UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

APROVECHAMIENTO, DESERCIÓN Y REPROBACIÓN DE ESTUDIANTES UNIVERSITARIOS DE TABASCO DESDE LA MINERÍA DE DATOS

TESIS PARA OBTENER EL GRADO DE:

Maestro en Administración de Tecnologías de la Información

PRESENTA:

I.S.C. Luis Manuel Juárez López

BAJO LA DIRECCIÓN DE:

Dra. Martha Patricia Silva Payró

EN CODIRECCIÓN DE**:**

Dr. Rubicel Cruz Romero

Cunduacán, Tabasco, A: agosto de 2024

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE
CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

APROVECHAMIENTO, DESERCIÓN Y REPROBACIÓN DE ESTUDIANTES UNIVERSITARIOS DE TABASCO DESDE LA MINERÍA DE DATOS

TESIS PARA OBTENER DEL GRADO DE:

Maestro en Administración
de Tecnologías de la Información

PRESENTA:
I.S.C. Luis Manuel Juárez López

BAJO LA DIRECCIÓN DE:
Dra. Martha Patricia Silva Payró

EN CODIRECCIÓN DE:
Dr. Rubicel Cruz Romero

JUARADO REVISOR:
Dr. Gerardo Arceo Moheno
Dr. Rafael Mena de la Rosa
Dr. Eric ramos Méndez
Dr. Pablo Payró Campos
Dr. Eddy Arquímedes García Alcocer

Cunduacán, Tabasco, A: agosto de 2024

Declaración de Autoría y Originalidad

En la Ciudad de Cunduacán, el día 1 del mes de agosto del año 2024, el que suscribe **Luis Manuel Juárez López** alumno del Programa de la **Maestría en Administración de Tecnologías de la Información** con número de matrícula **222H19003**, adscrito a la **División Académica de Ciencias y Tecnologías de la Información**, de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención del Grado de maestría y titulada **Aprovechamiento, deserción, y reprobación de estudiantes universitarios de Tabasco desde la minería de datos** dirigida por la Dra. Martha Patricia Silva Payró y el Dr. Rubicel Cruz Romero.

DECLARO QUE: La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando y aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente

Villahermosa, Tabasco a 1 de agosto de 2024.

Alumno: I.S.C. Luis Manuel Juárez López

Cunduacán, Tabasco a 10 de julio de 2024

Oficio No. 0817/DACYTI/CP/2024

Asunto: Autorización de impresión de Tesis

C. Luis Manuel Juárez López
Matricula: 222H19003

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional ***"Aprovechamiento, deserción y reprobación de estudiantes universitarios de Tabasco desde la minería de datos"***, para presentar examen y obtener el Grado de Maestro en Administración de Tecnologías de la Información.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

MTE. Óscar Alberto González González
Director

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

C.c.p. Dr. Eddy Arquímedes García Alcocer. - Encargado del Despacho de la Coordinación de Posgrado DACYTI
Archivo.
Consecutivo.
M.T.E. OAGG/EAGA

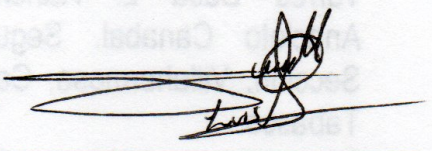
Carta de Cesión de Derechos

Villahermosa, Tabasco a 1 de agosto de 2024.

Por medio de la presente manifestamos haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Aprovechamiento, deserción y reprobación de estudiantes universitarios de Tabasco desde la minería de datos**.

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención, de que tenemos el derecho al reconocimiento como autores de la obra, y la Universidad Juárez Autónoma de Tabasco mantendrá en un 100% la titularidad de los derechos patrimoniales por un período de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADORES



Alumno: I.S.C. Luis Manuel Juárez López

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

F8: Cesión de Derechos

Cunduacán, Tabasco, a 08 de julio de 2024.

A quien corresponda:

Los que suscriben la presente, declaramos que el proyecto de obtención de grado denominado, **"Aprovechamiento, deserción y reprobación de estudiantes universitarios de Tabasco desde la minería de datos"** es de nuestra autoría intelectual y por lo tanto cedemos todos los **derechos** sobre este proyecto a la Universidad Juárez Autónoma de Tabasco, a la cual relevamos de cualquier sanción y asumimos responder a cualquier reclamo de derechos de autor ante las autoridades competentes.

Atentamente

Autores:

Nombre	Domicilio	Firma autógrafa
Luis Manuel Juárez López	Ranchería Cúlico.2da Sección, Cunduacán, Tabasco.	
Martha Patricia Silva Payró	Callejón del Pozo Privada Las Torres Casa 2. Ranchería Anacleto Canabal. Segunda Sección. Villahermosa, Centro Tabasco.	
Rubicel Cruz Romero	San Vicente #4 La Lima Parrilla. Villahermosa, Centro Tabasco.	

c.c.p. MTE. Oscar Alberto González González. - Director de la DACYTI
Dr. Eddy Arquímedes Garcia Alcocer. Encargada del despacho de la Coordinación de Posgrado.

Cunduacán, Tabasco a 17 de junio de 2024
Oficio No. 625/DACYTI/D/2024

Asunto: Dirección de Tesis

Dra. Martha Patricia Silva Payró
Profesora Investigadora

De conformidad con lo establecido en el Reglamento de Estudios de Posgrado Vigente, de la Universidad Juárez Autónoma de Tabasco, me permito informarle, que ha sido designada como Directora de la tesis titulada **"Aprovechamiento, deserción y reprobación de estudiantes universitarios de Tabasco desde la minería de datos"**, a realizar por el **C. Luis Manuel Juárez López**, para obtener el grado de Maestro en Administración de Tecnologías de la Información.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

UNIVERSIDAD JUÁREZ
AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE
CIENCIAS Y TECNOLOGÍAS
DE LA INFORMACIÓN

Atentamente

MTE. Óscar Alberto González González
Director

C.c.p. Dr. Eddy Arquímedes García Alcocer. Encargado del Despacho de la Coordinación de Posgrado
Alumno
Archivo
Consecutivo

MTE´OAGG/EAGA

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690.
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx

Cunduacán, Tabasco a 17 de junio de 2024
Oficio No. 626/DACYTI/D/2024

Asunto: Dirección de Tesis

Dr. Rubicel Cruz Romero
Profesor Investigador

De conformidad con lo establecido en el Reglamento de Estudios de Posgrado Vigente, de la Universidad Juárez Autónoma de Tabasco, me permito informarle, que ha sido designado como Director de la tesis titulada **"Aprovechamiento, deserción y reprobación de estudiantes universitarios de Tabasco desde la minería de datos"**, a realizar por el **C. Luis Manuel Juárez López**, para obtener el grado de Maestro en Administración de Tecnologías de la Información.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

Atentamente

MTE. Óscar Alberto González González
Director

C.c.p. Dr. Eddy Arquímedes García Alcocer. Encargado del Despacho de la Coordinación de Posgrado
Alumno
Archivo
Consecutivo

MTE´OAGG/EAGA

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690.
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS
Y TECNOLOGÍAS DE LA INFORMACIÓN

F7: Respuesta de jurado

Cunduacán, Tabasco, a 25 de julio de 2024.

MTE. Óscar Alberto González González
Director de la División Académica de Ciencias y Tecnologías de la Información
Presente

En atención a los oficios girados por usted, en los que se nos designa como parte del jurado para efectuar la revisión de la tesis titulada ***"Aprovechamiento, deserción y reprobación de estudiantes universitarios de Tabasco desde la minería de datos"***, realizada por el ***C. Luis Manuel Juárez López***, estudiante de la Maestría en Administración de Tecnologías de la Información, nos permitimos informarle que, en virtud de que ha atendido las observaciones realizadas, otorgamos nuestra aprobación para que continúe los trámites para la obtención del grado.

Sin otro particular, aprovechamos la ocasión para enviarle un cordial saludo.

Atentamente integrantes del jurado

Dr. Gerardo Arceo Moheno

Dr. Rafael Mena de la Rosa

Dr. Eric Ramos Méndez

Dr. Pablo Payró Campos

Dr. Eddy Arquímedes García Alcocer

c.c.p. Dr. Eddy Arquímedes García Alcocer. Encargada del despacho de la Coordinación de Posgrado.
Estudiante.

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
RECIBIDO
8 JUL. 2024
D.A.C.Y.T.I.
DIRECCIÓN

Agradecimientos

Mis agradecimientos a la Universidad Juárez Autónoma de Tabasco en especial a la División Académica de Ciencias y Tecnologías de la Información que me permitió continuar con mi preparación profesional en esta maestría.

De igual manera para los profesores del NAB que me proporcionaron sus conocimientos en las áreas correspondientes y a mis directores de tesis que sin su ayuda nada hubiera sido posible, por último, pero no menos importante a CONAHCYT por impulsar y fortalecer estas áreas de investigación en la División mediante sus apoyos económicos.

Dedicatorias

Con mucho cariño para todas las personas que me permitieron seguir adelante en este grado académico, a mis padres, mis hermanos.

De igual manera va con una dedicatoria especial a mi novia Cindy Yajaira, por estar siempre conmigo, apoyándonos en cada momento en esta etapa académica que hemos terminado.

Resumen

El desempeño escolar que hubo en tiempos de pandemia debido al Covid-19 fue un reto en el ambiente educativo para tomar las clases a distancia puesto que no todos los alumnos contaban con los medios de comunicación necesarios para recibirlas. Sin embargo, las Tecnologías de la Información fueron el pilar fundamental para llevar a cabo las clases con las diferentes herramientas y plataformas en línea. Para efectos de esta investigación se aplicaron técnicas de minería de datos con un enfoque cuantitativo las cuales fueron clasificación, agrupación y reglas de asociación, el *Dataset* se obtuvo al realizar una encuesta integrada por seis *ítems,* la cual se efectuó en el año 2020 por medio de cuestionarios realizados en formularios de *Google* enviados a estudiantes que cursaban el ciclo enero-junio 2020 de los diferentes programas de licenciatura y posgrado que se imparten en Tabasco, México. Derivado de lo anterior, se identificaron las variables de aprovechamiento académico, deserción y reprobación en tiempos de pandemia. De igual manera, a través del modelo *FURPS (*derivado de sus siglas en inglés *Funcionality, Usability, Reliability, Performance & Suportability*) con el cual se evaluó la mejor herramienta para analizar el *Dataset*. Para el desarrollo de esta investigación se utilizó la herramienta Weka y el proceso *KDD* (Descubrimiento de Conocimiento en Bases de Datos) para llevar a cabo la minería de datos. De acuerdo con la interpretación de los resultados la falta de internet no fue un factor que orillara a los alumnos a tener una deserción del ciclo escolar que cursaban, los principales patrones apuntan que profesores les ofrecieron alternativas de trabajo si no contaban con los recursos necesarios para realizar sus actividades lo que permitió hubiera un desempeño académico balanceado con base a la deserción y reprobación.

Palabras clave: educación, minería de datos, pandemia

Abstract

The school performance during the pandemic due to COVID-19 was a challenge in the educational environment to take distance classes since not all students had the necessary means of communication to receive them. However, information technologies were the fundamental pillar for the classes' different online tools and platforms. For this research, data mining techniques with a quantitative approach were performed, which were classification, grouping, and association rules; the Dataset was obtained by conducting a survey composed of six items, which was carried out in the year 2020 through questionnaires made in Google forms sent to students who were studying the January-June 2020 cycle of the different undergraduate and graduate programs taught in Tabasco, Mexico. As a result of the above, academic achievement, dropout, and failure variables during the pandemic were identified. Similarly, the best tool for analyzing the Dataset was evaluated through the FURPS (Functionality, Usability, Reliability, Performance, Supportability) model. For the development of this research, the Weka tool and the KDD (Knowledge Discovery in Databases) process were used to carry out the data mining. According to the interpretation of the results, the lack of internet was not a factor that led the students to drop out of the school cycle they were attending; the main patterns indicate that teachers offered them work alternatives if they did not have the necessary resources to carry out their activities, which allowed them to have a balanced academic performance based on dropout and failure.

Keywords: education, data mining, pandemic.

Introducción

La presente investigación está compuesta por cinco capítulos en la cual se describen las generalidades o antecedentes de la investigación con referencia al tema de investigación, objetivos, metodología, resultados y principales conclusiones con la finalidad de identificar patrones de aprovechamiento, deserción y reprobación de estudiantes de educación superior de Tabasco. A continuación, se presenta un resumen capitular de la Tesis.

El capítulo uno denominado Generalidades contiene información con antecedentes, planteamiento del problema del cual surgió la pregunta de investigación, seguidamente los objetivos, justificación y metodología con los procesos correspondientes para llevar a cabo la aplicación de minería de datos.

En el capítulo dos llamado Marcos de la investigación se presentan teorías con referencia a la educación en México haciendo un contraste con la educación en pandemia el cual está estructurado en dos subtemas respectivamente, de igual forma contiene el marco referencial con investigaciones referentes al tema y las soluciones mediante técnicas de minería de datos dependiendo de la problemática, seguidamente está el marco conceptual, tecnológico y legal en donde se describen diferentes conceptos de minería de datos y desempeño académico y tipos de licencia de herramientas de minería de datos.

El capítulo tres denominado Aplicación de la metodología y desarrollo contiene tres apartados: el primero con información general del *Dataset* en donde se describe el instrumento y se exploran los datos, seguidamente con el análisis del modelo FURPS para la elección de la herramienta de minería de datos a emplear en el proyecto, y finalmente con la metodología del proceso KDD el cual consta de seis etapas las cuales

fueron desarrolladas con la herramienta Weka con las técnicas de clasificación, agrupación y reglas de asociación.

En el capítulo cuatro llamado Resultados se realizaron las pruebas correspondientes con las técnicas de minería elegidas en el *software* Weka. Se ejecutaron tres pruebas por cada técnica de minería de datos con la finalidad de seleccionar el modelo que generó mejores resultados y de esa manera interpretar las reglas y patrones descubiertos.

El último capítulo se encuentra dividido por cuatro apartados, el primero llamado conclusiones en el cual se resumen los hallazgos y resultados obtenidos durante la investigación, así como una reflexión sobre su importancia y posibles implicaciones, la relación con los objetivos acordados y la pregunta de investigación. En el segundo apartado se apuntaron las recomendaciones con base a la experiencia obtenida dentro del desarrollo del proyecto de investigación y resultados de los patrones identificados. En el tercer apartado se presentan los trabajos futuros en donde se sugieren líneas de investigación adicionales que surgieron a partir de las limitaciones de la investigación actual. Finalmente, en el apartado cuatro se discuten y analizan las implicaciones de los resultados obtenidos, se ofrecen explicaciones para ver las relaciones y relevancia del estudio.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1. Generalidades

El Covid-19 es una de las enfermedades que causó muchas muertes alrededor del mundo. En 2024 se encuentra controlada con menos casos de defunciones a nivel mundial y sobre todo en México. Sin embargo, en 2020 que fue cuando dio inicio, generó diversas problemáticas en todos los ámbitos, y el sector educativo no se escapó a la incidencia negativa de esta pandemia.

García *et ál.,* (2022) refieren que la educación, que es considerada uno de los pilares del desarrollo de las sociedades, también fue trastocada y agravada por la crisis sanitaria y el confinamiento obligatorio, lo cual obligó a repensar la manera de aprovechar eficientemente la modalidad virtual y los recursos digitales en línea, así como a incorporarlos de forma definitiva en los procesos educativos subsecuentes.

La Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO, 2020) refiere que, en la esfera de la educación, esta emergencia dio lugar al cierre masivo de las actividades presenciales de instituciones educativas en más de 190 países con el fin de evitar la propagación del virus y mitigar su impacto y que las Tecnologías de la Información (TI) siempre han sido un reto en el proceso educativo, sin embargo, con la llegada de la pandemia se tuvo que hacer una migración de la noche a la mañana para mudar las clases de la modalidad presencial, a una modalidad virtual; lo que significó una revolución de las estrategias utilizadas para enseñar y sobre todo la forma de aprender de los alumnos.

1.1 Aprovechamiento, Deserción y Reprobación de Estudiantes

Debido a la pandemia por Covid-19 hubo un cambio radical que aceleró el ritmo de uso de las tecnologías y provocó de manera urgente la implementación de diversos medios para recibir las clases a distancia en las Instituciones de Educación Superior (IES). Sin embargo, hubo muchos factores que afectaron a los alumnos de educación superior de Tabasco a reprobar o desertar de las clases por no contar con las herramientas digitales o los medios de comunicación necesarios para poder recibir sus clases a distancia. Derivado de las medidas de distanciamiento social propuestas por las autoridades sanitarias, las instituciones educativas tuvieron que adecuar los métodos de enseñanza, que más allá de generar una solución, provocaron el surgimiento de algunos contratiempos relacionados con el uso de las TI, como el uso de plataformas de comunicación y colaboración en línea como *Microsoft Teams* para la adecuación de cursos presenciales hacia la modalidad virtual, capacitación a profesores y alumnos para el manejo de aplicaciones y sistemas para la gestión de la educación que se realizaban de manera física para adaptarse a la llamada “nueva normalidad” (López y Contreras 2022). De modo que la llamada nueva modalidad presentó muchos obstáculos sobre todo en alumnos de educación superior.

No solo fueron situaciones de estrés de parte de los alumnos, también hubo barreras afrontadas ante innovaciones tecnológicas por parte de los docentes de educación superior en tiempos de Covid-19, Yanes *et. ál.,* (2022) mencionan que ante la aparición de nuevas plataformas como *Microsoft Teams*, los docentes universitarios

requirieron de cursos debido a que su implementación fue inesperada puesto que cuentan con diversos segmentos dentro de la misma y resultó compleja en sus inicios.

En este sentido la crisis por Covid-19 impactó al mundo en todos los aspectos a nivel empresarial, económico y sobre todo en al ámbito educativo, sin embargo, esta crisis obligó a las diferentes entidades a evolucionar a una nueva realidad donde predomina la innovación digital. Vega-Rodríguez y Botero-Suaza (2020) refieren que el ritmo del desarrollo tecnológico y la introducción de nuevas tecnologías en entornos educativos se han acelerado dramáticamente durante la última década. Si bien desde los últimos años se ha venido evolucionando en el proceso de la trasformación digital, debido a la pandemia, este proceso aceleró el ritmo del cambio y sobre todo la velocidad en que fueron implementadas las tecnologías emergentes en educación superior para resolver los problemas de clases a distancia.

