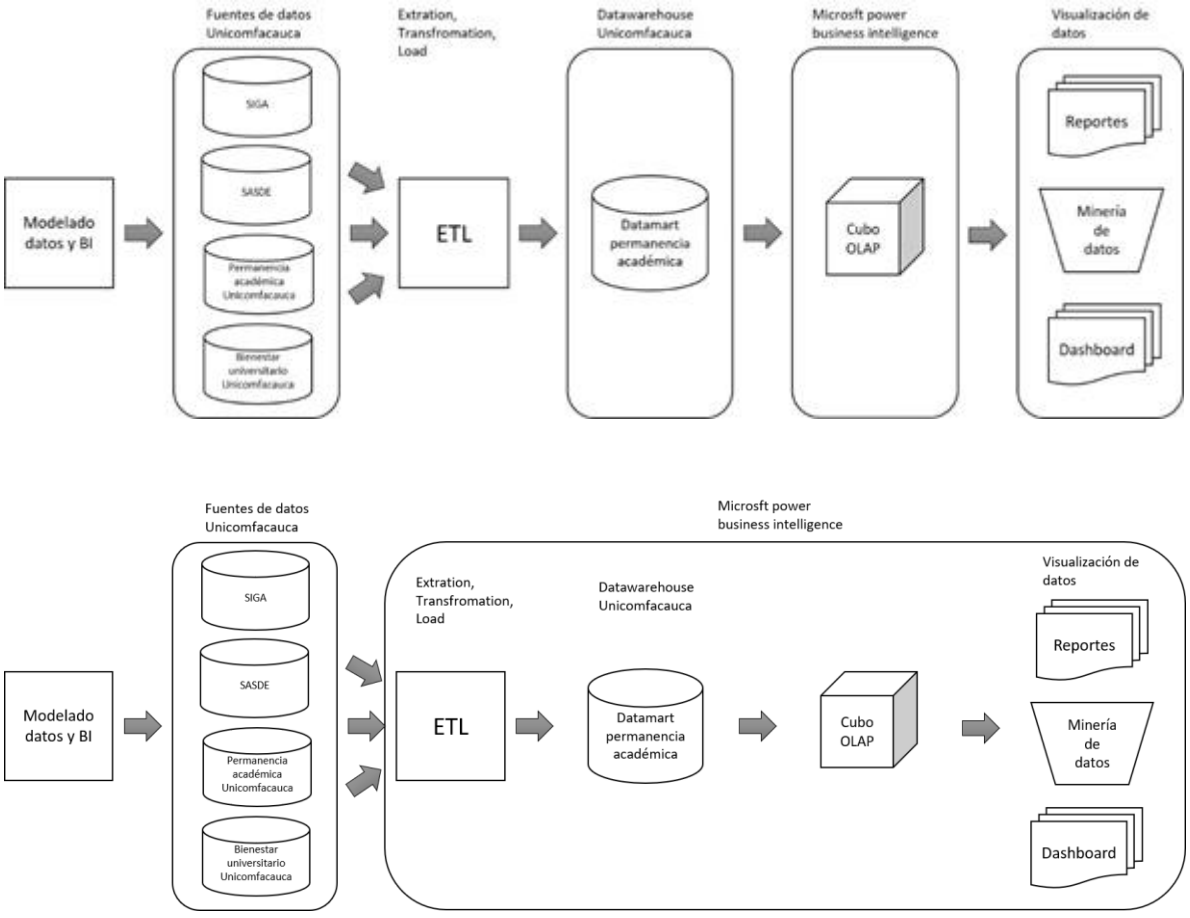

Figura 37: Diagrama extendido de la arquitectura de un SIN en el área de permanencia académica de Unicomfacauca

Como se ve en Figura 37: Diagrama extendido de la arquitectura de un SIN en el área de

permanencia académica de Unicomfacauca, la primera etapa es relacionada con las fuentes de datos, donde se establecen cuáles serán las distintas bases de datos, archivos o fuentes de información pertinentes para la implementación de la inteligencia de negocio, una vez se tiene las fuentes de datos identificadas se procede a realizar el denominado ETL (Extracción, Transformación, Carga). En este proceso se obtiene, limpia y organiza los datos que son más relevantes en la idea de negocio; en el cual se descartan los datos que no se necesitan logrando reducir el almacenamiento necesario.

Teniendo los datos listos se podrá integrarlos en un solo lugar (centralizar) conocido *DM* el cual almacenará lo referente a factores que inducen a la deserción de un estudiante, posteriormente con la ayuda de la plataforma especializada en SIN el cual en este caso será *Microsoft Power Business Intelligence* o *Power BI* se realizará el análisis OLAP (Proceso analítico en línea por sus siglas en inglés) de tal forma que sea más fácil establecer el análisis o generar los reportes que colaboren a la toma eficiente de decisiones en el área de permanencia académica de la corporación.

Culminando el proceso de la inteligencia de negocio se podrá generar los respectivos reportes, procesos de minería de datos que permitirán identificar patrones y tendencias utilizados por el área de permanencia académica en la detección temprana de estudiantes con posibilidad de desertar

4.1. Diseño centrado en el usuario - DCU

Teniendo en cuenta la importancia de entre la interfaz y el usuario dentro de cualquier sistema de información, se tuvo en cuenta que el diseño de la interfaz del *dashboard* cumpliera con los lineamientos básicos de una metodología DCU como lo son:

- Reparto apropiado de las funciones entre usuario y funciones
- Participación activa de los usuarios
- Iteraciones en las soluciones de diseño
- Equipos de diseño multidisciplinares

De esta manera ofrecer una mejor experiencia al usuario a la hora de interactuar con el sistema, no solo desde el punto de vista funcional en donde sea fácil acceder y consultar la información si no también desde el punto de vista grafico permitiendo que las diferentes graficas del sistema sean fáciles de comprender y analizar para personas que no está familiarizadas con graficas estadísticas.