Existen distintos motivos que influyeron a tomar ciertas decisiones en los estudiantes, por ejemplo, la falta de práctica de las innovaciones tecnológicas en los estudiantes de educación superior pudo traer como consecuencia una disminución en su rendimiento académico, provocando en el peor de los escenarios la deserción estudiantil. Debido al interés que existe por identificar los factores que provocan que un alumno deserte Reyes-Nava *et ál.,* (2021) efectuaron un estudio llamado Identificación de factores de riesgo que causan la deserción de alumnos que estudian a distancia por causa del Covid-19 usando técnicas de minería de datos, apuntan que haciendo uso de métodos de inteligencia artificial, minería de datos y aprendizaje automático en la mayoría

de los trabajos se hacía un énfasis en la educación presencial, sin embargo, actualmente por la situación que se vive es necesario analizar los factores de riesgo generados por los alumnos que estudian en línea, estos factores son almacenados en forma de datos y al final generan conocimiento que pueden entender las diferentes situaciones.

1.2 Planteamiento del Problema

1.2.1 Definición del Problema

La pandemia generada por Covid-19 causó el cierre de las instituciones educativas a nivel mundial, provocando un cambio inesperado en la forma de enseñar y evaluar a los alumnos de las Instituciones de Educación Superior, de tal modo que para recibir las clases a distancia se implementaran de forma particular diversos medios de comunicación como la televisión, radio y plataformas en línea para no perder el ciclo escolar.

La nueva modalidad virtual ocasionó que muchos alumnos de educación superior en Tabasco reprobaran o desertaran de clases, debido a que no contaban con las herramientas o recursos económicos necesarios para sustentar y mantener una conectividad a internet para realizar las actividades que les permitieran continuar sus estudios por esta modalidad, así lo refieren Ramírez-Melo *et ál.,* (2022) que los sistemas educativos no se encontraban preparados en su totalidad para sobrellevar aspectos técnicos, de ayuda psicológica y/o económica hacia los estudiantes, provocando que muchos de ellos enfrentaran el riesgo de desertar de sus estudios o ser vulnerables a otro tipo de afectaciones causadas por la pandemia. Al igual que los subsistemas

afectados como lo son los centros educativos que tenían una mayor proporción de prácticas de laboratorios y en consecuencia la virtualización no fue del todo satisfactoria en su proceso de enseñanza y provocó la baja de matrícula en educación superior (Martínez y Zamudio, 2021).

En Tabasco ante la situación que se vivió en tiempo de pandemia por Covid-19 evidentemente obligó a la población específicamente a la académica, a afrontar nuevos retos en el área educativa, causando que los estudiantes de nivel superior tomaran sus clases en línea ocasionando de alguna manera que los alumnos sufrieran diversos factores que lo llevaran a tener problemas en su aprovechamiento, que se reflejaran en reprobación o en el peor de los casos deserción de sus programas de licenciatura.

El estilo de afrontar la pandemia de los estudiantes de educación superior en Tabasco fue tomado de diferentes maneras dependiendo del ambiente en el que se mantuvieron, por consiguiente, es necesario identificar patrones de aprovechamiento, deserción y reprobación en el contexto de la pandemia por Covid-19 que hagan visible esta problemática.

1.2.2 Delimitación de la Investigación

Alcances.

Los alcances del estudio se enlistan a continuación:

- En la investigación se identificaron patrones de aprovechamiento, deserción y reprobación que afectaron a estudiantes de educación superior durante la pandemia por Covid-19 en Tabasco.

- Se hizo uso de herramientas de minería de datos que más expliquen el fenómeno a estudiar.
- La aplicación del instrumento se efectuó mediante formularios de *Google*.
- El número de personas encuestadas fue 8754 estudiantes de educación superior.

Limitaciones.

Las limitaciones del estudio se enlistan a continuación:

- Los resultados que se describen dependieron de las técnicas de minería de datos elegidas.
- Se utilizó la base de datos generada con las encuestas aplicadas.
- La investigación estuvo enfocada en las variables de aprovechamiento, deserción y reprobación de estudiantes de educación superior en el contexto de la pandemia por Covid-19.
- Mediante técnicas de minería de datos se identificaron patrones de aprovechamiento, deserción y reprobación que afectaron a los estudiantes de educación superior en Tabasco en el contexto de la pandemia por Covid-19.

1.2.3 Pregunta de Investigación

A raíz de la problemática planteada anteriormente surge la necesidad de responder la siguiente pregunta de investigación.

¿Qué patrones relacionados con el aprovechamiento, deserción y reprobación afectaron a los estudiantes de educación superior en Tabasco en el contexto de la pandemia por Covid-19?

1.3 Objetivos

1.3.1 Objetivo General

Identificar patrones de aprovechamiento, deserción y reprobación que afectaron a los estudiantes de educación superior de Tabasco en el contexto de la pandemia por Covid 19 empleando técnicas de minería de datos.

1.3.2 Objetivos Específicos

- Aplicar la técnica de árbol de decisión para la representación de patrones con relación en su aprovechamiento, deserción y reprobación de estudiantes de educación superior.
- Aplicar la técnica de agrupación para la representación e identificación patrones de aprovechamiento, deserción y reprobación de estudiantes de educación superior.
- Aplicar la técnica de reglas de asociación para la representación e identificación patrones de aprovechamiento, deserción y reprobación de estudiantes de educación superior.
- Analizar los resultados de patrones de aprovechamiento, deserción y reprobación obtenidos de estudiantes de educación superior en pandemia para la representación y exposición del conocimiento obtenido.

1.4 Justificación

La educación en tiempos de pandemia experimentó cambios en la manera de aprovechar eficientemente la modalidad virtual. Reyes-Nava *et ál.,* (2021) apuntan que,

en la mayoría de los trabajos de descubrimiento de conocimiento se hacía un énfasis en la educación presencial implementando métodos de inteligencia artificial, minería de datos o aprendizaje automático, sin embargo, actualmente por la situación que se vive es necesario analizar los factores de riesgo generados por los alumnos que estudian en línea y no solo quedarse en análisis de educación presencial.

En tal sentido en las Instituciones de Educación Superior es necesario implementar técnicas que permitan a estudios con grandes cantidades de datos recabados por medio de encuestas el análisis para encontrar información relevante. Sin embargo, esta investigación plantea usar técnicas de minería de datos porque como lo afirman Gutiérrez y Meza (2021) en la educación está siendo de gran utilidad para encontrar patrones en grandes conjuntos de datos, y con ello optimizar, predecir resultados que son de gran utilidad para la detección, prevención de problemas específicos y toma de decisiones, y de esta manera identificar patrones de aprovechamiento, deserción y reprobación que afectaron a los alumnos de educación superior en pandemia por Covid-19.

Entender este efecto ayudará a diseñar estrategias de educación, minimizar las consecuencias y tomar las medidas necesarias en función de las necesidades que se están presentado, ya que aportará nuevas evidencias sobre este fenómeno.

Desde el punto de vista tecnológico el proyecto de investigación tiene varios beneficios significativos al implementar minería de datos en desempeño académico como crear sistemas inteligentes que puedan predecir patrones y sugerir recomendaciones, del mismo modo la eficiencia en la gestión de datos educativos.

De igual forma en el ambiente educativo otro de sus fortalezas es identificar factores que afectan el rendimiento académico, con base a las necesidades identificadas adaptar métodos de enseñanza o abordar diferencias en el desempeño académico entre diferentes grupos de estudiantes generando nuevos conocimientos que puedan tener un impacto positivo a través de mejoras en la educación y desarrollo de tecnologías educativas que puedan innovar.

1.5 Metodología Utilizada

1.5.1 Enfoque de la Investigación

El enfoque metodológico que se le dio a la investigación fue el cuantitativo ya que al usar técnicas de minería de datos se analizan grandes cantidades de datos en donde se pueden hallar patrones y modelos en investigación educativa, de acuerdo con Dicovskiy y Pedroza (2018) las bases de datos académicas son un material importante en cualquier investigación educativa, y deberían ser estudiadas por minería de datos, como un método innovador dentro los métodos tradicionales de investigación cuantitativa.

1.5.2 Fuentes de Investigación

Como fuente primaria para el análisis de los datos en esta investigación se utilizó la base de datos del proyecto de investigación financiado por el Consejo de Ciencia y Tecnología del Estado de Tabasco (CCYTET) denominado “Diagnóstico Participativo Post Covid 19 en Tabasco”, y que contó con la participación de investigadores de diversas Divisiones Académicas de la UJAT. Este proyecto se efectuó en el año 2020 por medio de encuestas realizadas en formularios de *Google* enviadas a estudiantes que cursaban

el ciclo enero-junio 2020 de los diferentes programas de licenciatura y posgrado que se imparten en Tabasco, México. Se solicitó autorización al director de Investigación de la UJAT para hacer uso de la información recopilada. Del mismo modo se realizó la búsqueda de literatura como artículos, tesis y libros que le darán sustento teórico, conceptual y metodológico a esta investigación.

1.5.3 Técnicas de Recolección de Datos

Se realizó la aplicación de la encuesta titulada: “Aprovechamiento académico, deserción y reprobación en tiempos de pandemia. La perspectiva de los estudiantes de educación superior”. Se diseñó un instrumento con seis ítems para responderse con una escala tipo Likert, para describir la percepción sobre las variables de aprovechamiento académico, deserción y reprobación de estudiantes en el contexto de la pandemia por Covid-19 en Tabasco, se efectuó en el año 2020 por medio de formularios de *Google* enviado a estudiantes del ciclo enero-junio 2020, adscritos a 65 organizaciones, de las cuales 47 fueron IES públicas o privadas y 12 diversas organizaciones del sector público privado con un total de 8759 estudiantes de IES de estado de Tabasco México de todas las áreas de conocimiento y ciclos escolares de los diferentes programas de nivel licenciatura y posgrado.

De las respuestas obtenidas por los encuestados se extrajo un *Dataset* que fue de utilidad para generar patrones de aprovechamiento, deserción y reprobación.

1.5.4 Metodología para Aplicar

Para efectos de esta investigación, se consideraron el proceso *Knowledge Discovery in Databases* (KDD) para extraer patrones y reglas del *Dataset* y de igual forma el modelo FURPS (derivado de sus siglas en inglés Funcionality Usability Reliability Performance & Suportability) para la selección de la herramienta de minería de datos que evalúa el *software* mediante diferentes criterios y seleccionar el mejor.

Proceso KDD.

Con el fin extraer conocimiento del *Dataset* mediante el proceso KDD Timaran-Pereira *et ál.,* (2016) apuntan que el proceso descubrimiento de conocimientos en bases de datos es un proceso automático en el que se combinan descubrimiento y análisis.

El proceso que se utilizó consiste en extraer patrones que sean novedosos y útiles en forma de reglas o funciones, a partir de los datos, para que el usuario los analice y consta de las etapas que se muestran a continuación:

1. Selección
2. Preprocesamiento/limpieza
3. Transformación/reducción
4. Minería de datos (Data Mining)
5. Interpretación/evaluación
6. Utilizar el conocimiento descubierto

Modelo FURPS.

Este modelo se utilizó para evaluar el *software* de minería de datos mediante diferentes criterios y elegir el mejor con respecto a su desempeño. Su nombre se desglosa a continuación por sus siglas.

- Funcionalidad,
- Usabilidad,
- Confiabilidad (Reliability),
- Desempeño (Performance)
- Soporte.

Capítulo 2. Marcos de la Investigación

2.1 Marco Teórico

Fueron muchos factores que afectaron al sistema educativo a nivel nacional por Covid-19, provocando un desempeño académico diferente dependiendo del ambiente en que se mantuvieron. Miguel (2020), Ortega-Encinas *et ál*., (2022) y Ramírez-Melo *et ál.,* (2022) apuntan que la baja economía, estrés, problemas psicológicos, falta de herramientas o de recursos tecnológicos fueron los factores más destacados. Para efectos de esta investigación se contemplan los componentes relacionados con variables de aprovechamiento, deserción y reprobación de estudiantes a través de técnicas de minería de datos.

Han surgido modelos que permiten organizar y clasificar las variables para caracterizarlas, explicarlas y en el mejor de los casos predecirlas. En la actualidad se han diseñado una cantidad pronunciada de algoritmos que pueden ser utilizados para el análisis descriptivo y predictivo de un fenómeno, sin embargo, es importante darles relevancia a todas las etapas de la metodología de minería de datos, para garantizar la calidad e integridad de los datos, con la finalidad de que los resultados obtenidos sean confiables y pertinentes para la toma de decisiones (Beltrán 2022).

2.1.1 La Educación en México

En los sistemas educativos en México la reprobación escolar siempre ha estado presente, Diaz (2018) menciona que uno de los problemas que siempre ha enfrentado el sistema educativo nacional es la deserción escolar, al igual que el rezago académico y la

baja eficiencia terminal, que lejos de ser un asunto local, es una temática en el ámbito mundial; afirma que la deserción, la baja eficiencia terminal y el bajo rendimiento escolar son problemáticas asociadas a los niveles de reprobación en alumnos, los cuales están relacionados con elementos familiares, sociales, psicológicos, económicos, perfiles de ingreso limitados o falta de hábitos de estudio.

En contraste con lo anterior Durón y Oropeza (1999 como se citó en Izar *et ál*., 2011) mencionan la presencia de cuatro factores que influyen en el aprovechamiento académico los cuales son:

- Factores fisiológicos. Se sabe que afectan, aunque es difícil precisar en qué medida lo hace cada uno de ellos, ya que por lo general están interactuando con otro tipo de factores. Entre los que se incluyen en este grupo están: cambios hormonales por modificaciones endocrinológicas, padecer deficiencias en los órganos de los sentidos, desnutrición y problemas de peso y salud.
- Factores pedagógicos. Son aquellos aspectos que se relacionan con la calidad de la enseñanza. Entre ellos están el número de alumnos por maestro, los métodos y materiales didácticos utilizados, la motivación de los estudiantes y el tiempo dedicado por los profesores a la preparación de sus clases.
- Factores psicológicos. Entre estos se cuentan algunos desórdenes en las funciones psicológicas básicas, como son la percepción, la memoria y la conceptualización, los cuales dificultan el aprendizaje.

- Factores sociológicos. Son aquellos que incluyen las características familiares y socioeconómicas de los estudiantes, tales como la posición económica familiar, el nivel de escolaridad y ocupación de los padres y la calidad del ambiente que rodea al estudiante.

En esta investigación las variables estudiadas será el aprovechamiento, deserción y reprobación de estudiantes de educación superior de Tabasco.

2.1.2 La Educación en México en Tiempos de Pandemia

La enseñanza virtual tuvo que implementarse por causa de la pandemia, ya que ésta última provocó que el ecosistema educativo tuviera dificultades en el área tecnológica para asumirse la enseñanza en la modalidad virtual, aunado a lo anterior hubo incidencia en muchos estudiantes en relación con su aprovechamiento, deserción o reprobación académica.

Aquino (2020) indica que, las principales dificultades presentadas en pandemia fueron las académicas y personales, ya que la mayoría de los profesores no utilizaban la plataforma de Teams y por lo tanto no se tenía la sesión completa, las dificultades a nivel personal que se enfrentaron fueron con respecto al estrés provocado por el confinamiento en términos de restricción de actividades sociales, recreativas y deportivas; y efectos socioemocionales como la tristeza, inconformidad, angustia, temor. Finalmente explica que otro de los problemas enfrentados fue el factor económico ya que algunos alumnos no podían seguir trabajando para ayudarse en sus gastos personales.

Las razones para darse de baja fueron de aspectos tecnológicos y conectividad, puesto que había complicaciones para mantenerse comunicado y enviar sus tareas, darse de baja o desertar fueron decisiones que se tomaron por estas razones, fueron muchos factores que afectaron el rendimiento escolar debido al cambio radical por mover las clases a diversas plataformas en línea. En este sentido Ortega-Encinas *et ál*., (2022) apuntan que un gran porcentaje de alumnos y maestros no estaban preparados para la enseñanza-aprendizaje a distancia ya que carecían de habilidades digitales y de los dispositivos necesarios para su correcta implementación dada la premura derivada de la contingencia. Así mismo mencionan que un porcentaje de estudiantes no tendrían los recursos suficientes para afrontar la nueva realidad y debido a esto, la pandemia modificó diferentes hábitos de estudiantes.

Del mismo modo refieren que la educación virtual o a distancia posee características que la diferencian en gran medida de la educación presencial, las cuales se señalan a continuación.

- Existe una mayor autonomía e independencia de parte del alumnado para el desarrollo de su proceso de aprendizaje, siendo el estudiante quien marca su ritmo de trabajo.
- Muchos de los estudiantes conceden un carácter más práctico a sus objetivos de aprendizaje.

Ante los cambios que implica el ajuste de clases presenciales a virtuales Miguel (2020) indica que algunas de las inconformidades de los estudiantes ante este cambio

radicaron en la mala comunicación con los profesores; las clases se basaban en cargas de tareas, sin explicación previa o retroalimentación; en algunos casos, la conectividad representaba un problema. Por su parte, quienes manifiestan estar bien y cómodos con el cambio mencionaron que el estar en sus casas les brinda paz, y ahorran tiempo, lo que se puede traducir en optimización del tiempo. Por consiguiente, puede decirse que fueron perspectivas y realidades diferentes que causaron un desempeño académico distinto, aunque el tener una buena conectividad o recursos tecnológicos se ve ligada al factor económico.

Las características que definen qué intervino para tener un desempeño académico reflejado en aprovechamiento, deserción y reprobación académico va dependiendo de diferentes autores como Aquino (2020), Martínez y Zamudio (2021), Miguel (2020), Ortega-Encinas *et ál*., (2022), Ramírez-Melo *et ál.*, (2022), Reyes-Nava (2021), Vega-Rodríguez y Botero-Suaza (2020), en las que se especifican las características principales que llevaron a influir el desempeño académico de los estudiantes desde antes de pandemia y en pandemia. A continuación, se presentan los factores que influyeron en el desempeño académico en alumnos (ver tabla 1).

Tabla 1

Factores que Influyeron en el Desempeño Académico en Alumnos

Antes de pandemia por Covid-19	**En pandemia por Covid-19**
Factores fisiológicos	Dificultades académicas
Pedagógicos	Dificultades personales
Psicológicos	Estrés
Sociológicos	Tristeza
Familiares	Recursos económicos

Sociales	Recursos tecnológicos
Económicos	Conectividad
Falta de hábitos de estudio	Carencia de habilidades digitales

Nota: Elaboración propia.

Como se muestra en la tabla anterior, es notorio que los muchos factores se involucran en el desempeño académico de estudiantes de educación superior, desde siempre han existido barreras que interfieren en el aprovechamiento académico de los estudiantes en pandemia y antes de pandemia.

2.2 Marco Referencial

A continuación, se describen las investigaciones más relevantes en orden descendentes por año, con referencia a tecnologías y desempeño académico de estudiantes con diferentes técnicas de ciencia de datos, de igual forma se hace referencia a aspectos tecnológicos y económicos que influyeron en el desempeño académico de los estudiantes.

2.2.1 Análisis de los índices de reprobación en la carrera de ITICS utilizando técnicas de inteligencia artificial y minería de datos en el Tecnológico Nacional de México Campus Conkal

Pech *et ál*., (2022) realizaron una investigación identificando factores que influyen en los índices de reprobación y deserción de las y los estudiantes de la carrera de ITIC, utilizando técnicas de inteligencia artificial y minería de datos mediante el *software* WEKA. Se aplicaron algoritmos de evaluación de atributos y de clasificación como árboles de decisión. Se identificaron variables influyentes en los índices de reprobación y deserción, así como su relación con el desempeño académico finalmente entre los resultados más

destacados se observó que las materias de programación y electrónica son un alto referente en los índices de reprobación y deserción de las y los estudiantes.

2.2.2 El impacto de la pandemia por Covid-19 en estudiantes mexicanos de educación media superior

López (2022) realizó una investigación cuyo objetivo fue describir el impacto de la enfermedad por coronavirus de 2019 (Covid-19) en estudiantes mexicanos de educación media superior utilizando una metodología que consistió en una investigación exploratoria con un enfoque cuantitativo de diseño no experimental, de corte transversal y con un alcance descriptivo. Para la recolección de datos, se aplicó un cuestionario a modo de encuesta que permitió caracterizar el comportamiento del fenómeno a partir de cuatro dimensiones: infraestructura tecnológica, capacitación, ámbito social-económico y de salud y competencias digitales. Entre los resultados, destaca que, en la mayoría de los casos, los estudiantes cuentan con la infraestructura tecnológica adecuada para continuar con los procesos de aprendizaje de manera virtual, así como con las competencias digitales para el uso de aplicaciones que promuevan la comunicación. Respecto a los efectos ocasionados a la salud provocados por las medidas sanitarias, se detectó que mayoritariamente los estudiantes presentaron algún tipo de afectación, principalmente de índole emocional.

2.2.3 Aplicación de técnicas de minería de datos para la caracterización de estudiantes bajo el efecto de la Covid-19

Ramírez-Melo et al., (2022) integraron técnicas de minería de datos, tales como el análisis de clúster jerárquico y la regresión logística, para caracterizar alumnos de un programa educativo de la Universidad Autónoma del Estado de Hidalgo (UAEH). Se realizó un sondeo a los alumnos para obtener datos acerca de las condiciones sociodemográficas, económicas, técnicas, de salud mental y académicas que permitan encontrar patrones que inciden en el desempeño académico del alumnado, mediante el modelo de aprendizaje no supervisado se detectó que si el alumno tiene calificaciones más bajas es porque está en riesgo de reprobar, lo mismo sucede si el alumno accede a las clases mediante un dispositivo de acceso público o de renta. En cambio, que se conecte por medio de una laptop o computadora de escritorio disminuye sus probabilidades de ser vulnerable.

2.2.4 Identificación de factores de riesgo que causan la deserción de alumnos que estudian distancia por causa del Covid-19 usando técnicas de minería de datos

Reyes-Nava *et ál*., (2021) realizaron esta investigación donde se muestran los factores que influyen en la deserción escolar en la carrera de Ingeniería en Sistemas Computacionales del Tecnológico de Estudios Superiores de Jocotitlán, durante el periodo de contingencia, se muestran los hechos la situación actual del Tecnológico por el desempeño de los alumnos en las clases en línea explorando los datos mediante el

uso del algoritmo *A priori de reglas de asociación,* mediante minería de datos con información recabada mediante la aplicación de formularios con 75 atributos y 250 registro de alumnos mediante el uso del *software* Weka con el cual se interpretan los patrones de cada factor.

De las reglas obtenidas se llega a los resultados con el algoritmo *A priori* se concluyeron algunas deducciones que se muestran a continuación.

1. Estudiantes que no tienen hijos y han cursado en una institución pública la preparatoria no han experimentado ninguna situación de penuria en los últimos seis meses.
2. Aquellos estudiantes que viven solo con su madre y que no han experimentado ningún evento que haya ocurrido, han alterado sus hábitos son alumnos regulares.
3. La mayoría de los alumnos mencionan haber elegido la carrera de sistemas como su primera opción ya que es el área que siempre les ha gustado, de igual forma son alumnos que tienen internet en cada uno y no pasan por ningún evento que ha alterado sus hábitos normales en el último año. También se identificó que la mayoría de los estudiantes tienen un promedio actual entre 80 y 89.
4. Tienen una computadora personal pero su mayor forma de conectarse a internet es mediante datos celulares. En cuanto a los factores de apoyos económicos, no saben si el tecnológico cuenta con un departamento de becas, así como tampoco están muy bien informados sobre los apoyos que ellos pudieran recibir.

2.2.5 Patrones que identifican a estudiantes universitarios desertores aplicando minería de datos educativa

Urbina-Nájera (2021) realizó un análisis cuantitativo no experimental de carácter exploratorio del fenómeno de la deserción universitaria utilizando algoritmos de la minería de datos educativa. Se utilizó un conjunto de datos de 10 635 instancias, adquiridas en el período 2014-2019, de 53 programas de licenciatura de una institución privada del estado de Puebla (México). Los resultados muestran que el modelo obtenido por los árboles de decisión ofrece mayor desempeño que otros algoritmos, así como una fácil interpretación de éste mediante reglas de decisión.

Los métodos de selección de características permitieron encontrar los atributos más importantes que identifican a un potencial desertor, tales como: el período, el último semestre cursado, créditos cursados, asistencia, materias reprobadas y programa. Utilizando los atributos y reglas de decisión encontradas se podrían crear mecanismos que favorezcan la prevención de la deserción.

2.2.6 Modelo Basado en Árbol de Decisiones para Determinar los Factores de Deserción de Estudiantes en una Institución de Educación Superior Mexicana

Gutiérrez y Meza (2021) realizaron una investigación en la que se analizan los datos obtenidos de una encuesta aplicada a 1,582 estudiantes para determinar los principales factores que influyen en la deserción escolar en una etapa pre-Covid-19. Con

esta información se desarrolló un análisis del árbol de decisión, detectando las principales rutas que influyen en la deserción escolar con la herramienta R.

Mencionan que se pueden representar diferentes situaciones personales en los alumnos, algunas de ellas podrían ser las siguientes: El modelo indica que, si el alumno no tiene problemas de salud, pero si tiene problemas económicos, y falta de tiempo, la probabilidad de desertar es la más alta.

Otra situación personal que no afectaría considerablemente la decisión de deserción sería que, si alguien tiene problemas de salud, problemas personales, es altamente probable que repruebe una o más asignaturas.

2.2.7 Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico

Pérez-Gutiérrez (2020) realizó un estudio de comparación de técnicas para apoyar la identificación de deserción estudiantil a partir del registro académico de los estudiantes de una Universidad en Colombia, se recopilaron datos de 762 estudiantes matriculados en el Programa de Ingeniería de Sistemas para ese trabajo se abordó una tarea de analítica basada en una clasificación binaria, utilizando la metodología CRISP-DM. Árboles de decisión, regresión logística y Redes bayesianas, fueron comparados para lograr establecer la mejor técnica de detección de desertores. Adicionalmente, la herramienta Watson Analytics de IBM fue utilizada para comparar su usabilidad y precisión. A partir del resultado del modelo Árbol de Decisión se destacan dos casos:

1. Caso 1: Un estudiante cuyo promedio en los cursos sea menor o igual a 3,505, y que el número de veces en que ha fallado cursos sea menor o igual a 0,1389 (escalado entre 0 y 5), y un GPA menor o igual a 3,5164, tendrá el 100% de probabilidad de desertar. Esto equivale a 143 estudiantes del *Dataset* utilizado.
2. Caso 2: Un estudiante cuyo promedio en los cursos) sea menor o igual a 3,505 y que el número de veces en que ha fallado cursos de SE sea menor o igual a 0,1389 (escalado entre 0 y 5), y un GPA mayor a 3,5164, tendrá 88.8% de probabilidad de desertar. Esto equivale a 48 estudiantes del *Dataset* utilizado.

2.2.8 Baja o deserción en tiempos de la pandemia del Covid-19. Experiencia de un estudiante de pregrado

Aquino (2020) analiza el caso de un estudiante de la Licenciatura en Idiomas de la División Académica de Educación y Artes (DAEA) de la Universidad Juárez Autónoma de Tabasco (UJAT), quien decidió darse de baja a punto de concluir el ciclo escolar 2020_02. Indica que, las principales dificultades presentadas en pandemia fueron las académicas y personales, ya que la mayoría los profesores no utilizaban la plataforma de Teams y por lo tanto no se tenía la sesión completa, las dificultades a nivel personal que se enfrentaron fueron con respecto al estrés provocado por el confinamiento en términos de restricción de actividades sociales, recreativas y deportivas; y efectos socioemocionales como la tristeza, inconformidad, angustia, temor. Finalmente explica que otro de los problemas enfrentados fue el factor económico ya que alumnos no podían

seguir trabajando para ayudarse en sus gastos personales. Las razones para darse de baja fueron de aspectos tecnológicos y conectividad, puesto que había complicaciones para mantenerse comunicado y enviar sus tareas, darse de baja o desertar fueron decisiones que se tomaron por estas razones.

2.2.9 Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos

Miranda y Guzmán (2017) realizaron un estudio de tipo descriptivo mediante el análisis de datos cuantitativos con tres tipos de clasificadores: clasificador de redes bayesiana, clasificación basada en árboles de decisión y clasificación basada en redes neuronales para determinar cuáles son y cuál es la importancia de las variables que llevan a un estudiante a abandonar sus estudios universitarios, usando técnicas de minería de datos. Los resultados obtenidos a partir de los datos proporcionados por las carreras de Ingeniería de la Universidad Católica del Norte en Antofagasta y Coquimbo (Chile) determinan que las variables que mejor explican la deserción de un estudiante son, las razones socioeconómicas y el puntaje de ingreso a la universidad.

2.2.10 Estudio de variables que influyen en la deserción de estudiantes universitarios de primer año, mediante minería de datos

Zarria *et ál.,* (2017) realizaron un estudio donde analizaron la deserción de los estudiantes mediante técnicas de minería de datos y poder obtener un modelo que fuese capaz de clasificar estudiantes desertores a partir de los datos socioeconómicos y académicos, se utilizó CRISP-DM como metodología para guiar las etapas del proyecto

y se analizaron tres diferentes modelos de clasificación: árboles de decisión, métodos bayesianos y redes neuronales, con el fin de evaluar su comportamiento, encontrándose que arboles de decisión es el algoritmo de mejor desempeño general, con un 88.9% de exactitud, mientras que el algoritmo redes bayesianas resultó ser el más adecuado para dar respuesta a los objetivos del proyecto, se determinó que las variables académicas de ingreso de los estudiantes no resultan significativas para explicar la deserción de primer año.

2.3 Marco Conceptual

2.3.1 Aprovechamiento

El aprovechamiento escolar puede concebirse como el nivel de conocimientos, habilidades y destrezas que el alumno adquiere durante el proceso enseñanza-aprendizaje, se considera que el rendimiento escolar es el nivel de conocimientos demostrado en un área o materia comparada. Así, tal rendimiento no es sinónimo de capacidad intelectual, de aptitudes o de competencias, sin embargo, con las clases a distancia pudieron intervenir muchos factores que no permitieron tener un desempeño correcto por alumnos de educación superior (Osorio 2012).

Se considera que el aprovechamiento es la capacidad o productividad que tienen los estudiantes para el logro de objetivos establecidos en un grado académico especifico así lo apunta Garbanzo (2007) que el rendimiento académico es la suma de diferentes y complejos factores que actúan en la persona que aprende, y ha sido definido con un valor atribuido al logro del estudiante en las tareas académicas. De igual forma los indicadores

de desempeño se plantean como referentes de logro para los estudiantes, son medibles y demuestran qué tanto se ha alcanzado en el proceso de instrucción, es decir, demuestran el desempeño académico.

De igual manera, Jiménez (2000 como se citó en Saucedo *et ál.*, 2014), entiende como rendimiento escolar el nivel de conocimientos demostrados en un área o materia comparado con la norma de edad y nivel académico y Saucedo *et ál*., (2014) lo definen como el resultado cuantitativo obtenido durante el proceso de aprendizaje conforme a las evaluaciones que realiza el docente mediante pruebas objetivas y otras actividades complementarias.

2.3.2 Deserción

La deserción escolar es un proceso de alejamiento sucesivo de la escuela que culmina con el abandono por parte del alumno como consecuencia de la separación su programa educativo. De acuerdo con Corzo (s.f.) en el plano educativo, se utiliza el término para hablar de aquellos alumnos que abandonan sus estudios por diferentes causas. De igual forma hay muchos factores que intervienen en la deserción escolar como factores socio-económicos, factores personales, psicológicos, historia académica personal, institucionales, factores pedagógicos, factores familiares y factores sociales.

Por otra parte, Rico (2006 como se citó en Chalpartar *et. ál.*, 2022) apunta que la deserción estudiantil se entiende como la interrupción de las actividades académicas; puede ser el resultado de la interacción de una serie de características o variables, tales

como el contexto, factores económicos, tecnológicos, psicológicos sociales, demográficos, familiares, individuales, entre otros.

En sí la deserción es un abandono, alejamiento temporal o definitivo del programa educativo que cursa el estudiante por diversas circunstancias ajenas al alumno matriculado en la institución así lo afirman Ramón *et ál.,* (2023) que el concepto se asocia como fuga o retiro del estudio, desvinculación temporal o definitiva, problema multicausal, abandono del aprendizaje, fin al proceso de formación, ausentismo de clases, interrupción del periodo siguiente, decisión de alterar la continuidad del estudio renuncia voluntaria entre otros.

2.3.3 Reprobación

De acuerdo con Saucedo *et ál.*, (2014) el fenómeno académico de la reprobación no es más que la manifestación del bajo aprovechamiento escolar por falta de sus habilidades, destrezas, aptitudes, ideales, intereses.

Villalobos (2021) refiere que la educación virtual presenta desventajas sobre la educación presencial, este fenómeno puede tener diferentes causas, tales como la dificultad del estudiante para adaptarse al nivel universitario, la deficiencia en los conocimientos previos no adquiridos en el nivel de educación anterior, problemas en sus hábitos de estudio, organización de actividades y tareas, problemas económicos y provocar una reprobación.

Corso (s.f. cómo se citó en Diaz y Ruiz 2018) apunta que la reprobación escolar consiste en no aprobar una o varias materias en un determinado grado o nivel, lo cual se

relaciona a elementos sociales y familiares, psicológicos, económicos, perfiles de ingreso limitados y falta de hábitos de estudio. De igual manera Diaz y Ruiz (2018) refieren que reprobar es expresión de un bajo aprovechamiento escolar y signo claro de una desigualdad en el aprendizaje.

Sin embargo, para que la educación en línea sea efectiva requiere de una serie de situaciones para lograrlo, en pandemia se generó un caos en la manera de impartir las clases puesto que se utilizaron muchas alternativas que llevaron a la educación a tener un cambio de forma radical por no contar con los medios alternativos suficientes para continuar a distancia y optar por la vía de reprobar.

De tal manera que los factores académicos en tiempos de pandemia por Covid-19 son relevantes para el estudio de las variables de esta investigación ya que fueron elementos que causaron un desempeño diferente y se vieron reflejados en un aprovechamiento, deserción o reprobación de estudiantes de educación superior.

2.3.4 Modelo FURPS

Bajo el acrónimo de FURPS: funcionalidad (Functionality), usabilidad (Usability), confiabilidad (Reliability), desempeño (Performance) y capacidad de soporte (Supportability); Olsina (2007 como se citó en Constanzo 2014) menciona que este modelo fue desarrollado por Hewlett-Packard en el año 1987. En él se desarrollan un conjunto de factores de calidad de *software* que permiten evaluarlo bajo diferentes criterios para utilizar el de mejor desempeño. A continuación, se presentan los factores y criterio asociados al modelo FURPS (Ver tabla 2).

Tabla 2

Factores y Criterios Asociados al Modelo FURPS.

Factores	Criterios
Funcionabilidad	Adaptabilidad Exactitud Interoperabilidad Seguridad
Usabilidad	Comprensibilidad Aprendizaje Operatividad Atractivo
Confiabilidad *Realiability*	Análisis Cambio Estabilidad Prueba Madurez Tolerancia a fallos Recuperabilidad
Desempeño *Performance*	Comportamiento del tiempo Uso de los recursos
Soporte	Adaptabilidad Instalación Coexistencia Remplazo

Nota: Elaboración propia con información de Constanzo (2014).

2.3.5 5.3.4 Minería de Datos

De acuerdo con Troche (2014) la minería de datos es el campo de las ciencias orientadas a la informática referido al proceso que intenta descubrir conocimiento a través de patrones en grandes volúmenes de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos.

De la misma manera Cendejas-Valdez *et ál*., (2020) apuntan que la minería de datos es un conjunto de técnicas y herramientas de *software* que permiten extraer información de grandes cantidades de datos al identificar y reconocer una serie de patrones y algoritmos como los utilizados y ligados con la inteligencia artificial.

Con las definiciones anteriores se puede decir que el objetivo general del proceso de minería de datos consiste en extraer conocimiento por medio de la identificación de patrones y tendencias de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.