Todo lo anterior se logra en gran medida a la herramienta de inteligencia de negocio en la cual se implementó el sistema (Power BI), debido a que esta

predeterminadamente ofrece funciones que cumplen con los principios del DCU, como la personalización de interfaz, la interactividad y facilidad en su uso.

No obstante, se consulto acerca del tipo de gráfica y colores óptimos a la hora de presentar los informes a la corporación, por lo cual, se decidió utilizar colores institucionales y graficas de tipo pastel, barras e histogramas por recomendación de un diseñador gráfico.

4.2. Construcción del prototipo

La construcción del prototipo fue realizada enteramente con PowerBI, ya que esta herramienta ofrece una interfaz gráfica amigable, dinámica y de rápido aprendizaje con un gran nivel de personalización. Pensando en personas que tengan muy poco o ningún conocimiento de este tipo de herramientas tecnológicas, se presentó el *dashboard* mediante el servicio de uso embebido en la nube que ofrece PowerBI, el cual genera un enlace web que permite interactuar con el sistema con facilidad.

Dentro del dashboard se generaron pestañas de introducción al sistema, los cuales pretenden servir de guía al usuario dentro del sistema, indicando brevemente el contenido de sus diferentes pestañas (véase Figura 38).

Figura 38. Pantalla de resumen del sistema

Además de esto se presentó la información de manera organizada en categorías,

las cuales van desde reportes de deserción por variables personales hasta reportes según el número de reprobados según programa, facultad, semestre y materia entre otras. Los diferentes informes se pueden filtrar según se seleccione alguna variable de interés presente en los gráficos, además de filtrarse según el periodo académico (véase Figura 39).

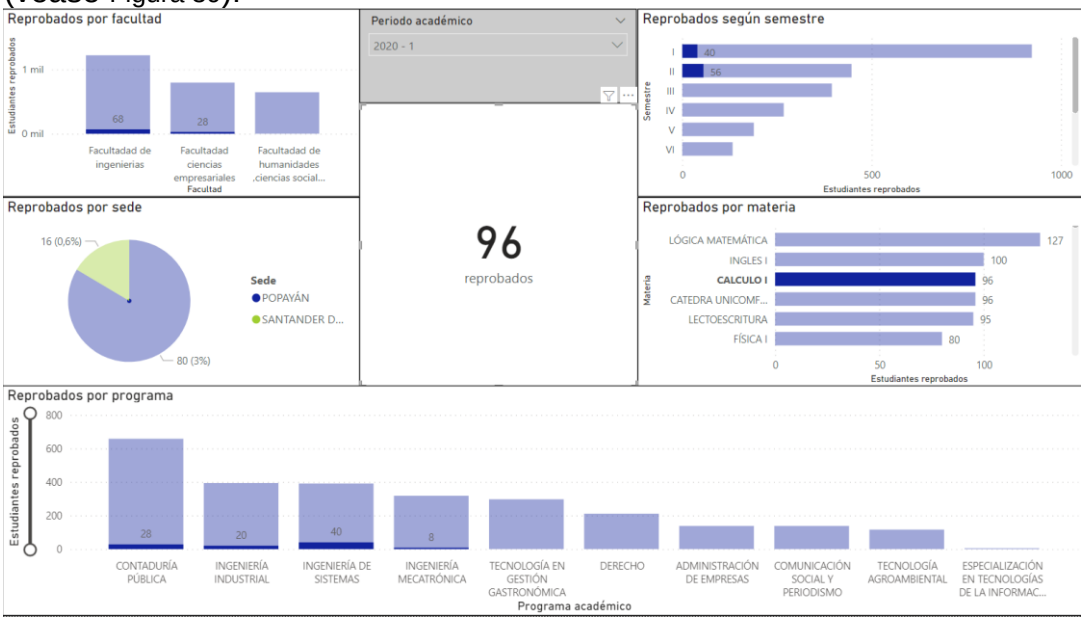

Figura 39. Informes reprobados Calculo I durante el periodo 2020 - 1

Dentro del *dashboard* se generó un breve reporte acerca de la probabilidad de deserción de los estudiantes según su estrato económico (véase Figura 40), el cual fue generado mediante la predicción del algoritmo KNN tratado anteriormente en este documento.

La implementación del prototipo de inteligencia de negocio propuesta para el área de permanencia académica se llevó a cabo tomando en cuenta las 4 fases que componen la definición de implementar, en las que están, diseñar, en donde se planteó un modelo dimensional, una caracterización de variables y un sistema que respondiera con las necesidades presentadas; se construyó una bodega de datos, un protocolo de limpieza y un prototipo de sistema de inteligencia de negocio en *PowerBI* el cual permite visualizar e interactuar con la información generada mediante el diseño; se evaluó el prototipo del sistema finalizado frente a los usuarios finales a los cuales va destinado el producto, constatando la potencial utilidad presentada por el sistema. Además de esto, cabe señalar que durante el proceso de la implementación se siguió la metodología iterativa y creciente [68] con fin de generar un diseño el cual cumpliera con los estándares planteados para el sistema. Recapitulando, se implementó la solución planteada hasta el punto de despliegue

de un prototipo en una plataforma accesible para los interesados, la cual es entregada junto con un protocolo de tratamiento de datos a los administrativos responsables de su integración dentro de la Corporación, de tal manera que dichos responsables puedan tomar como guía el protocolo de tratamiento de datos para la implementación eficaz del *datamart* planteado.