2.3.6 Proceso KDD

Riquelme *et ál.,* (2006) apuntan que el descubrimiento de conocimiento en bases de datos (KDD) se define como el proceso de identificar patrones significativos en los datos que sean válidos, novedosos, potencialmente útiles y comprensibles para un usuario. El proceso KDD es interactivo e iterativo conteniendo los siguientes pasos:

1. Comprender el dominio de aplicación: este paso incluye el conocimiento relevante previo y las metas de la aplicación.
2. Extraer la base de datos objetivo: recogida de los datos, evaluar la calidad de los datos y utilizar análisis exploratorio de los datos para familiarizarse con ellos.
3. Preparar los datos: incluye limpieza, transformación, integración y reducción de datos. Se intenta mejorar la calidad de los datos a la vez que disminuir el tiempo requerido por el algoritmo de aprendizaje aplicado posteriormente.

4. Minería de datos: esta es la fase fundamental del proceso. Está constituido por una o más de las siguientes funciones, clasificación, regresión, agrupamiento, resumen, recuperación de imágenes, extracción de reglas.
5. Interpretación: explicar los patrones descubiertos, así como la posibilidad de visualizarlos.
6. Utilizar el conocimiento descubierto: hacer uso del modelo creado.

Este proceso se puede ejecutar de manera iterativa por lo cual es aplicable las veces que sea necesario, este se implementará para la extracción de automatizada de conocimiento en el *Dataset* obtenido.

2.4 Marco Tecnológico

En la actualidad existen diversas herramientas tecnológicas para minería. En este apartado se describen algunas herramientas que han sido implementadas para aplicar técnicas de minería de datos.

2.4.1 Herramientas para Minería de Datos

SAS Enterprise Miner.

Jaramillo y Paz-Arias, (2015) refieren que es una herramienta de minería de datos comercializada, crea modelos predictivos y descriptivos precisos sobre grandes volúmenes de datos a través de diferentes fuentes mediante un proceso transparente, lo que permite colaborar de manera más eficiente, incluye una interfaz de usuario intuitiva que incorpora los principios de diseño comunes establecidos para el *software* de SAS y herramientas de navegación adicionales para mover fácilmente

alrededor del área de trabajo. Esta herramienta permite implementar diferentes algoritmos como reglas de asociación y evaluar modelos estadísticos con su principal cualidad de poder manejar grandes bases de datos y facilitar el trabajo de minería.

Orange

De acuerdo con Zupan (2012 como se citó en Cendejas-Valdez *et ál*., 2017) Orange es una herramienta de minería de datos de código abierto para el análisis de datos de forma interactiva basado en componentes de procedimientos de minería de datos. Su uso principal es la exploración de datos en la cual que se probaron y anotaron diferentes combinaciones de algoritmos de preprocesamiento y aprendizaje mediante validación cruzada y modelado predictivo.

Waikato Environment for Knowledge Analysis

Waikato Environment for Knowledge Analysis (WEKA) es un proyecto desarrollado en la Universidad de Waikato de Nueva Zelanda. Esta plataforma de *software* provee una colección de algoritmos para el preprocesamiento, filtrado de datos y de minería de datos, incluyendo algoritmos de regresión, clasificación, agrupamiento, reglas de asociación y selección de atributos (Martínez, 2017).

Sin embargo, Sudhir (2013 como se citó en Cendejas-Valdez *et ál*., 2017) refieren que Weka es un *software* libre disponible bajo General Public License (GNU) el aplicativo se caracteriza por tener una colección de herramientas de visualización y algoritmos como técnicas de procesamiento clasificación o agrupamiento para el análisis de los datos y el modelado predictivo, junto con interfaces gráficas de usuario para un acceso

fácil a esta funcionalidad. Esta serie de algoritmos de aprendizaje automático es bastante versátil para resolver problemas de minería de datos en el mundo real.

Lenguaje R

Actualmente R es considerado la lengua franca de la estadística, debido a algunas de sus características que lo sitúan muy por encima de prácticamente todos sus competidores. De acuerdo con (Lizana 2020) como tal R es un lenguaje de programación empleado primordialmente para efectuar análisis estadístico de datos y construcción de gráficos, R es gratuito y libre, es muy versátil, permite realizar una cantidad insospechable de procedimientos estadísticos y gráficos, permite construir gráficos de calidad inmejorable.

Python

En los últimos años Python se ha vuelto un lenguaje con mucha preferencia para los usuarios, es utilizado para ciencia de datos a diferencia de R, es interpretado y busca mejorar el código legible a través de la sintaxis. De acuerdo con la Universidad de Alcalá (s.f.) al soportar orientación a objetos, la programación funcional y la programación imperativa, se trata de un lenguaje multiparadigma que usa tipado multiplataforma y dinámico.

Dicho lenguaje de programación es utilizado en ciencia de datos para analizar textos mediante el procesamiento de lenguajes naturales. A su vez, se usa para extraer información valiosa de diferentes bases de datos y programar algoritmos de aprendizaje en Machine Learning.

RapidMiner

De acuerdo con Jaramillo y Paz-Arias (2015) RapidMiner es una herramienta de minería de datos desarrollado en Java, permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de 500 operadores a través de un entorno gráfico, permite utilizar los algoritmos incluidos en Weka, contiene técnicas de preprocesamiento de datos, modelación predictiva y descriptiva, métodos de entrenamiento y prueba de modelos, visualización de datos, aprendizaje automático.

2.5 Marco Legal

2.5.1 Licencia de Software Libre

En este caso se destacan las herramientas Weka, Orange, Python y R por ser herramientas de minería de datos libres y se caracterizan de esta manera porque pueden ser descargadas gratuitamente desde la web, además que en términos generales es un *software* que puede ser modificado.

2.5.2 Licencia de Software Comercial

En este caso se destacan las herramientas RapidMiner y SAS Enterprise Miner, por ser *software* comercial se tiene que pagar una licencia de uso del programa para ser ejecutado en el ordenador.

2.5.3 Ley de Protección de Datos Personales

Para llevar a cabo el análisis de los datos se hará uso de un *Dataset* o conjunto de datos que se recabó al realizar una encuesta a alumnos de educación superior del

estado de Tabasco, cabe mencionar que existe una ley que protege los datos personales en México llamada Ley Federal de Protección de Datos Personales en Posesión de los Particulares.

Así mismo la Ley Federal de Protección de Datos Personales en Posesión de los Particulares (2010) afirma que tiene por objeto la protección de los datos personales en posesión de los particulares, con la finalidad de regular su tratamiento legítimo, controlado e informado, a efecto de garantizar la privacidad y el derecho a la autodeterminación informativa de las personas.

De tal manera que esta ley busca sancionar a todas aquellas personas que incumplan las normas de los derechos ya que se busca garantizar el derecho a la privacidad de los datos personales con el fin de evitar un mal uso de la información de las personas y en caso de no cumplir se llevará a cabo un procedimiento que sancione al infractor de los datos (Ley Federal de Protección de Datos Personales en Posesión de los Particulares, 2010).

Capítulo 3. Aplicación de la Metodología y Desarrollo

3.1 Información general

3.1.1 Descubrimiento de la información

Para efectos de esta investigación es necesario conocer la cantidad y calidad de información obtenida en la encuesta aplicada que se utilizó para el análisis mediante el proceso KDD y minería de datos.

El Dataset de PRODECTI (2020) que se utilizó en este trabajo se obtuvo de la recolección de datos mediante encuestas aplicadas el cual tuvo como objetivo describir la percepción sobre las variable de aprovechamiento académico, deserción y reprobación de estudiantes en el contexto de la pandemia por Covid-19, se efectuó en el año 2020 por medio de formularios de *Google* enviado a estudiantes del ciclo enero-junio 2020, adscritos a 65 organizaciones, de las cuales 47 fueron IES públicas o privadas y 12 diversas organizaciones del sector público privado.

De acuerdo con el total de registros recabados se obtuvo una muestra de 8759 estudiantes de IES del estado de Tabasco, México de todas las áreas de conocimiento y ciclos escolares de los diferentes programas de nivel licenciatura y posgrado encuestados, considerando aspectos como la institución académica, género, edad, semestre, municipio, así como preguntas referentes al desempeño académico, si aprendieron lo mismo en las clases virtuales que en las presenciales, si los profesores ofrecían alternativas al no contar con recursos tecnológicos, darse de baja del ciclo escolar por problemas derivados de la contingencia, si las formas de evaluación de forma

virtual fueron las adecuadas o si esa modalidad afectó la calidad de la educación . Con estas preguntas se buscó describir la percepción sobre el aprovechamiento deserción y reprobación de estudiantes de educación superior en Tabasco, México.

A continuación, se presenta la descripción de los estudiantes encuestados por institución y género (ver tabla 3).

Tabla 3

Estadístico de Estudiantes por Institución y Género.

Institución	Masculino	Femenino	Total
Universidad Juárez Autónoma de Tabasco	2905	4098	7003
Universidad del Sureste	6	11	17
Universidad Popular de la Chontalpa	34	45	79
Instituto Tecnológico Superior de Centla	4	2	6
Instituto Tecnológico Superior de Comalcalco	88	100	188
Instituto Tecnológico Superior Campus Huimanguillo	1	2	3
Universidad Politécnica del Golfo de México	2	4	6
Universidad Tecnológica de Tabasco	449	457	906
Universidad Politécnica Mesoamericana	2	1	3
Universidad IEU	0	2	2
Universidad La Salle	2	1	3
Universidad Intercultural del Estado de Tabasco	8	17	25
Universidad Autónoma de Guadalajara Campus Tabasco	0	2	2
Instituto Tecnológico Superior de la Sierra	1	5	6
Instituto Tecnológico Superior de los Ríos	1	1	2
Instituto Tecnológico Superior de Villa La Venta	48	62	110
Universidad de México	1	0	1
Instituto tecnológico Superior Zona Olmeca	1	0	1
Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias	6	6	12
Colegio de la Frontera Sur	5	1	6
Instituto Universitario Puebla	29	61	90
Instituto Universitario de Yucatán Campus Tabasco	18	21	39
Colegio de Postgraduados Campus Tabasco	19	17	36
Universidad Tecnológica del Usumacinta	61	85	146

Instituto Tecnológico Superior de Macuspana	15	12	27
Universidad Valle del Grijalva	3	0	3
Escuela Normal de Educación Preescolar "Rosario María Gutiérrez"	0	1	1
Universidad Olmeca	1	3	4
Instituto Tecnológico y de Estudios Superiores de Occidente	0	1	1
Universidad CINDEHU	3	4	7
No especificó	11	13	24
Total	**3724**	**5035**	**8759**
Total, general (%)	**42.51%**	**57.48%**	**100%**

Nota: Elaboración propia con información de PRODECTI (2020).

De acuerdo con los datos mostrados en la tabla 1 los estudiantes encuestados se encontraban inscritos en diversas universidades del estado de Tabasco haciendo un total de 30 instituciones y uno con datos que no especificaron de manera correcta en el formulario de la encuesta aplicada, participaron 3724 personas del sexo masculino y 5035 del sexo femenino que corresponden al 42.51% y 57.48% respectivamente.

3.1.2 Diseño y Descripción del Instrumento

Para la elaboración del Dataset se aplicó una encuesta elaborada con la participación de investigadores de diversas Divisiones Académicas de la UJAT. Se utilizó la base de datos del proyecto de investigación financiado por el Consejo de Ciencia y Tecnología del Estado de Tabasco (CCYTET) denominado “Diagnóstico Participativo Post Covid 19 en Tabasco”. Con realización del instrumento aplicado a estudiantes de educación superior se podrá determinar el desempeño académico mediante patrones que influyeron a tomar ciertas decisiones que se vieron reflejadas en un aprovechamiento, deserción o en el peor de los casos la reprobación.

La estructura del cuestionario cuenta con 6 preguntas relacionadas con su aprendizaje y desempeño académico de igual forma con 17 cuestionamientos de tipo sociodemográficos del alumno como la institución, división académica, tipo de institución, género, edad, semestre, municipio, localidad, trabaja, beca, tipo de beca, número de personas en casa, baños, automóviles, internet, de las personas que viven en casa cuántas trabajan y número de cuartos, haciendo un total de 23 preguntas siendo 5 de ellas preguntas abiertas.

Cabe destacar que este cuestionario fue aplicado de forma virtual mediante un vínculo para acceder al formulario en *Google forms.* En la siguiente tabla se presentan las preguntas clasificadas (ver tabla 4).

Tabla 4

Clasificación de Preguntas del Dataset.

Criterios	Preguntas	No. de preguntas
Datos del alumno	1. Institución 2. División académica 3. Tipo de institución 4. Genero 5. Edad 6. Semestre 7. Municipio 8. Localidad o colonia 9. Trabaja 10. Beca 11. Tipo de beca	11
Aspectos sociodemográficos	1. Número de personas en casa 2. Número de baños 3. Número de automóviles 4. Internet	6

	5. Personas que viven en casa y que trabajan 6. Número de cuartos	
Relacionadas con el desempeño académico	1. Aprendió lo mismo en las clases virtuales que en las clases presenciales. 2. Los profesores les ofrecieron alternativas de trabajo si no contaba con los recursos (internet o computadora) para realizar sus tareas y actividades. 3. Por problemas derivados de la contingencia, tuvo que darse de baja del ciclo escolar. 4. A pesar de la contingencia, consideró que pudo aprobar todas las asignaturas del ciclo escolar. 5. las formas de evaluación fueron las adecuadas. 6. El cambio de modalidad afectó la calidad de la educación en tu entorno inmediato (amigos, vecinos, familiares, etc.).	6

Nota: Elaboración propia con información de PRODECTI (2020).

Las preguntas relacionadas con el desempeño académico se encuentran en la parte 3 de criterios de la tabla 2 haciendo un total de 6 ítems cabe destacar que cada una de esas preguntas fueron respondidas con una escala de opción múltiple tipo Likert, en otras secciones fueron rellenadas por el alumno dependiendo de las necesidades de la pregunta.

3.1.3 Exploración de los Datos

Para la confiabilidad de los datos se destaca que la fuente de información de donde se obtuvieron los datos es confiable y exacta. Los datos deben mantenerse sin corrupción y no deben ser alterados de manera no autorizada para garantizar la integridad de los datos.

El Dataset se obtuvo en el formato estándar de Excel .xlsx para utilizarlo en la hoja de cálculo de Excel, una vez realizado eso se procedió a aplicar filtros para conocer los datos y así comenzar con la limpieza y transformación. Cabe destacar que los datos se exploraron con la finalidad de conocer la calidad y tipo de errores tal como palabras mal escritas en las preguntas abiertas o palabras no acorde con el campo específico. En la siguiente figura se muestra un tipo de error (ver figura 1).

Figura 1

Errores de Datos no Acorde a lo que se Pedía.

654	Universidad Juárez Autónoma de Tabasco	División A	1	
655	Universidad Juárez Autónoma de Tabasco	División A	1	
656	Ujat	División a	1	
657	Universidad Juárez Autónoma de Tabasco	División A	1	
658	Universidad Juárez Autónoma de Tabasco	Division A	1	
659	Pemex	División a	1	
660	Ujat	Dacyti	1	
661	Universidad Juárez Autónoma de Tabasco	División A	1	
662	Universidad Juárez Autónoma de Tabasco	DAIA, Lic.	1	
663	Universidad Juárez Autónoma de Tabasco	División A	1	
664	Universidad Juárez Autónoma de Tabasco	División d	1	

Nota: Elaboración propia.

Estos errores se deben a que no comprendieron la pregunta lo que provocó que hubiera desfases en la columna por lo que en la etapa de limpieza de datos se corrigieron, de igual manera al ser muchas universidades se exploraron y se etiquetaron con su respectivo nombre para tener un mejor orden.

3.2 Análisis y Selección de Herramientas de Minería de Datos

Para medir la calidad de un sistema existen varios modelos como el FURPS, modelo de calidad BOHEMN o las normas ISO que permiten realizar el análisis del *software* en el cual cada modelo cuenta con sus características que sirven para evaluar el *software* mediante diferentes criterios y métricas así lo apunta Callejas-Cuervo y Alarcón-Aldana (2017) existen diferentes modelos de calidad del *software* es importante conocer los conceptos y características acerca de la calidad de *software* de igual manera su estructura y enfoque ya que permitirá contar con los requisitos funcionales necesarios y un rendimiento adecuado del sistema a ejecutar.

Para seleccionar la herramienta que se utilizó en esta investigación se ejecutó el modelo FURPS el cual utiliza ingeniería de *software* para describir los requisitos de un sistema. Este modelo cuenta con cinco factores y sus respectivos criterios en los que se avaluaran las diferentes herramientas (ver tabla 5).

Tabla 5

Modelo FURPS desglosado.

Factores	Criterios
Funcionabilidad	Adaptabilidad Exactitud Interoperabilidad Seguridad
Usabilidad	Comprensibilidad Aprendizaje Operatividad Atractivo
Confabilidad *Reliability*	Análisis Cambio Estabilidad Prueba Madurez Tolerancia a fallos Recuperabilidad
Desempeño *Performance*	Comportamiento del tiempo Uso de los recursos
Soporte	Adaptabilidad Instalación Coexistencia Remplazo

Nota: Elaboración propia con información de Constanzo (2014)

Evaluar las diferentes herramientas de minería de datos permitió dar un respaldo y sustento a las herramientas seleccionadas ya que expone cuál fue el proceso de selección de estas y así agregarle más valor metodológico a la investigación.

3.2.1 Procedimiento para Evaluar las Herramientas de Minería de Datos

Para efectos de esta investigación es indispensable conocer las mejores herramientas en las que se trabajaron se tomó de guía el trabajo de Guzmán (2020) y

Azuara (2023) para el análisis considerando los criterios de evaluación y otorgando una calificación del cero al 23 en la cual el 23 representa que la herramienta cumple con todos los atributos evaluados del modelo FURPS. En la siguiente tabla (ver tabla 6) se muestran los resultados de la evaluación.

Tabla 6

Herramientas Utilizadas para Minería de Datos.

Factores	Herramientas utilizadas para minería de datos															
	Rapid miner	Orange	WEKA	Lenguaje R	Python	Matlab	Microsoft SQL server	Microsoft excel	Microsoft análisis service	Oracle Bussines Intelligence Server	SPSS Modeleer	DAAL	Amazon EMR	Google Cloud Data	Microsoft Azure Analysis Services	SAS Enterprise Miner
Funcionalidad	3	3	3	3	2	2	3	3	3	3	3	2	3	3	3	3
Usabilidad	3	2	4	4	2	2	3	3	4	2	3	1	3	3	4	3
Confiablidad	2	2	4	4	1	1	3	3	3	3	2	0	1	2	4	2
Rendimiento	3	4	5	4	4	4	5	5	5	5	4	2	5	5	5	5
Capacidad de soporte	5	5	5	6	3	3	6	5	3	4	4	3	6	3	6	5
Total	**16**	**16**	**21**	**21**	**12**	**12**	**20**	**19**	**18**	**17**	**16**	**8**	**18**	**16**	**22**	**18**

Nota: Guzmán (2020).