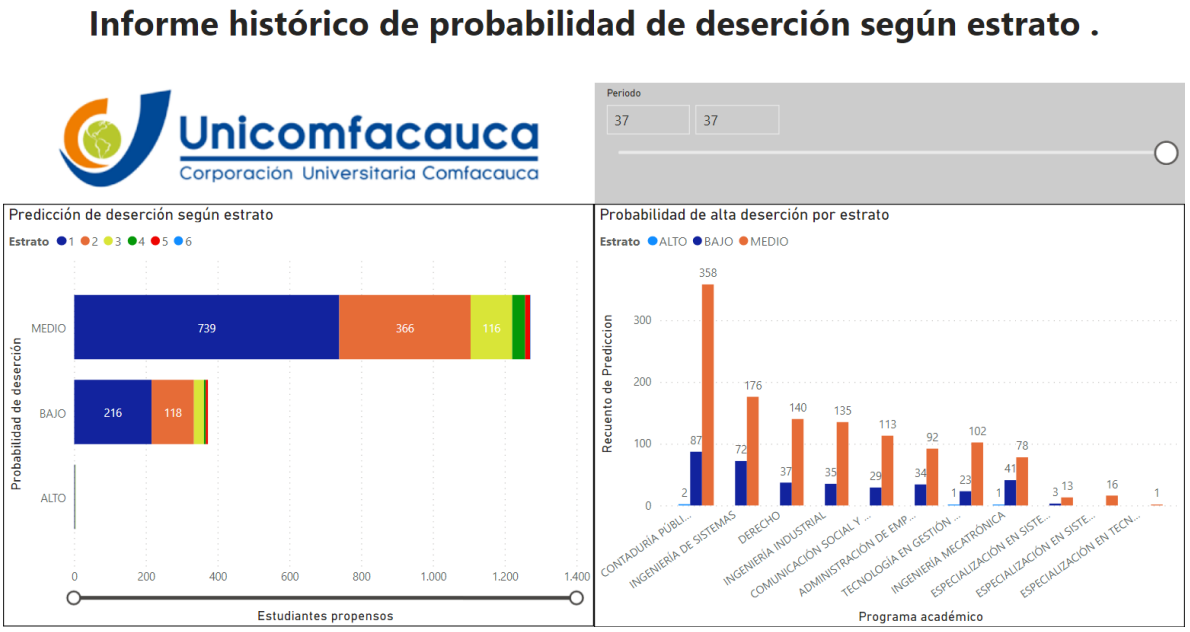

Figura 40. Predicción de deserción según estrato socioeconómico

La implementación de este prototipo permitió responder a las preguntas formuladas desde el área de permanencia académica, con las cuales se dieron origen a los requerimientos del sistema formulados en este documento; para obtener más detalles revisar el (**Anexo A**) en el documento de anexos presentado.

Para finalizar, cabe aclarar que los reportes y variables relacionadas en este prototipo pueden ser modificadas por los mismos usuarios según sus necesidades.

4.3. Evaluación de la experiencia de usuario del sistema

La evaluación del sistema se realizó mediante la metodología SUS, en la cual se propuso preguntas a medida que el usuario experimentaba con el *dashboard* (véase **Anexo C**), de acuerdo con las preguntas realizadas se categorizaron en 5 secciones presentadas en la Tabla 16. Medición de escala de usabilidad (SUS) , posteriormente se realizó la escala de medición arrojando los porcentajes expuestos a continuación.

Tabla 16. Medición de escala de usabilidad (SUS)

Cargo administrativo del participante	Uso intuitivo del sistema	Facilidad de comprensión del sistema	Eficiencia del sistema	Efectividad del sistema	Idoneidad del sistema
Gestor de permanencia académica Unicomfacauca	4	4	4	5	4
Coordinador de sistemas de información Unicomfacauca	5	4	5	5	4
Directora del programa de ingeniería de sistemas Unicomfacauca	5	5	5	5	5
Porcentaje	93.3%	86.6%	93.3%	100%	86.6%
TOTAL					91.9%

De acuerdo con lo evaluado durante la reunión con los administrativos de Unicomfacauca, se pudo establecer mediante las preguntas formuladas (véase **Anexo C**) la efectividad con que cuenta el sistema, en la cual se constató que el prototipo propuesto ofrece un menor tiempo de respuesta frente a sistemas transaccionales y, por consiguiente, un mayor desempeño en la visualización de la información requerida en el área de permanencia.

Además de lo anterior, se exponen las ventajas que brinda la implementación de una bodega de datos resaltando el análisis de información relacionada a temas en concretos, en este caso la deserción académica, y se evidencia la amplia superioridad a la ofrecida actualmente por los sistemas de la corporación.

La factibilidad del proyecto se evidencio en la medición de la satisfacción de los

usuarios finales mediante experiencia de usuario por medio de un focus group, donde se tomó en cuenta tanto la factibilidad operacional como la técnica (no se tomó en cuenta la factibilidad económica debido a la naturaleza del proyecto y a la actual disponibilidad de todas las herramientas necesarias en la corporación). Como lo dice [69] la factibilidad operacional cuenta con 4 aspectos a considerar; primero, la complejidad de un sistema, referente a su manejo por el usuario; segundo, resistencia a la aceptación del nuevo sistema debido al modo de trabajo del usuario, interés en el sistema antiguo u otras razones; tercero, resistencia a la aceptación debido a la poca adaptación al cambio por parte del usuario, causado por cambios repentinos del sistema; como último, la probabilidad de cambios posteriores a la implementación de sistema que afecten su factibilidad.