Después de realizar el conteo de los criterios de las herramientas se suman y se tiene un total de puntaje por cada herramienta de acuerdo con cada factor que propone el modelo: funcionalidad, usabilidad, confiabilidad, rendimiento y soporte.

De acuerdo con Azuara (2023) las 16 herramientas clasificadas en la tabla 3 (ver tabla 7) hay herramientas que son de acceso libre y con licencia de pago por uso. Para dividir esas diferencias se clasificaron por tipo: gratuita, comerciales y herramientas de programación (ver tabla 7).

Tabla 7

Clasificación de Herramientas por Acceso Libre y Licencia de Pago.

Herramientas de minería	Factores del modelo FURPS					
	Funcionalidad	Usabilidad	Confiabilidad	Rendimiento	Capacidad de soporte	Calificación
Herramientas de uso gratuitas						
Orange	3	2	2	4	5	16
Weka	3	4	4	5	5	21
Programación						
Lenguaje R	3	4	4	4	6	21
Python	2	2	1	4	3	12
Herramientas de pago por licencia						
RapidMiner	3	3	2	3	5	16
Microsoft Excel	3	3	3	5	5	19
SAS Enterprise Miner	3	3	2	5	5	18
SPSS Modeler	3	3	2	4	4	16
Google Cloud Datalab	3	3	2	5	3	16
Microsoft SQL Server	3	3	3	5	6	20
Microsoft Analisys Services	3	4	3	5	3	18
Oracle Bussinnes Inteligence Server	3	2	3	5	4	17
Amazon EMR	3	3	1	5	6	18
Microsoft Azure Analysis Service	3	4	4	5	6	22
Programación						
DAAL	2	1	0	2	3	8
Matlab	2	2	1	4	3	12

Nota: Elaboración propia con base en datos de Azuara (2022).

Al realizar la ponderación de las herramientas clasificadas por gratuitas, de pago de licencia y lenguaje de programación, se optó por las herramientas de uso gratuito de acuerdo con Azuara (2023) las cuatro herramientas que contempló dentro de la investigación fueron Orange, Weka y Lenguaje R y Python de estos *softwares* mencionados se utilizó la de mejor calificación.

Después de evaluar y clasificar las herramientas mediante el modelo *FURPS* se determinó que la herramienta seleccionada para implementar técnicas de minería de

datos en esta investigación fue WEKA por ser la mejor ponderada dentro de las herramientas de uso gratuitas e interactivas de acuerdo con la tabla anterior.

3.2.2 Descripción de las Herramientas Seleccionadas

WEKA

Waikato Environment for Knowledge Analysis (*WEKA*), es una herramienta de *software* de código abierto popular y ampliamente utilizada para el aprendizaje automático y la minería de datos. Desarrollado en la Universidad de *Waikato* en Nueva Zelanda, de acuerdo con Martínez (2017) esta plataforma de *software* provee una colección de algoritmos para el preprocesamiento, filtrado y minería de datos, ya que proporciona un conjunto completo de herramientas y algoritmos para diversas tareas relacionadas con análisis de datos, clasificación, regresión, agrupamiento, reglas de asociación, selección de características y más. Está diseñado para ser fácil de usar tanto para principiantes como para científicos de datos experimentados.

La interfaz gráfica de usuario de *Weka* ofrece una interfaz gráfica fácil de usar que permite a los usuarios interactuar con datos, preprocesarlos, aplicar algoritmos de aprendizaje automático y visualizar los resultados. Esto lo hace accesible a usuarios con distintos niveles de experiencia técnica.

En general, es una herramienta versátil y poderosa para el análisis de datos y el aprendizaje automático que ha sido ampliamente adoptada en el mundo académico y la industria para tareas que van desde fines educativos hasta proyectos de minería de datos

del mundo real. Su interfaz fácil de usar y su amplio conjunto de funciones lo convierten en un activo valioso para cualquier persona interesada en trabajar con datos.

3.3 Metodología del Proceso KDD

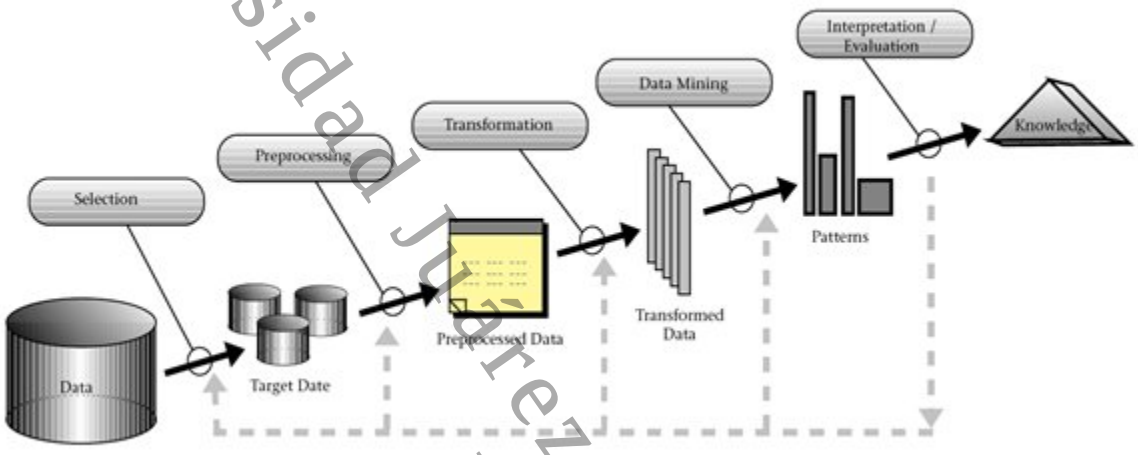
La minería de datos es un proceso que implica descubrir patrones, tendencias y conocimientos útiles a partir de conjuntos de datos. Para llevar a cabo este proceso de manera efectiva, se han utilizado diferentes metodologías y enfoques. Aquí se presentan algunas de las metodologías más comunes utilizadas en la minería de datos: *Semma (Sample, Explore, Modify, Model, Assess), CRISP-DM (Cross-Industry Standard Process for Data Mining)* y *KDD (Knowledge Discovery in Databases).*

Para efectos de esta investigación se ocupó el proceso *KDD* que en si es una metodología y proceso integral amplio para descubrir patrones útiles, información y conocimientos previamente desconocidos a partir de grandes conjuntos de datos. Esta metodología no se limita únicamente a la minería de datos, sino que engloba todo el proceso de descubrimiento de conocimiento, del cual la minería de datos es solo una parte.

En la siguiente figura se muestra de forma gráfica los procesos que se consideran durante la aplicación de minería de datos en esta investigación con la metodología *KDD*. (Ver figura 2).

Figura 2

Proceso KDD.



Nota: Vázquez y Hispass (2014).

De acuerdo con Vázquez y Hispass (2014) todas las fases del proceso se relacionan entre ellas en un flujo secuencial no rígido en el cual las flechas representan las dependencias más importantes y frecuentes. Aunado a lo anterior la figura 2 muestra los seis procesos de la metodología original que se llevaron a cabo esta investigación y que fueron descritos de acuerdo con las actividades correspondientes.

3.3.1 Selección.

En la primera etapa de la metodología llamada selección se obtuvo el conjunto de datos para comenzar con la ejecución del proceso *KDD*. En el apartado de información general se describió como se obtuvo el Dataset el cual se integró efectuando una encuesta en el año 2020 por medio de formularios de Google enviado a estudiantes del

ciclo enero-junio 2020, adscritos a 65 organizaciones, de las cuales 47 fueron IES públicas o privadas y 12 diversas organizaciones del sector público privado.

De acuerdo con el total de registro recabados se obtuvo una población de 8759 estudiantes de IES del estado de Tabasco, México de todas las áreas de conocimiento y ciclos escolares de los diferentes programas de nivel licenciatura y posgrado.

3.3.2 Preprocesamiento.

En esta etapa de limpieza se analizó y se estudió el Dataset como se mencionó en el punto 3.1.3 nombrada exploración de los datos, esto con el fin de ver y verificar la calidad y estructura de los datos los tipos de errores se clasificaron de la siguiente manera:

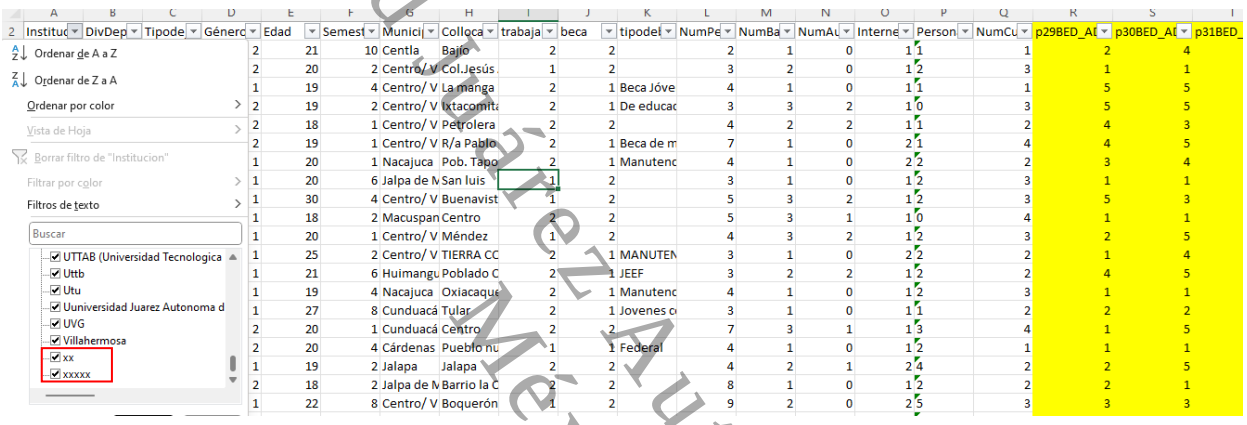
- Datos nulos
- Datos duplicados
- Datos perdidos
- Campos vacíos
- Información fuera de rango

Para mejorar la calidad de la información se remplazaron esos errores por valores más próximos. Reyes y García (2005) mención que en esta etapa se debe reducir el ruido del Dataset y ver como remplazar los datos faltantes en esta etapa de limpieza todos estos valores se ignoran, se reemplazan por un valor por omisión, o por el valor más cercano, es decir, se usan métricas de tipo estadístico como media, moda, mínimo y máximo para reemplazarlos. A partir de los tipos de errores mencionados anteriormente, se realizó la limpieza del Dataset retomando los aspectos más relevantes.

Al abrir el Dataset en formato .xlsx (formato de hoja de cálculo de Excel) se procedió a revisar la revisar el número de hojas del Dataset e integridad de los datos seguidamente se aplicó un filtro para revisar las variables que contenía cada pregunta (ver figura 3).

Figura 3

Revisión de Variables que tenía el Dataset.



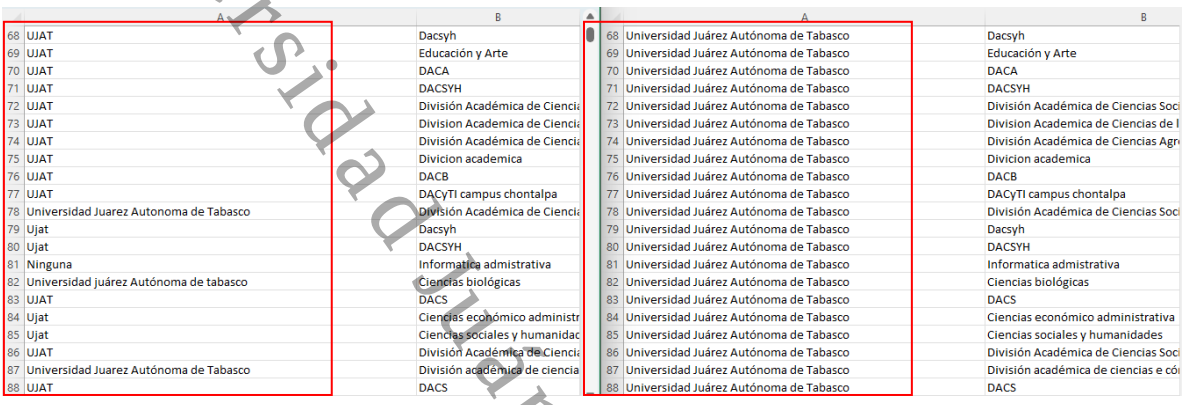
Nota: Elaboración propia.

Como se muestra en la figura 3 al realizar filtro todos los campos del Dataset y revisar el primero llamado institución se cumple los tipos de errores mencionados previamente se puede notar que en ese campo existe información fuera de rango.

Debido a lo anterior se procedió a resolver todas las inconsistencias en ese apartado llamado institución lo cual permitió etiquetar esos valores con información más clara y valida con respecto a sus demás respuestas que permitiera entender mejor la información. En la siguiente figura se compara el campo antes y después de realizar la resolución de problemas (ver figura 4).

Figura 4

Corrección de Problemas en el Dataset.



	A	B
68	UJAT	Dacsyh
69	UJAT	Educación y Arte
70	UJAT	DACA
71	UJAT	DACSYH
72	UJAT	División Académica de Ciencia
73	UJAT	Division Academica de Ciencia
74	UJAT	División Académica de Ciencia
75	UJAT	Divicion academica
76	UJAT	DACB
77	UJAT	DACyTI campus chontalpa
78	Universidad Juarez Autonoma de Tabasco	División Académica de Ciencia
79	Ujat	Dacsyh
80	Ujat	DACSYH
81	Ninguna	Informatica admistrativa
82	Universidad juárez Autónoma de tabasco	Ciencias biológicas
83	UJAT	DACS
84	Ujat	Ciencias económico administr
85	Ujat	Ciencias sociales y humanidac
86	UJAT	División Académica de Ciencia
87	Universidad Juarez Autónoma de Tabasco	División académica de ciencia
88	UJAT	DACS

	A	B
68	Universidad Juárez Autónoma de Tabasco	Dacsyh
69	Universidad Juárez Autónoma de Tabasco	Educación y Arte
70	Universidad Juárez Autónoma de Tabasco	DACA
71	Universidad Juárez Autónoma de Tabasco	DACSYH
72	Universidad Juárez Autónoma de Tabasco	División Académica de Ciencias Soci
73	Universidad Juárez Autónoma de Tabasco	Division Academica de Ciencias de l
74	Universidad Juárez Autónoma de Tabasco	División Académica de Ciencias Agr
75	Universidad Juárez Autónoma de Tabasco	Divicion academica
76	Universidad Juárez Autónoma de Tabasco	DACB
77	Universidad Juárez Autónoma de Tabasco	DACyTI campus chontalpa
78	Universidad Juárez Autónoma de Tabasco	División Académica de Ciencias Soci
79	Universidad Juárez Autónoma de Tabasco	Dacsyh
80	Universidad Juárez Autónoma de Tabasco	DACSYH
81	Universidad Juárez Autónoma de Tabasco	Informatica admistrativa
82	Universidad Juárez Autónoma de Tabasco	Ciencias biológicas
83	Universidad Juárez Autónoma de Tabasco	DACS
84	Universidad Juárez Autónoma de Tabasco	Ciencias económico administrativa
85	Universidad Juárez Autónoma de Tabasco	Ciencias sociales y humanidades
86	Universidad Juárez Autónoma de Tabasco	División Académica de Ciencias Soci
87	Universidad Juárez Autónoma de Tabasco	División académica de ciencias e cór
88	Universidad Juárez Autónoma de Tabasco	DACS

Nota: Elaboración propia.

En esta etapa del proceso KDD se corrigieron alrededor de 40 datos que estaban fuera de rango sin eliminar ningún registro del Dataset haciendo el mismo total de 8759 estudiantes.

3.3.3 Transformación.

Esta etapa consistío en asignar categorías rangos, simplificando valores, normalizando o etiquetando datos, dependiendo de lo que se necesite se realizan los cambios así lo menciona Carracedo y Terrádez (2016) que no siempre los datos se encuentran en la forma más adecuada para poder aplicar las técnicas y algoritmos correctos. Se concluyó que de las 23 variables de Dataset a las únicas que se le realizó un cambio de etiqueta fue a la variable número tres, ocho, y a las últimas seis que corresponden de la variable 18 a la 23 con la finalidad de tener mejor accesibilidad y

entendimiento al proyectar el Dataset con cualquier herramienta de minería. En la siguiente tabla se muestran los cambios realizados (ver tabla 8).

Tabla 8

Cambios de Variables Realizados al Dataset.

No.	Variable	Nuevo nombre
1	Institución	
2	DivDepCampus	
3	Tipode_institucion	Tipo_institución
4	Género	
5	Edad	
6	Semestre	
7	Municipio	
8	Collocalidad	Colonia_localidad
9	Trabaja	
10	Beca	
11	Tipodebeca	
12	NumPersonasCasa	
13	NumBaños	
14	NumAutomoviles	
15	Internet	
16	Personasviventrabajan	
17	NumCuartos	
18	p29BED_ADR	P_1
19	p30BED_ADR	P_2
20	p31BED_ADR	P_3
21	p32BED_ADR	P_4
22	p33BED_ADR	P_5
23	p34BED_ADR	P_6

Nota: Elaboración propia.

3.3.4 Minería de Datos.

Para la ejecución de esta etapa se recurrió a elegir las técnicas y algoritmos de minería de datos que se van a implementar para analizar el Dataset. Debido al interés de conocer el aprovechamiento, deserción y reprobación de estudiantes de educación superior existen relaciones entre variables por lo cual se emplearon las siguientes

técnicas. De acuerdo con aportaciones de Hernández *et ál.,* (2004), Carracedo y Hernández (2016) se describen a continuación.

- Reglas de asociación: Identifica las asociaciones frecuentes, que ayudan a mostrar la probabilidad de las relaciones entre los elementos de datos.
- Arboles decisión. La idea de esta técnica utiliza una estructura de tipo árbol para representar y categorizar decisiones basadas en datos, en sí divide el conjunto de datos en subconjuntos más pequeños y homogéneos de manera recursiva, tomando decisiones basadas en características específicas de los datos.
- Agrupamiento. Esta técnica agrupa datos formando un conjunto de datos con base en variables o características similares.