En [69]" La factibilidad técnica evalúa si el equipo, software están disponibles o si es posible desarrollarse y si los responsables tienen las capacidades técnicas requeridas para su implementación". Entendiendo lo anterior, la factibilidad del prototipo se logró medir teniendo en cuenta ambos aspectos, la factibilidad operacional, realizando preguntas sobre la facilidad de adaptación en el primer uso del prototipo por parte del usuario, nivel entendimiento de la visualización presentada, nivel de agrado de la interfaz gráfica del prototipo y pertinencia de la información generada, entre otras. La factibilidad técnica de igual manera se puede medir con las preguntas mencionadas, debido a que se pudo desarrollar el prototipo del sistema de manera oportuna y pertinente, visualizando información relevante con la que actualmente no cuenta el área de permanencia académica.

Capítulo 5: CONCLUSIONES Y TRABAJOS FUTUROS

5.1. Conclusiones

- Para una implementación de un sistema de inteligencia de negocios es de suma importancia contar con un conjunto de datos específicos para poder aplicar su respectivo proceso analítico, sino se cuenta con los datos suficientes, se puede implementar un algoritmo adicional como el KNN para encontrar datos faltantes, esto conlleva a que el proyecto tome más tiempo para su desarrollo e implementación, en caso de que los datos faltantes sea mayor a un 10% de su total, el proyecto de inteligencia de negocios no sería viable.

- Una vez terminado este trabajo, podemos llegar a la conclusión que es factible la implementación de un SIN que identifique estudiantes propensos a desertar, ya que la universidad cuenta con una fuente de datos (SIGA) los cuales se pueden tratar para aplicar un debido proceso analítico haciendo de este proyecto viable y de gran impacto para la corporación. Este proceso de análisis

a los datos conllevo a generar información tal como indicadores de deserción desde distintos factores como el individual, académico, económico, entre otros, que da soporte a los encargados de permanencia académica y otras dependencias para la toma de decisiones. Adicional a esto, los resultados arrojados tras aplicar el SUS indican el grado de satisfacción de los encargados de permanencia académica respecto al SIN. Cabe señalar que la universidad no cuenta con un sistema de inteligencia de negocios para la toma de decisiones, se cuenta es con sistemas cuantitativos, que deben hacer análisis de forma manual, pero este SIN desarrollado es automático y tiene historial en el tiempo.

- La universidad CORPORACION UNIVERSITARIA UNICOMFACAUCA debería implementar estrategias que permitan llevar un mejor control de la información, porque los datos proporcionados para el desarrollo del SIN eran incompletos y no actualizados, esto trajo como consecuencia dejar de lado determinadas variables de deserción, a su vez produjo un esfuerzo adicional de los encargados del desarrollo del sistema para poder obtener información completa de las variables utilizadas.

- El modelo dimensional (modelo estrella, modelo copo de nieve, entre otros) implementado para el desarrollo del DW, puede estar sujetos a cambios, debido a que en el transcurso del tiempo se detectan pequeños errores, falencias, se identifica la falta de datos y/o también por la relación de las dimensiones no se puede sacar de manera correcta una medida, o al menos eso paso con frecuencia en la realización de este proyecto, quien el modelo dimensional sufrió una serie de cambios a medida que se avanzaba en el desarrollo del SIN.

5.2. Protocolo de tratamiento de datos

Para el desarrollo del SIN, específicamente en el proceso de limpieza de los datos, se realizó un extenso trabajo con fin de transformar los datos faltantes e inconsistentes en datos limpios que puedan ser usados para el proceso analítico propuesto en este documento, debido a lo anterior mencionado, se hace necesario implementar el uso de buenas prácticas por parte de la Corporación Universitaria Comfacauca - Unicomfacauca que permitan un registro de datos más completos y consistentes, para ello, se sugieren prácticas como:
Validación de datos personales de los estudiantes, establecer como obligatorio el diligenciamiento de datos relevantes como cantidad aproximada del ingreso familiar, ya que muchos de los datos relacionados a esta variable se encontraban vacíos o con información no validada.
Establecer en campos de datos personales, donde la información lo permita,

listas desplegables o respuestas predefinidas, así, de tal forma reducir errores digitación, ortografía y significados ambiguos en variables importantes para el análisis de datos.
Priorizar la recolección de datos personales que favorezcan el análisis y generación de información útil para los distintos procesos de la Corporación Universitaria Comfacauca – Unicomfacauca, algunos de los datos personales más destacables de los cuales no se utilizaron por falta calidad en los datos fueron:

- Ingreso económico
- Origen Étnico
- Empleo
- Número Hijos
- Personas a cargo
- Perfil psicológico
- Estudios padres
- Escolaridad previa
- Promedio
- Apoyo económico (becas, créditos, contado)
- Ambiente y convivencia
- Seguimiento /tutorías
- Lugar Procedencia
- Intensidad Horaria
- Vocación
- Situación incómoda (social, familiar, económica)
- Ocupación padre y madre
- Independencia económica
- Adaptación académica
- Motivación
- Hábitos de estudio
- Bases académicas previas
- Colegio procedencia
- Promedio pre U
- Ingreso familiar

I. Recibir datos: se obtienen los datos suministrados por el Coordinador de Sistemas de la Corporación Universitaria Comfacauca - Unicomfacauca, este conjunto de datos como primera medida se almacenan en una base de datos provisional llamada data-staging o data-lake, donde los datos son almacenados tal como vienen de las fuentes de datos para posteriormente realizar su debido tratamiento.