Como se ha mencionado a lo largo de la investigación la minería de datos tiene como objetivo analizar los datos para extraer conocimiento en forma de relaciones o patrones. De acuerdo con Hernández *et ál*., (2004) los modelos pueden ser de dos tipos:

- Predictivos: pretenden estimar valores futuros o desconocidos
- Descriptivos: identifican patrones que explican o resumen los datos.

Cabe mencionar que existen dos tipos de clasificación para las técnicas y algoritmos de minería de datos que se describen a continuación.

- Supervisadas: se entrena el algoritmo mediante los datos previamente cargados con el Dataset para actuar como guía.
- No supervisadas: no existe el conjunto de datos de entrenamiento y se entra al problema de manera directa.

En la siguiente tabla (ver tabla 9) se especifican las características de las técnicas usadas dentro de la investigación, cabe destacar que el algoritmo de clasificación (reglas de asociación) es la técnica que ayuda a predecir eventos futuros del fenómeno mientras que las restantes describirán a fondo lo que pasa actualmente al analizar la información del Dataset.

Tabla 9

Técnica, Modelo y Algoritmo Usado en esta Investigación.

Técnica	Tipo de modelo	Clasificación	Algoritmo usado
Arboles de decisión (Clasificación)	Predictivo	Supervisado	J48
Agrupación (*Clustering*)	Descriptivo	No supervisado	SimpleKMeans
Reglas de asociación (*Association Rules*)	Descriptivo	No supervisado	Apriori

Nota: Elaboración propia.

Weka

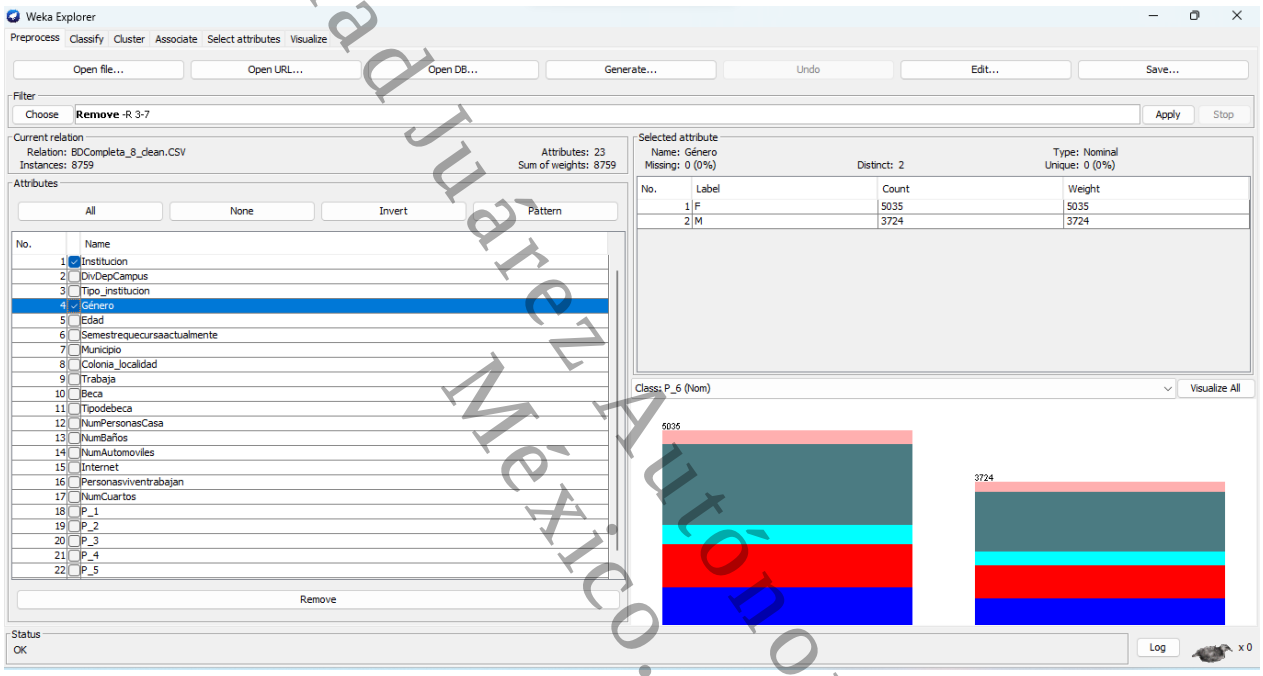
Este apartado presenta la ejecución de los algoritmos con la herramienta Weka. Se inició con la carga e inicio de la aplicación para ver el entorno de trabajo, en la siguiente (ver figura 5) se muestra la pantalla principal al cargar el Dataset, la cual muestra el nombre de las variables y su representación gráfica de forma individual o en conjunto.

Se puede notar que en cada apartado se muestra información relevante, en el caso de *Selected attribute* el cual describe cada variable seleccionada y muestra su nombre, el tipo de variable, los valores perdidos entre otros datos, lo que es de mucha utilidad al

aplicar los algoritmos ya que permite estar al tanto de todos los valores que pueden influir en los resultados finales.

Figura 5

Carga de Datos en WEKA.

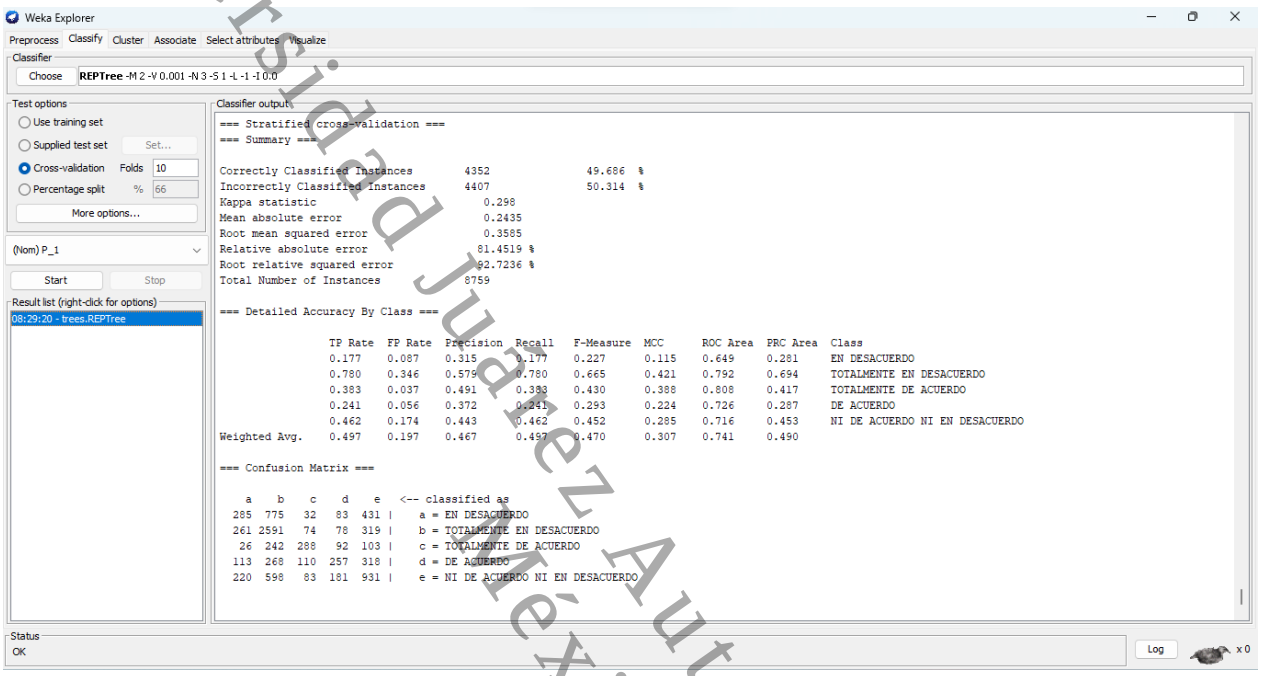


Nota: Elaboración propia.

Para continuar se procedió con la ejecución de los algoritmos comenzando con la técnica de árbol de decisión en el cual se seleccionó en la pestaña *Classify, Choose, trees, REPtree* cabe destacar que para ejecutar diferentes algoritmos y técnicas hay que hacer uso de filtros para seleccionar los índices que se utilizaran del contrario pues no se podrá marcar el algoritmo para ejecutarlo en su apartado correspondiente. En la siguiente figura se muestra la ejecución de la técnica árbol de decisión (ver figura 6).

Figura 6

Aplicación de Árbol de Decisión en WEKA.



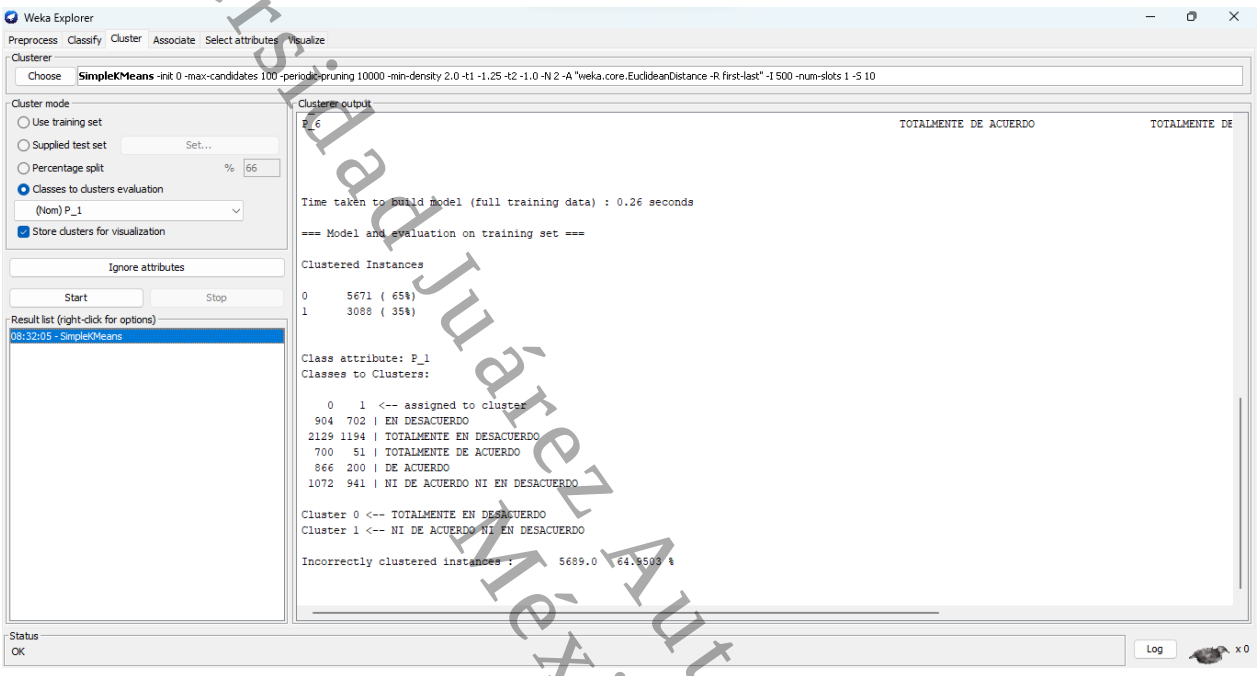
Nota: Elaboración propia.

Al ejecutar el algoritmo se visualiza los datos que genera la técnica de árbol de decisión, así como la clasificación de sus datos, métricas y matriz de confusión.

Para continuar se ejecutó la técnica de agrupación (*Clustering*) en el cual se seleccionó la pestaña *Cluster* mediante el algoritmo *SimpleKmeans*. Al presionar en el apartado de *Choose* se puede configurar distintas partes del modelo como el método de inicialización, el número de *Cluster* e iteraciones que se van a crear con la herramienta, en la siguiente figura se muestras los resultados de las pruebas ejecutadas ver figura 7.

Figura 7

Aplicación de Clustering en WEKA.

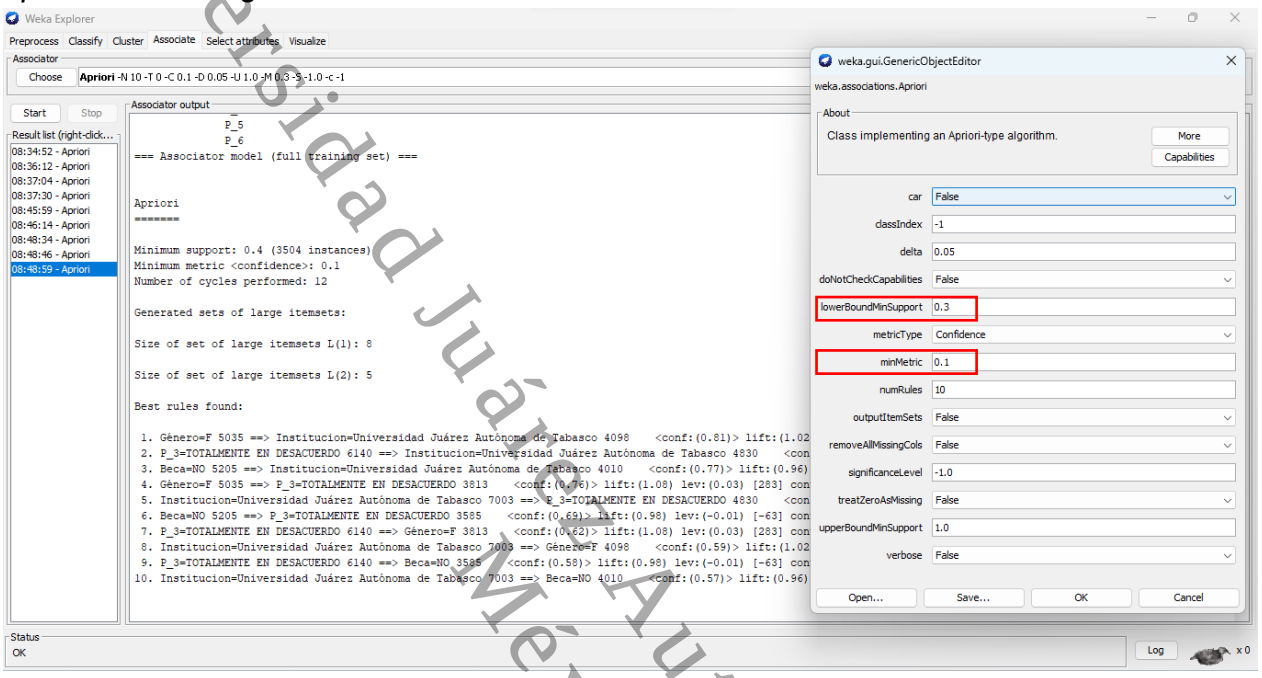


Nota: Elaboración propia.

Para finalizar se ejecutó la técnica de reglas de asociación mediante el algoritmo *Apriori* el cual se seleccionó en la pestaña de *Associate*, cabe destacar que al ejecutar la técnica con sus variables correspondientes el algoritmo no mostraba ninguna regla, pues siempre hay que tener en cuenta los valores de soporte y confianza del algoritmo, luego de probar con diferentes valores si mostró las reglas encontradas para dichos valores, en la siguiente figura se muestra la ejecución de la técnica con sus resultados ver figura 8.

Figura 8

Aplicación de Reglas de Asociación en WEKA.



Nota: Elaboración propia.

3.3.5 Interpretación Evaluación.

De acuerdo con las pruebas realizadas para la obtención de los patrones de aprovechamiento, deserción y reprobación fue necesario retomar los pasos anteriores del proceso *KDD* en la etapa especifica de preprocesamiento y transformación, se tuvieron que reorganizar los datos para tener mejores resultados, de igual forma los algoritmos se repitieron algunas veces por errores de datos.

Aunado a lo anterior se concluye que en esta etapa se refleja un proceso en donde se hacen muchas pruebas y experimentaciones realizadas con la herramienta de minería de datos con la finalidad de obtener resultados definidos y legibles que permitan tener

una buena interpretación de la información obtenida para los usuarios finales de esta investigación.

3.3.6 Conocimiento Obtenido.

Cómo se mencionó previamente Hernández *et ál*., (2004) apunta que los modelos descriptivos permiten identificar patrones que explican o resumen los datos. Por lo cual en esta investigación se aplicó minería de datos descriptiva, que permitió obtener un conocimiento relevante en el Dataset y de utilidad para la descripción de reportes y toma decisiones informadas que permitan describir la problemática planteada con respecto al aprovechamiento, deserción y reprobación en estudiantes de educación superior en Tabasco por Covid 19.

Capítulo 4. Resultados

Una vez concluida la obtención de los datos y el análisis de la información, se procedió a la interpretación de resultados los cuales se presentan a continuación.

En el apartado anterior se desglosó la metodología del proceso *KDD* en el cual se hizo la transformación y pasos referentes a la interpretación y uso del conocimiento, en este capítulo se realizaron con detalle las pruebas con las técnicas y algoritmos correspondientes con los que se obtuvieron resultados confiables que permitieron responder la pregunta de investigación y exponer el conocimiento descubierto.

Este capítulo se encuentra dividido por dos apartados, en el primero se muestran las pruebas realizadas con la herramienta Weka con las técnicas de clasificación (árboles de decisión), agrupación y reglas de asociación que fueron las seleccionadas para ejecutar las técnicas, y finalmente en el apartado dos se describen la interpretación de los patrones encontrados.

4.1 Pruebas realizadas con la herramienta de minería de datos.

A continuación, se muestran las pruebas realizadas mediante la herramienta Weka para la identificación de los patrones con el orden siguiente: técnica de árboles de decisión, clasificación y agrupamiento, cabe destacar que se realizaron tres pruebas por técnica de minería de datos.

4.1.1 Weka.

Clasificación Árboles de decisión.

Mediante la herramienta *Weka* se realizaron pruebas para interactuar con las técnicas, algoritmos y gráficos que proporciona, para comenzar se optó por usar el algoritmo *J48* ya que de acuerdo con Guzmán (2020) este algoritmo ofrece mejores resultados, de este modo se realizó la presentación mediante este algoritmo.

Para realizar las pruebas se consideraron diferentes parámetros para ejecutar el algoritmo, por ejemplo para el máximo de ramificaciones con un valor mínimo de 50 y máximo de 100, el cual permite conocer el largo del diagrama, de igual forma el porcentaje de confidencia que entre menor número sea, mayor proporción de clasificación correcta habrá, se utilizó un porcentaje de 10% al 25%, es necesario encontrar un equilibrio entre la confidencia, largo del diagrama y el porcentaje de clasificación para obtener información valiosa.