II. Identificar y tratar inconsistencia: se inspeccionan los datos proporcionados con el fin de identificar inconsistencias (mala ortografía, diferentes nombres empleados para una única materia, errores tipo gráficos, entre otras), una vez se establecen que variables son las afectadas, mediante el uso de la herramienta de Power BI se realiza el proceso de transformación del data-lake, aumentando la calidad de los datos de tal manera de facilitar el proceso analítico.

III. Tratamiento de registros vacíos: en esta fase se determina que variables

dentro del dataset son prioritarias (basado en la investigación realizada) y cuales no representan un aporte significativo para la implementación del proyecto, teniendo en cuenta lo anterior, si las variables prioritarias presentan menos del 10% de registros vacíos sobre la totalidad de los datos, se realiza un proceso de llenado mediante el algoritmo CBR basado en distancia euclidiana.

IV.Depuración del dataset: Debido a la inconsistencia de los datos se identificaron registros con gran cantidad de campos nulos, donde las variables presentan más del 10% de registros vacíos sobre la totalidad de los datos, afortunadamente, la mayoría de estas variables no representan un aporte significativo en el proceso analítico , por lo cual, se descarta su utilización con el fin de no afectar los resultados del proceso, debido a que con una mayor cantidad de campos vacíos la confiablidad de la implementación del algoritmo CBR decrece.

V. Generar variables implícitas: durante el desarrollo del SIN se vio la necesidad de utilizar variables no incluidas dentro del data set recibido, pero que a su vez es posibles generarlas a partir de otras variables suministradas, como es el caso de la variable edad, donde este registro se puede obtener a partir de la fecha de nacimiento de la persona. Otro claro ejemplo de este proceso es la conversión del id para el último periodo cursado por un estudiante a una fecha en concreto, gracias al registro de periodos académicos los cuales cuentan con una fecha correspondiente para cada id.

5.3. Trabajos futuros

Como trabajos futuros se podría implementar un algoritmo avanzado de minería de datos que ayude en el proceso de hallar anomalías, patrones y que correlacione los datos que existen en la DW de la universidad corporación universitaria Unicomfacauca, haciendo una comparación de variable de estudiantes desertores con estudiantes nuevos, conllevando a la identificación detalla de estudiantes que tienen un nivel alto y medio de deserción. De esta manera la universidad tendrá conocimientos previos y podrá tomar mejores decisiones con anterioridad.

Por otra parte, el DW desarrollado en este trabajo comprende solo el área de permanencia académica, más adelante otras dependencias pueden de igual forma adoptar estos mecanismo de inteligencia de negocios que los ayuden a obtener información valiosa para el apoyo de toma de decisiones, una vez estas dependencias implementen sus DW se puede obtener como resultado final la

gran BODEGA DE DATOS de donde se puede suministrar información para toda la universidad Corporación Universitaria Comfacauca - Unicomfacauca.

BIBLIOGRAFÍA

[1] C. N. C. Parra, M. A. G. Duarte, J. S. M. Rueda, J. T. O. Bernal, and L. Y. C. Chacón, "Sistema de información para la generación de alertas tempranas de estudiantes con riesgo de deserción," *Rev. Matices Tecnológicos*, vol. 10, no. 0, pp. 38–46, 2018, [Online]. Available: http://publicaciones.unisangil.edu.co/index.php/revista-matices-tecnologicos/article/view/408.

[2] O. SIERRA, Helvy HERNÁNDEZ, "SISTEMA DE ALERTAS TEMPRANAS COMO HERRAMIENTA DE INNOVACIÓN TECNOLÓGICA EN LA UNIVERSIDAD SANTO TOMÁS PARA EL FORTALECIMIENTO DE LA PERMANENCIA ESTUDIANTIL Y GRADUACIÓN OPORTUNA," 2014.

[3] L. E. Durán Rafael, "Implementacion De Un Data Mart Para El Seguimiento Académico De Los Estudiantes En La Escuela Académico Profesional De Ingeniería De Sistemas De La Universidad Nacional De Cajamarca," p. 148, 2017.

[4] J. Santos and M. Benites, "Business intelligence and its impact on university management of the Faculty of Engineering of the National University of Trujillo," *Rev. Cienc. y Tecnol.*, vol. 16, no. 3, pp. 87–104, 2020, doi: 10.17268/rev.cyt.2020.03.09.

[5] V. Duro Novoa and C. M. Pérez Cuevas, "Inteligencia De Negocios Y Sistema De Soporte a Las Decisiones De La Gestión Económica En La Universidad De La Habana," *3C TIC Cuad. Desarro. Apl. a las TIC*, vol. 5, no. 4, pp. 38–54, 2016, doi: 10.17993/3ctic.2016.54.38-54.

[6] G. Pascal, E. Grillo, D. Servetto, and A. Redchuk, "Sistema de Apoyo a las Decisiones (DDS) para la productividad de las universidades: implementación de tableros de control," *XIX Work. Investig. en Ciencias la Comput.*, pp. 349–353, 2017.

[7] J. Castillo, A. González, and L. Muñoz, "Inteligencia de Negocios como apoyo a sistemas de información de egresados de instituciones de educación superior," *AmITIC*, pp. 81–88, 2018.

[8] H. SUYUTI, "No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title," pp. 5–10, 2019.

[9] C. A. Cardoza Timana, "Elaboración De Un Data Mart Para Evidenciar El Retraso Académico En Los Alumnos De Pregrado De La Fii-Unp," pp. 1–132,

2015, [Online]. Available: http://repositorio.unp.edu.pe/handle/UNP/645.

[10] A. M. Arenas Lopez, Maria Camila ; Gomez Montes, “Inteligencia de negocios aplicada a los procesos... - Google Académico,” 2017. https://scholar.google.es/scholar?hl=es&as_sdt=0%2C5&q=Inteligencia+de+negocios+aplicada+a+los+procesos+de+autoevaluación+de+la+Universidad+de+Manizales&btnG= (accessed Nov. 20, 2020).

[11] J. J. Avila Espejo, “Universidad Peruana de Ciencias Aplicadas,” *Univ. Peru. Ciencias Apl.*, p. 2020, 2016, Accessed: Jun. 18, 2021. [Online]. Available: https://repositorioacademico.upc.edu.pe/handle/10757/621002.

[12] U. Nacional and J. M. Arguedas, “UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS Implementación de un Sistema de Soporte de Decisiones para la Gestión Académica de la Universidad Nacional José María Arguedas Jeans Diego Ramos Peñaloza,” 2015.

[13] N. Tang *et al.*, “Comparación de herramientas□□□□□□,” بيبيب, vol. ث ققثق, no. ثق ثقثقثق. p. 2018, ثقثقثقثق.

[14] I. Ruiz-Quintero, “INTELIGENCIA DE NEGOCIOS AL PROCESO DE LA EVALUACION DOCENTE,” *Vestig. Ire*, vol. 8, no. 1, pp. 206–214, 2015.

[15] Q. Fernando Medina, M. Francisco Fariña, and W. Castillo-Rojas, “Data mart to obtain indicators of academic productivity in a university,” *Ingeniare*, vol. 26, pp. 88–101, 2018, doi: 10.4067/S0718-33052018000500088.

[16] Y. Reyes, D. Y. Cu, and L. N. Maturel, “La inteligencia de negocio como apoyo a la toma de decisiones en el ámbito académico,” *L*, vol. 3, no. 2, pp. 2013–2014, 2015.

[17] L. Asto Huamán and M. R. Arangüena Yllanes, “INTELIGENCIA DE NEGOCIOS EN LA GESTIÓN ACADÉMICA DE LA EDUCACIÓN SUPERIOR UNIVERSITARIA,” *Rev. Investig.*, vol. 7, no. 2, pp. 526–536, May 2018, doi: 10.26788/riepg.2018.2.77.

[18] O. Cerrón and J. José, “Modelos estocásticos e inteligencia de negocios en la Oficina de Admisión de la Universidad Nacional José María Arguedas,” 2017.

[19] N. A. H. M. Rodzi, M. S. Othman, and L. M. Yusuf, “Significance of data integration and ETL in business intelligence framework for higher education,” in *2015 International Conference on Science in Information Technology (ICSITech)*, Oct. 2015, pp. 181–186, doi: 10.1109/ICSITech.2015.7407800.

[20] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, “Analyzing undergraduate students’ performance using educational data mining,” *Comput. Educ.*, vol. 113, pp. 177–194, Oct. 2017, doi: 10.1016/j.compedu.2017.05.007.

[21] S. Dong and M. S. Lucas, “An Analysis of Disability, Academic Performance, and Seeking Support in One University Setting,” *Career Dev. Transit. Except. Individ.*, vol. 39, no. 1, pp. 47–56, Feb. 2016, doi: 10.1177/2165143413475658.

[22] D. Kabakchieva, “Business Intelligence Systems for Analyzing University Students Data,” *Cybern. Inf. Technol.*, vol. 15, no. 1, pp. 104–115, Mar. 2015, doi: 10.1515/cait-2015-0009.

[23] K. A. Fakeeh, “Decision Support Systems (Dss) in Higher Education System,” *Int. J. Appl. Inf. Syst.*, 2015.

[24] V. K. Ong, “Business Intelligence and Big Data Analytics for Higher Education: Cases from UK Higher Education Institutions,” 2016. Accessed: Jun. 19, 2021. [Online]. Available: http://www.iaiai.org/journals/index.php/IEE/article/view/63.

[25] Y. M. Pérez-Pérez, A. A. Rosado-Gómez, and A. M. Puentes-Velásquez, “Application of business intelligence in the quality management of higher education institutions,” *J. Phys. Conf. Ser.*, vol. 1126, p. 012053, Nov. 2018, doi: 10.1088/1742-6596/1126/1/012053.

[26] K. Sin and L. Muthu, “"APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS – A LITERATURE REVIEW ",” *ICTACT J. Soft Comput.*, vol. 05, no. 04, pp. 1035–1049, Jul. 2015, doi: 10.21917/ijsc.2015.0145.

[27] H. Arturo Combita Niño, J. Patricia Cómbita Niño, and R. Morales Ortega, “Business intelligence governance framework in a university: universidad de la costa case study,” 2018. Accessed: Jun. 19, 2021. [Online]. Available: http://repositorio.cuc.edu.co/handle/11323/2256.

[28] V. Khatibi, A. Keramati, F. S.-S. S. & H. Open, and undefined 2020, “Deployment of a business intelligence model to evaluate Iranian national higher education,” *Elsevier*, Accessed: Jun. 19, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590291120300450.