A continuación, se presentan las tres pruebas realizadas mediante la técnica de árbol de decisión con el algoritmo J48 con las variables Internet, P_3 y Beca como clase principal e identificación del modelo. Se utilizó el porcentaje de entrenamiento del 66% que es el que utiliza la herramienta Weka.

Weka permite visualizar los parámetros generados en la pantalla principal, así mismo dentro de las métricas que genera se muestra la matriz de confusión, el porcentaje de clasificación correcta, el porcentaje de precisión por cada clase que genera y el modelo que ejecuta.

En la figura 9 se presentan los resultados obtenidos con la primera prueba realizada con la variable Internet como clase principal seguidamente Tipo_Institucion, Género, Edad, Internet, P_2 y P_3 como variables de apoyo. Se observa que el 61.57% de las instancias se clasificaron correctamente lo que es igual a 5393 de las cuales 5340 pertenecen a la clase “a” y 53 a la clase “b” del Dataset.

Figura 9

Resumen de Métricas de Evaluación de Árbol de Decisión con Pregunta Raíz Internet.

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

Correctly Classified Instances        5393               61.571  %
Incorrectly Classified Instances      3366               38.429  %
Kappa statistic                          0.0042
Mean absolute error                      0.4714
Root mean squared error                  0.4866
Relative absolute error                 99.7672 %
Root relative squared error            100.1231 %
Total Number of Instances             8759

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.988     0.984     0.618       0.988   0.760      0.014    0.513     0.627     SI
               0.016     0.012     0.442       0.016   0.031      0.014    0.513     0.393     NO
Weighted Avg.  0.616     0.612     0.551       0.616   0.481      0.014    0.513     0.538

=== Confusion Matrix ===

    a    b   <-- classified as
 5340   67 |    a = SI
 3299   53 |    b = NO
```

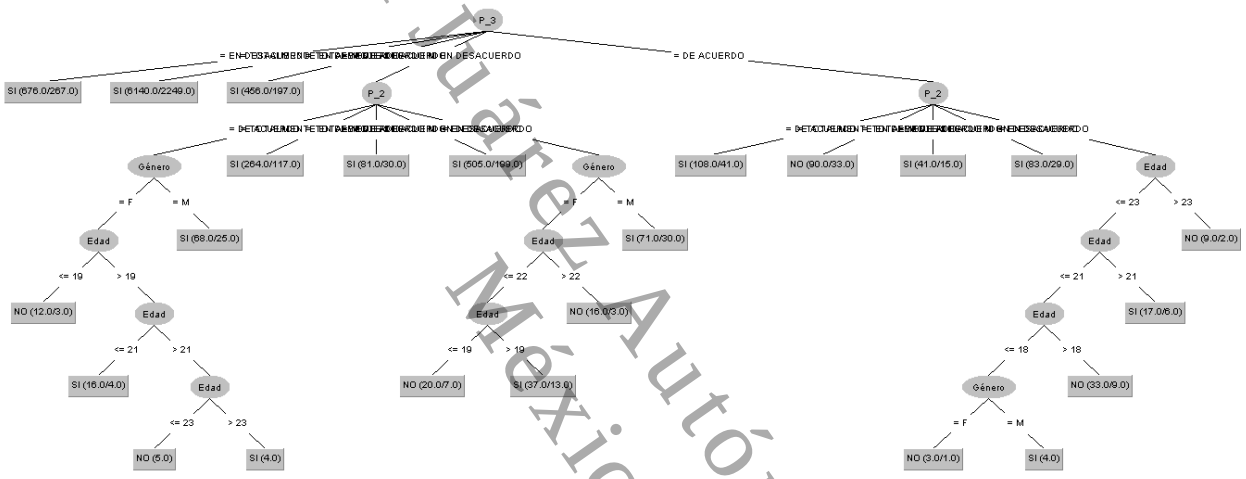
Nota: Elaboración propia.

Como se puede ver el porcentaje correcto de instancias clasificadas fue de 61.57% porque es un modelo con métricas aceptables y se va a considerar para la interpretación de patrones dependiendo de los resultados de las pruebas siguientes.

En la siguiente figura 10 se muestra el árbol de decisión generado de acuerdo con todos los parámetros establecidos y el conjunto de datos. El árbol tiene un total de 14 nodos y 24 hojas, por lo cual se presenta el diagrama completo.

Figura 10

Visualización del Diagrama de Árbol con Pregunta Raíz Internet.



Nota: Elaboración propia.

De acuerdo con el diagrama de árbol se puede decir que la clasificación “a” hace referencia al ítem sí mientras que la clasificación “b” hace referencia al ítem no. Con la información obtenida de esta prueba se puede interpretar inicialmente en los patrones obtenidos de acuerdo con la pregunta: Internet ¿cuenta con internet en casa?

El árbol muestra un balance en que los estudiantes no se dieron de baja del ciclo escolar por problemas derivados de la pandemia pues los profesores ofrecieron alternativas de trabajo si no contaban con los recursos para realizar sus actividades.

Cabe destacar que los estudiantes encuestados que respondieron a no darse de baja y que los profesores no les ofrecieron alternativas de trabajo sí contaban con internet en sus hogares.

La segunda prueba realizada fue con la variable P_3 como clase principal y Tipo_institución, Género, Edad, Trabaja, Beca, Internet, P_1, P_2, P_3 y P_6 como variables de apoyo, con la cual se obtuvo un 72% de instancias clasificadas correctamente las cuales equivalen a 6370 (ver figura 11).

Figura 11

Resumen de Métricas de Evaluación de Árbol de Decisión con Pregunta Raíz P_3.

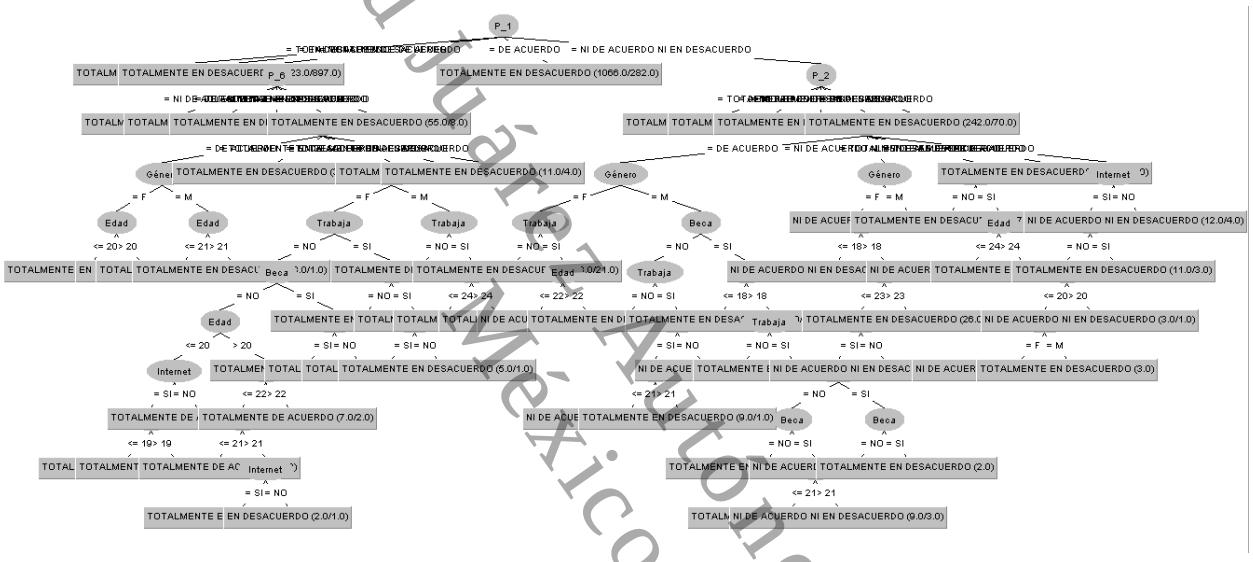
```
=== Summary ===

Correctly Classified Instances        6307               72.0059 %
Incorrectly Classified Instances      2452               27.9941 %
Kappa statistic                          0.18
Mean absolute error                      0.1767
Root mean squared error                  0.2996
Relative absolute error                 91.5657 %
Root relative squared error             96.4762 %
Total Number of Instances             8759

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.001    0.000    0.250      0.001    0.003      0.014    0.604     0.104     EN DESACUERDO
               0.967    0.840    0.730      0.967    0.832      0.227    0.594     0.754     TOTALMENTE EN DESACUERDO
               0.232    0.010    0.573      0.232    0.331      0.344    0.681     0.238     TOTALMENTE DE ACUERDO
               0.237    0.022    0.612      0.237    0.341      0.332    0.685     0.324     NI DE ACUERDO NI EN DESACUERDO
               0.003    0.001    0.167      0.003    0.005      0.016    0.580     0.061     DE ACUERDO
Weighted Avg.  0.720    0.592    0.645      0.720    0.644      0.221    0.610     0.592

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
    1  655    3   16    1 |    a = EN DESACUERDO
    2 5939   64  131    4 |    b = TOTALMENTE EN DESACUERDO
    0  346  106    4    0 |    c = TOTALMENTE DE ACUERDO
    0  834    5  260    0 |    d = NI DE ACUERDO NI EN DESACUERDO
    1  365    7   14    1 |    e = DE ACUERDO
```

Nota: Elaboración propia.

El porcentaje de precisión fue del 72% por lo cual las métricas indican que este modelo es aceptable. A continuación, se presenta el diagrama de árbol generado con un total de 45 nodos y 61 hojas (ver figura 12).

Figura 12

Visualización del Diagrama de Árbol con Pregunta Raíz P_3.



Nota: Elaboración propia.

De acuerdo con el diagrama de árbol en este ejemplo la pregunta objetivos es: P_3 ¿Por problemas derivados de la contingencia tuve que darme de baja del ciclo escolar? de los estudiantes que están de acuerdo que durante la contingencia aprendieron lo mismo en las clases virtuales que en las clases presenciales existe un total desacuerdo en darse de baja del ciclo escolar.

Sin embargo, los alumnos que estuvieron totalmente en desacuerdo que aprendieron lo mismo en clases virtuales que en presenciales, de igual forma están

totalmente en desacuerdo que tuvieron que darse de baja del ciclo escolar. Por lo tanto, el desempeño académico en los estudiantes de educación superior en clases virtuales no influyó a que hubiera una deserción estudiantil.

La tercera prueba se realizó con la variable Beca como clase principal y Beca, Internet, P_2, P_3, P_4, P_5 y P_6 como variables de apoyo, pues las demás variables no tenían un grado de clasificación correcta acertada para que fuera válida, entonces con la variable Beca se obtuvo un 59% de instancias clasificadas correctamente lo que corresponde a 5201 de las cuales 4878 pertenecen a la clase “a” y 323 a la clase “b” (ver figura 13).

Figura 13

Resumen de Métricas de Evaluación de Árbol de Decisión con Pregunta Raíz Beca.

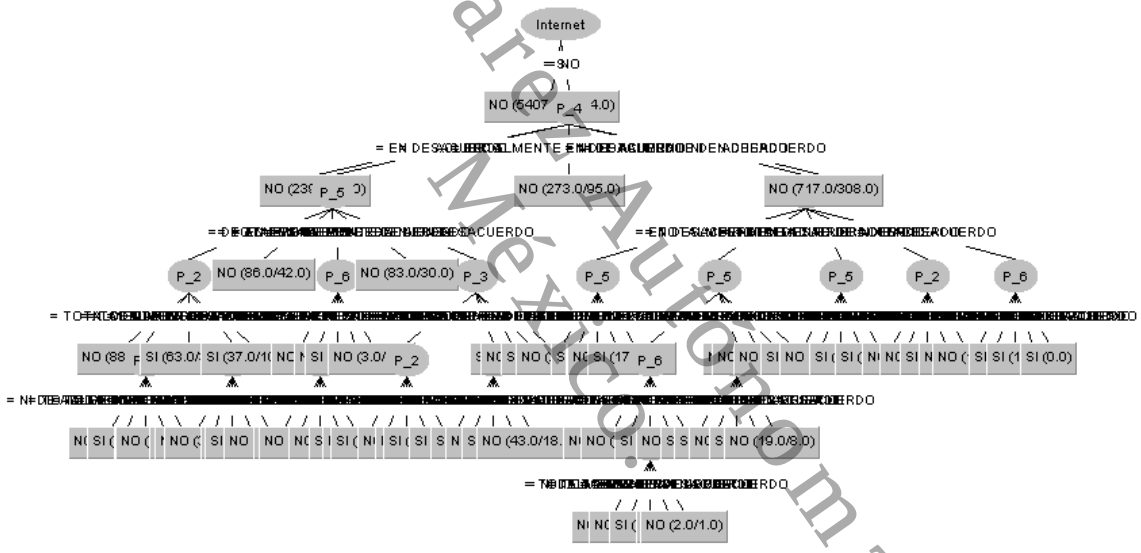
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5201               59.3789 %
Incorrectly Classified Instances      3558               40.6211 %
Kappa statistic                          0.0291
Mean absolute error                      0.478
Root mean squared error                  0.4938
Relative absolute error                 99.2798 %
Root relative squared error            100.6457 %
Total Number of Instances             8759

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.934    0.909    0.603      0.934    0.733      0.047    0.524     0.606     NO
               0.091    0.066    0.484      0.091    0.154      0.047    0.524     0.429     SI
Weighted Avg.  0.594    0.568    0.555      0.594    0.499      0.047    0.524     0.535

=== Confusion Matrix ===

    a    b   <-- classified as
 4878  345 |    a = NO
 3213  323 |    b = SI
```

Nota: Elaboración propia.

Como se puede ver el porcentaje correcto de instancias clasificadas fue de 59.37% por lo cual es un modelo con métricas aceptables por lo que se van a consideran para la interpretación de resultados. A continuación, se presenta el diagrama de árbol generado con un total de 20 nodos y 78 hojas (ver figura 14).

Figura 14

Visualización del Diagrama de Árbol con Pregunta Raíz Beca.



Nota: elaboración propia.

De acuerdo con el diagrama de árbol se puede decir que la clasificación “a” hace referencia al ítem no, mientras que la clasificación “b” hace referencia al ítem sí. Con la información obtenida de esta prueba se puede interpretar inicialmente en los patrones obtenidos de acuerdo con el diagrama de árbol en este ejemplo la pregunta objetivos es: Beca. ¿Cuenta con Beca? De los estudiantes que no cuentan con beca están en

desacuerdo que a pesar de la contingencia consideraban que podrían aprobar todas sus asignaturas del ciclo escolar mientras que los que están de acuerdo influyó a que tuvieran que darse de baja del ciclo escolar que cursaban.

Agrupación.

El segundo modelo de minería de datos que se utilizó con la herramienta *Weka* fue el de agrupación, para encontrar mejores agrupaciones se realizaron diferentes pruebas con el algoritmo *SimpleKMeans.*

En el siguiente apartado se muestran las tres pruebas con los diferentes números de *clústeres,* el parámetro considerado para realizar los ejemplos fueron las métricas que presenta Weka mediante la matriz de confusión, ésta es igual que el modelo de clasificación con árboles de decisión ya que ofrece el porcentaje de instancias agrupadas correctamente y así ve qué tan confiable fue la ejecución con ciertas variables.

Cabe destacar que la herramienta *Weka* permite visualizar las agrupaciones con diferentes gráficas y combinaciones de variables. La primera prueba se realizó con dos clústeres como internet como clase principal y Tipo_Institución, Género, Edad, Internet, P_2 y P_3 como variables de apoyo, en la figura 15 se puede ver el modelo que se clasificó en 5170 en el grupo cero y 3589 en el grupo uno, cabe destacar que el porcentaje de instancias clasificadas incorrectamente fueron de 4187 que corresponden a un 47.80% del Dataset (ver figura 15).

Figura 15

Modelo de Entrenamiento con Dos Clústeres en Weka

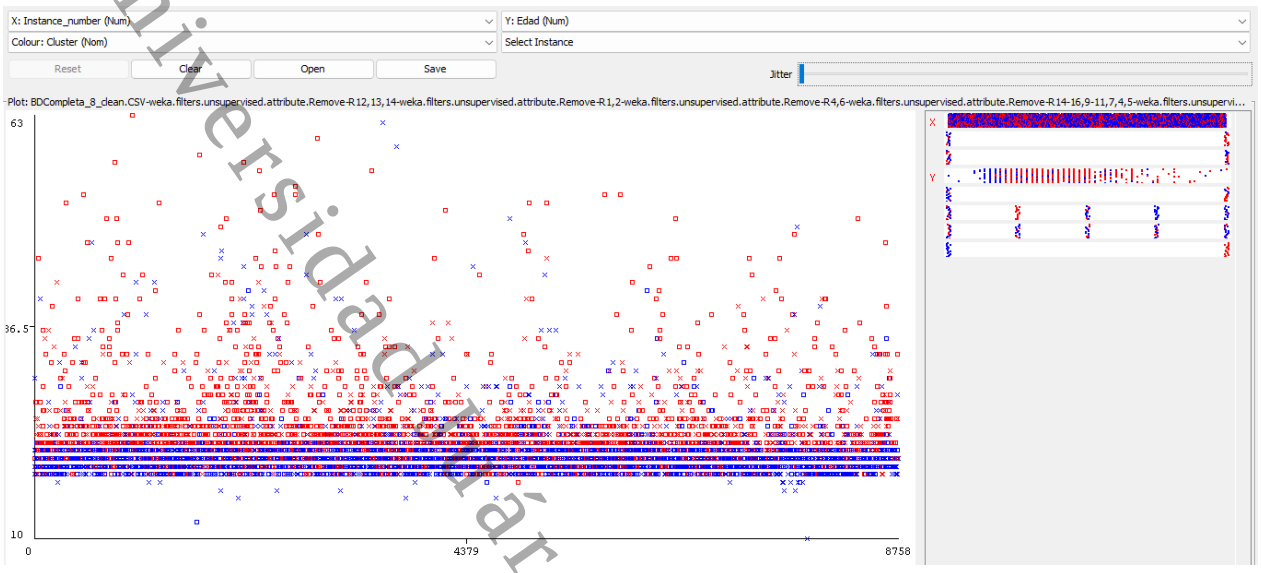
```
Time taken to build model (full training data) : 0.05 seconds

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

Clustered Instances

0      5170 ( 59%)
1      3589 ( 41%)


Class attribute: Internet
Classes to Clusters:

    0    1  <-- assigned to cluster
 3195 2212 | SI
 1975 1377 | NO

Cluster 0 <-- SI
Cluster 1 <-- NO

Incorrectly clustered instances :       4187.0   47.8023 %
```

Nota: Elaboración propia.