[29] E. Ojeda, M. Santiago, and G. Gomez, “ANÁLISIS DE LA DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD DE CUNDINAMARCA SEDE FUSAGASUGÁ UTILIZANDO HERRAMIENTAS DE INTELIGENCIA DE NEGOCIOS (Caso de Uso-Muestra de Datos suministrado: Programa de Ingeniería de Sistemas Pensum (Ingeniería de Sistemas 2013) ,” 2019. Accessed: Nov. 19, 2020. [Online]. Available: http://repositorio.ucundinamarca.edu.co/handle/20.500.12558/2239.

[30] J. S. Caicedo Chacón, Luz Yamile; Cárdenas Parra, Cristian Noé; Müller Rueda and J. T. Ortiz Bernal, “Aplicación para la gestión y el análisis de informaci... - Google Académico,” 2019. https://scholar.google.es/scholar?hl=es&as_sdt=0%2C5&q=Aplicación+para+la+gestión+y+el+análisis+de+información+relacionada+con+la+deserción+estudiantil+universitaria&btnG= (accessed Nov. 19, 2020).

[31] R. ; Timaran and Jimenez, *Detección de patrones de deserción estudiantil en programas de pregrado de instituciones de educación superior con CRISP-DM*. .

[32] C. Guzmán Ruiz *et al.*, *Educación Superior Colombiana*. 2009.

[33] M. Heredia Alarcón *et al.*, “Deserción estudiantil en las carreras de ciencias de la salud en el Perú,” *An. la Fac. Med.*, vol. 76, no. SPE, p. 57, Feb. 2015, doi: 10.15381/anales.v76i1.10972.

[34] A. B. Urbina-Nájera, J. C. Camino-Hampshire, and R. Cruz Barbosa, “University dropout: Prevention patterns through the application of educational data mining,” *Reli. - Rev. Electron. Investig. y Eval. Educ.*, vol. 26, no. 1, pp. 1–19, Oct. 2020, doi: 10.7203/relieve.26.1.16061.

[35] “Vista de DESERCIÓN UNIVERSITARIA EN ESTUDIANTES DE UNA UNIVERSIDAD PRIVADA DE IQUITOS.” https://revistas.upc.edu.pe/index.php/docencia/article/view/42/11 (accessed Jun. 18, 2021).

[36] R. A. Pazmiño Maji, C. E. Solis Benavides, F. J. García Peñalvo, and M. Á. Conde González, “investigación de pregrado en la Escuela Superior Politécnica de Chimborazo: Mapeo Sistemático y Analíticas.,” *Rev. Científica Ecociencia*, vol. 6, no. 1, pp. 1–25, 2019, doi: 10.21855/ecociencia.61.183.

[37] C. Zarria Torres, C. Arce Ramos, and J. Lam Moraga, “Estudio de variables que influyen en la deserción de estudiantes universitarios de primer año, mediante minería de datos,” *Cienc. Amaz.*, vol. 6, no. 1, p. 73, Jun. 2016, doi: 10.22386/ca.v6i1.110.

[38] E. Sergio and F. Angulo, “Modelo para la automatización del proceso de determinación de riesgo de deserción en alumnos universitarios,” Universidad de Chile, 2012. Accessed: Jun. 18, 2021. [Online]. Available: http://repositorio.uchile.cl/handle/2250/111188.

[39] C. Choque, “Universidad nacional de moquegua,” no. 052, pp. 1–18, 2015.

[40] L. Canchila and J. Sánchez, “Análisis de la Deserción Estudiantil de Cecar Utilizando Herramientas de Inteligencia de Negocios con Licencia Libre,” pp. 1–125, 2016.

[41] K. B. Eckert and R. Suénaga, “Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos,” *Form. Univ.*, vol. 8, no. 5, pp. 3–12, 2015, doi: 10.4067/S0718-50062015000500002.

[42] “Analysis of Student Desertion in a Systems and Computing Engineering Undergraduate Program - Dialnet.” https://dialnet.unirioja.es/servlet/articulo?codigo=6939723 (accessed Jun. 18, 2021).

[43] J. Argote, I., Jimenez, R. y Gómez, “Detección de patrones de deserción en los programas de pregrado de la Universidad Mariana de San Juan de Pasto, aplicando el proceso de descubrimiento de conocimiento sobre base de datos (KDD) y su implementación en modelos matemáticos de predicción,” *Cuarta Conf. Latinoam. sobre el Abandon. en la Educ. Super.*, pp. 1–7, 2014.

[44] C. Gonzales Cam and C. Rodriguez Dominguez, “Propuesta de un Modelo de

Business Intelligence para Identificar el perfil de deserción estudiantil en la Universidad Científica del Sur," *Univ. Peru. Ciencias Apl.*, p. 147, 2017, [Online]. Available: https://repositorioacademico.upc.edu.pe/handle/10757/622749.

[45] M. Castillo-Sánchez, R. Gamboa-Araya, and R. Hidalgo-Mora, "Factors that influence student dropout and failing grades in a university mathematics course," *Uniciencia*, vol. 34, no. 1, pp. 219–245, Jan. 2020, doi: 10.15359/ru.34-1.13.

[46] "Deserción universitaria en Colombia | Academia y Virtualidad." https://revistas.unimilitar.edu.co/index.php/ravi/article/view/5461 (accessed Jun. 18, 2021).

[47] G. B. Astudillo, "Modelo BI para Adaptar la Oferta de Programas de Apoyo Estudiantil a Perfiles de Riesgo de Deserción en la Universidad Masiva," 2016. Accessed: Jun. 18, 2021. [Online]. Available: https://repositorio.usm.cl/handle/11673/23122.

[48] E. Ojeda, M. Santiago, and G. Gomez, "ANÁLISIS DE LA DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD DE CUNDINAMARCA SEDE FUSAGASUGÁ UTILIZANDO HERRAMIENTAS DE INTELIGENCIA DE NEGOCIOS (Caso de Uso-Muestra de Datos suministrado: Programa de Ingeniería de Sistemas Pensum (Ingeniería de Sistemas 2013) Períodos (IPA 2013-IPA 2018))," 2019.

[49] "Revista ESPACIOS | Vol. 41 (Nº 06) Año 2020." http://es.revistaespacios.com/a20v41n06/20410615.html (accessed Jun. 18, 2021).

[50] G. R. Rivadera, "La metodología de Kimball para el diseño de almacenes de datos (Data warehouses)."

[51] M. E. Mendoza, L. D. Meneses, and N. Rivera Ortíz, "MBD 1.0-MetoDología De Desarrollo De BoDegas De Datos Para Micro, Pequeñas y MeDianas eMPresas," 2010. Accessed: Nov. 23, 2020. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=6299689.

[52] K. S. Pratt, "Design Patterns for Research Methods: Iterative Field Research," *AAAI Spring Symp. Exp. Des. Real*, no. 1994, pp. 1–7, 2009.

[53] A. Mendez and P. Martínez., Garcia, R. Mártire, A. Britos, "Fundamentos de Data Warehouse," *Reportes Técnicos en Ing. del Softw.*, vol. 5, no. 1, pp. 19–26, 2003, [Online]. Available: http://artemisa.unicauca.edu.co/~ecaldon/docs/bd/fundamentosdedatawarehouse.pdf.

[54] sinnexus, "¿Qué es Business Intelligence?," 2018.

[55] "Diferencia entre Data Mart y Data warehouse. ¡Descúbrela!" https://revistadigital.inesem.es/informatica-y-tics/diferencia-entre-data-mart-y-data-warehouse/ (accessed Jun. 19, 2021).

[56] "Data Warehouse y Data Marts - Conoce las diferencias." https://blogs.solidq.com/es/business-analytics/data-warehouse-y-data-marts-esquema-en-estrella-11/ (accessed Jun. 19, 2021).

[57] "Los gestores de bases de datos (SGBD) más usados." https://revistadigital.inesem.es/informatica-y-tics/los-gestores-de-bases-de-datos-mas-usados/ (accessed Jun. 19, 2021).

[58] "Gartner Magic Quadrant for Operational Database Management Systems." https://www.gartner.com/en/documents/3975492/magic-quadrant-for-operational-database-management-syste (accessed Jun. 19, 2021).

[59] M. A. Romero and J. R. García, "COMPARISON OF OPTIONS FOR INTELLIGENCE BUSINESS IN MAJOR SYSTEMS MANAGERS MARKET DATABASES," *lamjol.info*, 2016, Accessed: Jun. 19, 2021. [Online]. Available: https://www.lamjol.info/index.php/EyA/article/view/4289.

[60] "Write a Review About an IT Solution Reviews 2021 | Gartner Peer Insights." https://www.gartner.com/reviews/market/operational-dbms/vendors (accessed Jun. 19, 2021).

[61] "Introducción a la Minería de Datos - Google Libros." .

[62] H. R. Álvarez and G. Avendaño, "Comparación de las metodologías de análisis discriminante robusto y redes neuronales," *Rev. Ontare*, vol. 2, no. 2, pp. 35–64, 2015.

[63] D. J. Matich, "Redes Neuronales: Conceptos Básicos y Aplicaciones.," *Historia Santiago.*, p. 55, 2001.

[64] S. Orea, A. Vargas, M. A.- Ene, and undefined 2005, "Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos," *academia.edu*, Accessed: Jun. 19, 2021. [Online]. Available: https://www.academia.edu/download/34203825/e1.pdf.

[65] "Arboles de decisiones en la minería de datos - Conecta Software." https://conectasoftware.com/analytics/arboles-de-decisiones-en-la-mineria-de-datos/ (accessed Jun. 19, 2021).

[66] J. E. Sotomonte-Castro, C. C. Rodríguez-Rodríguez, C. E. Montenegro-Marín, P. A. Gaona-García, and J. G. Castellanos, "Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos - Towards the construction of a predictive model of academic desertion based on data mining techniques," *Rev. científica*, vol. 3, no. 26, p. 35, Oct. 2016, doi: 10.14483/23448350.11089.

[67] "El enfoque de Ralph Kimball." https://blog.bi-geek.com/arquitectura-el-enfoque-de-ralph-kimball/ (accessed Jun. 06, 2021).

[68] "Desarrollo iterativo y creciente - Wikipedia, la enciclopedia libre." https://es.wikipedia.org/wiki/Desarrollo_iterativo_y_creciente (accessed Dec. 06, 2021).

[69] "Diseño de Sistemas de Información: Estudio de la Factibilidad." http://sismeca12009.blogspot.com/2009/04/estudio-de-la-factibilidad.html?m=1 (accessed Dec. 06, 2021).

CONTROL DE CAMBIOS

VERSIÓN	FECHA	MOTIVO CAMBIO
1	20 de Febrero de 2020	Creación
2	31 de Octubre del 2021	Finalizado