En la siguiente figura se muestra la gráfica realizada con la variable *Instance_number* y la variable Edad. A simple vista se observa una mayor densidad de color por debajo de la gráfica lo que significa que la mayoría de los estudiantes son menores de 24 años, (ver figura 16).

Figura 16

Prueba de Dos Agrupaciones por Edad con K-Means.



Nota: Elaboración propia.

La segunda prueba realizada se hizo con cuatro agrupaciones con las mismas variables que en la prueba uno. En la figura 17 se puede notar que las agrupaciones quedaron de la siguiente manera: en el grupo cero 2187, 1675 en el grupo uno, 1767 en el grupo dos y 3130 en el grupo tres. Del mismo modo se aprecia que 6022 instancias fueron clasificadas incorrectamente que corresponden al 68.75% del total de registros del Dataset, en comparación con las dos agrupaciones anteriores se puede notar que la calidad de los grupos formados disminuyó (ver figura 17).

Figura 17

Modelo de Entrenamiento con Cuatro Clústeres en Weka

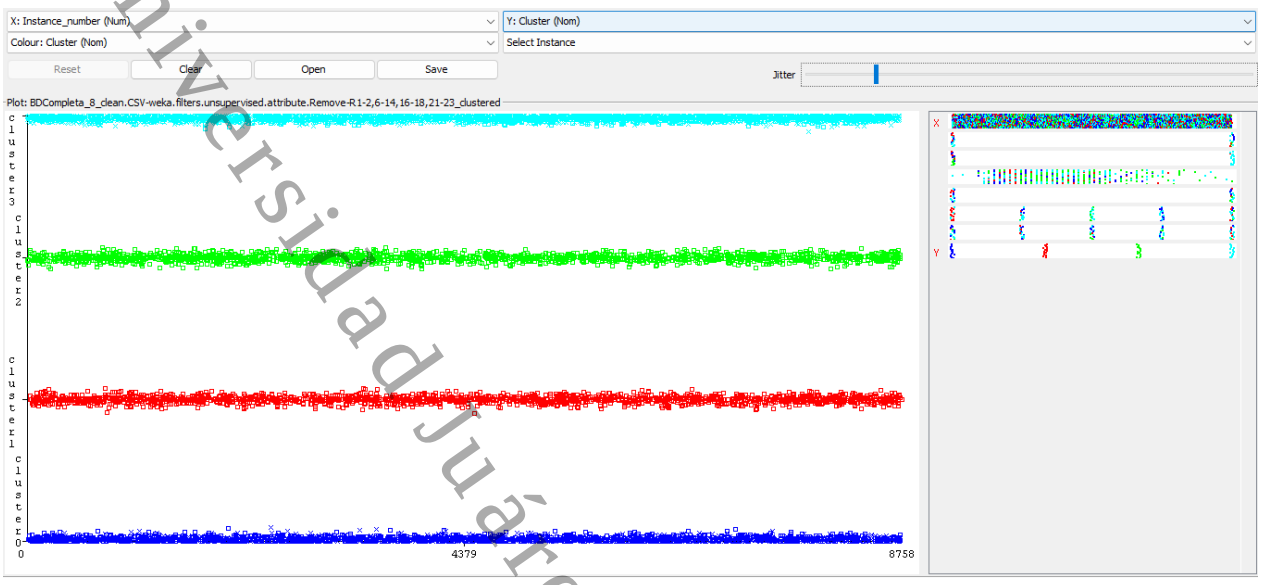
```
=== Model and evaluation on training set ===

Clustered Instances

0      2187 ( 25%)
1      1675 ( 19%)
2      1767 ( 20%)
3      3130 ( 36%)


Class attribute: Internet
Classes to Clusters:

    0    1    2    3  <-- assigned to cluster
 1352 1064 1089 1902 | SI
  835  611  678 1228 | NO

Cluster 0 <-- NO
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- SI

Incorrectly clustered instances :      6022.0    68.7521 %
```

Nota: Elaboración propia.

Para continuar, en la figura 18 se muestra la visualización de los clústeres mediante las variables *Instace_number* y Clúster, se puede notar a simple vista la separación de las cuatro agrupaciones que creó el algoritmo.

Figura 18

Prueba de Cuatro Agrupaciones por Clústeres con K-Means.



Nota: Elaboración propia.

La tercera prueba se hizo con cinco agrupaciones, de igual manera se utilizaron las mismas variables que la primera prueba, en la siguiente figura las agrupaciones fueron realizadas de la siguiente manera: 2169 instancias clasificadas en el grupo cero, 1304 en el grupo uno, 1696 en el grupo dos, 2551 en el grupo tres y finalmente 1039 instancias clasificadas pertenecientes al grupo cuatro (ver figura 19).

Figura 19

Modelo de Entrenamiento con Cinco Clústeres en Weka.

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

Clustered Instances

0       2169 ( 25%)
1       1304 ( 15%)
2       1696 ( 19%)
3       2551 ( 29%)
4       1039 ( 12%)


Class attribute: Internet
Classes to Clusters:

    0    1    2    3    4  <-- assigned to cluster
 1376  856 1027 1542  606 | SI
  793  448  669 1009  433 | NO

Cluster 0 <-- SI
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- NO
Cluster 4 <-- No class

Incorrectly clustered instances :      6374.0   72.7709 %
```

Nota: Elaboración propia.

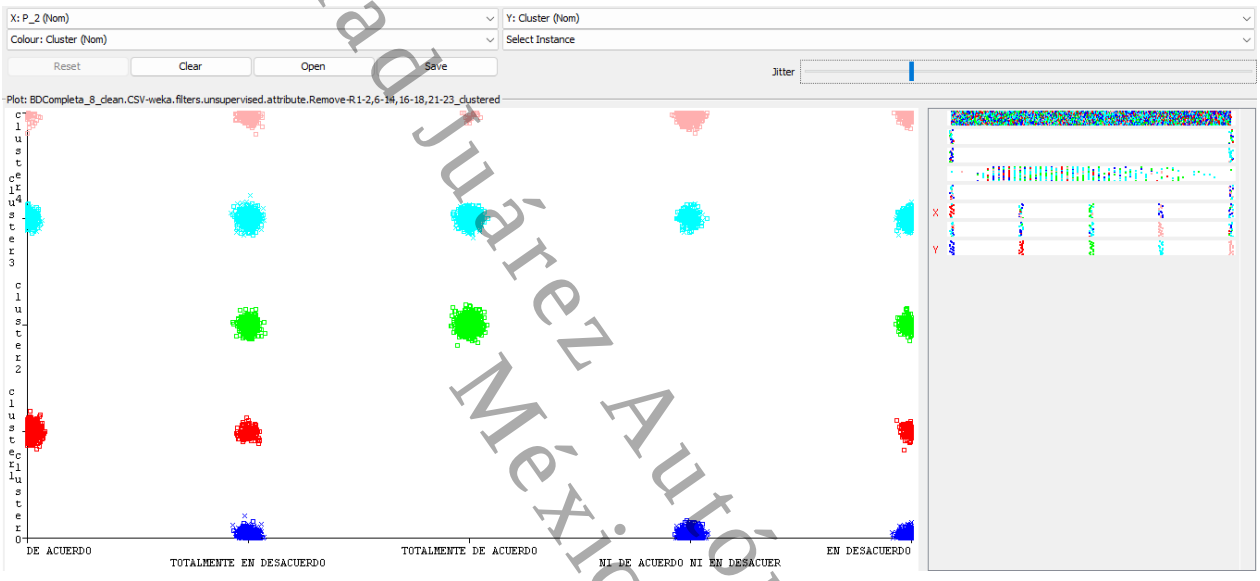
Como se puede ver en esta agrupación de cinco clústeres, todos quedaron con una diferente distribución. Se puede notar que el porcentaje de clasificación incorrecta es del 72.77% por lo cual mediante las métricas expuestas por *Weka,* esta prueba fue la que obtuvo los resultados más bajos y menos confiables.

A continuación, se presenta el siguiente diagrama con la variable P_2 y Clúster, el grafico representa las agrupaciones con respecto a si los profesores ofrecieron alternativas de trabajo si los estudiantes no contaban con los recursos, por ejemplo, internet, o computadora para realizar sus actividades a distancia por lo cual en la

siguiente gráfica se puede apreciar que hay un balance entre los clústeres creados con el algoritmo (ver figura 20).

Figura 20

Prueba de Cinco Agrupaciones con la Variable P_22 en Weka.



Nota: Elaboración propia.

con las pruebas de agrupamiento se mostraron los modelos, métricas de calidad y la funcionalidad de las gráficas de agrupación con el algoritmo *SimpleKMeans* con la finalidad de verificar qué tan confiable fueron los modelos creados.

Reglas de asociación.

Para continuar se contemplan tres ejemplos de pruebas realizadas con la técnica de reglas de asociación en la herramienta Weka, las cuales fueron ejecutadas con el algoritmo *A priori* con las variables Tipo_institución, Trabaja, Internet, P_1 y P_3, cabe

destacar que existieron cambios en las métricas como lo fue el parámetro de soporte y confianza.

Al realizar el primer ejemplo se obtuvieron 2190 instancias con un soporte del 25% y una confianza del 90% con un total de 15 ciclos requeridos para ejecutar la técnica. Por defecto la herramienta *Weka* muestra las primeras diez reglas. Se detectaron seis patrones de tamaño uno, 12 patrones de tamaño dos, ocho patrones de tamaño tres y un patrón de tamaño cuatro (ver figura 21).

Figura 21

Métricas del Algoritmo Apriori con Noventa por Ciento de Confianza.

```
Apriori
=======

Minimum support: 0.25 (2190 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 12

Size of set of large itemsets L(3): 8

Size of set of large itemsets L(4): 1
```

Nota: Elaboración propia.

En la figura 22 se puede observar que Weka generó las primeras diez reglas, la segunda regla se puede describir de la siguiente manera: Los alumnos están totalmente

en desacuerdo de darse de baja del ciclo escolar, aunque no tuvieran internet en casa y la mayoría pertenecen a una universidad pública.

Figura 22

Reglas de Asociación en Weka con Noventa por Ciento de Confianza.

```
Best rules found:

 1. Trabaja=NO Internet=NO 2465 ==> Tipo_institucion=PUBLICA 2450    <conf:(0.99)> lift:(1.02) lev:(0.01) [45] conv:(3.76)
 2. Internet=NO P_3=TOTALMENTE EN DESACUERDO 2249 ==> Tipo_institucion=PUBLICA 2223    <conf:(0.99)> lift:(1.01) lev:(0) [28] conv:(2.04)
 3. Internet=NO 3352 ==> Tipo_institucion=PUBLICA 3312    <conf:(0.99)> lift:(1.01) lev:(0) [41] conv:(2)
 4. Trabaja=NO P_1=TOTALMENTE EN DESACUERDO 2537 ==> Tipo_institucion=PUBLICA 2489    <conf:(0.98)> lift:(1.01) lev:(0) [13] conv:(1.26)
 5. Trabaja=NO P_3=TOTALMENTE EN DESACUERDO 4868 ==> Tipo_institucion=PUBLICA 4771    <conf:(0.98)> lift:(1) lev:(0) [21] conv:(1.21)
 6. Trabaja=NO 6720 ==> Tipo_institucion=PUBLICA 6582    <conf:(0.98)> lift:(1) lev:(0) [26] conv:(1.18)
 7. P_1=TOTALMENTE EN DESACUERDO 3323 ==> Tipo_institucion=PUBLICA 3246    <conf:(0.98)> lift:(1) lev:(0) [4] conv:(1.04)
 8. P_1=TOTALMENTE EN DESACUERDO P_3=TOTALMENTE EN DESACUERDO 2426 ==> Tipo_institucion=PUBLICA 2368    <conf:(0.98)> lift:(1) lev:(0) [1] conv:(1)
 9. P_3=TOTALMENTE EN DESACUERDO 6140 ==> Tipo_institucion=PUBLICA 5990    <conf:(0.98)> lift:(1) lev:(0) [0] conv:(0.99)
10. Trabaja=NO Internet=SI P_3=TOTALMENTE EN DESACUERDO 3151 ==> Tipo_institucion=PUBLICA 3064    <conf:(0.97)> lift:(1) lev:(-0) [-10] conv:(0.87)
```

Nota: Elaboración propia.

En cuanto al primer ejemplo la regla se cumple en 2465 registros de Dataset y la regla completa se cumple en 2450 del Dataset completo. Al visualizar las métricas de las reglas obtenidas se mantienen en el margen por lo cual las reglas de obtenidas son confiables y de calidad.

La segunda prueba realizada fue con una confianza del 70% con las mismas variables que el ejemplo anterior, la cual generó 4380 instancias con un soporte mínimo del 50% haciendo un total de 4380 instancias con diez ciclos requeridos para la ejecución del algoritmo, al realizarle cambios a los parámetros del algoritmo se le pidieron las primeras 25 reglas lo cual se podrá interpretar como un mayor número de patrones (ver figura 23).

Figura 23

Métricas del Algoritmo Apriori con Setenta por Ciento de Confianza.

```
Apriori
=======

Minimum support: 0.5 (4380 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Size of set of large itemsets L(2): 4

Size of set of large itemsets L(3): 1
```

Nota: Elaboración propia.

En la figura 24 se muestran las 25 reglas que se le solicitaron al algoritmo debido a la configuración de los parámetros ya que el soporte es muy alto y afecta los resultados mostrados. Mediante los resultados obtenidos de estas reglas se observa que los resultados son diferentes en algunas variables, pero no cambian mucho.

Figura 24

Reglas de Asociación en Weka con Setenta por Ciento de Confianza.

```
Best rules found:

 1. Trabaja=NO P_3=TOTALMENTE EN DESACUERDO 4868 ==> Tipo_institucion=PUBLICA 4771    <conf:(0.98)> lift:(1) lev:(0) [21] conv:(1.21)
 2. Trabaja=NO 6720 ==> Tipo_institucion=PUBLICA 6582    <conf:(0.98)> lift:(1) lev:(0) [26] conv:(1.18)
 3. P_3=TOTALMENTE EN DESACUERDO 6140 ==> Tipo_institucion=PUBLICA 5990    <conf:(0.98)> lift:(1) lev:(0) [0] conv:(0.99)
 4. Internet=SI 5407 ==> Tipo_institucion=PUBLICA 5233    <conf:(0.97)> lift:(0.99) lev:(-0) [-41] conv:(0.75)
 5. Tipo_institucion=PUBLICA P_3=TOTALMENTE EN DESACUERDO 5990 ==> Trabaja=NO 4771    <conf:(0.8)> lift:(1.04) lev:(0.02) [175] conv:(1.14)
 6. P_3=TOTALMENTE EN DESACUERDO 6140 ==> Trabaja=NO 4868    <conf:(0.79)> lift:(1.03) lev:(0.02) [157] conv:(1.12)
 7. P_3=TOTALMENTE EN DESACUERDO 6140 ==> Tipo_institucion=PUBLICA Trabaja=NO 4771    <conf:(0.78)> lift:(1.03) lev:(0.02) [157] conv:(1.11)
 8. Tipo_institucion=PUBLICA 8545 ==> Trabaja=NO 6582    <conf:(0.77)> lift:(1) lev:(0) [26] conv:(1.01)
 9. Tipo_institucion=PUBLICA Trabaja=NO 6582 ==> P_3=TOTALMENTE EN DESACUERDO 4771    <conf:(0.72)> lift:(1.03) lev:(0.02) [157] conv:(1.09)
10. Trabaja=NO 6720 ==> P_3=TOTALMENTE EN DESACUERDO 4868    <conf:(0.72)> lift:(1.03) lev:(0.02) [157] conv:(1.08)
11. Trabaja=NO 6720 ==> Tipo_institucion=PUBLICA P_3=TOTALMENTE EN DESACUERDO 4771    <conf:(0.71)> lift:(1.04) lev:(0.02) [175] conv:(1.09)
12. Tipo_institucion=PUBLICA 8545 ==> P_3=TOTALMENTE EN DESACUERDO 5990    <conf:(0.7)> lift:(1) lev:(0) [0] conv:(1)
```

Nota: Elaboración propia.

El último ejemplo se ejecutó con las mismas variables que el ejemplo anterior en el cual se obtuvieron 5255 instancias con un soporte mínimo del 60%, una confianza del 50%, solo mostró las principales reglas de las cuales hubo cuatro patrones de tamaño uno y dos patrones de tamaño dos (ver figura 25).

Figura 25

Métricas del Algoritmo Apriori con Cincuenta por Ciento de Confianza.

```
Apriori
=======

Minimum support: 0.6 (5255 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 8

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Size of set of large itemsets L(2): 2
```

Nota: Elaboración propia.

Como se observa en este último ejemplo el algoritmo solo generó cuatro reglas, en la cual la segunda regla dice que los alumnos están totalmente en desacuerdo que por problemas de la contingencia tuvieron que darse de baja del ciclo escolar que cursaban y además, pertenecen a una institución de educación superior pública.

Figura 26

Reglas de Asociación en Weka con Cincuenta por Ciento de Confianza.

```
Best rules found:

 1. Trabaja=NO 6720 ==> Tipo_institucion=PUBLICA 6582    <conf:(0.98)> lift:(1) lev:(0) [26] conv:(1.18)
 2. P_3=TOTALMENTE EN DESACUERDO 6140 ==> Tipo_institucion=PUBLICA 5990    <conf:(0.98)> lift:(1) lev:(0) [0] conv:(0.99)
 3. Tipo_institucion=PUBLICA 8545 ==> Trabaja=NO 6582    <conf:(0.77)> lift:(1) lev:(0) [26] conv:(1.01)
 4. Tipo_institucion=PUBLICA 8545 ==> P_3=TOTALMENTE EN DESACUERDO 5990    <conf:(0.7)> lift:(1) lev:(0) [0] conv:(1)
```

Nota: Elaboración propia.

De acuerdo con las pruebas realizadas con anterioridad a mayor precisión de confianza, mejores patrones obtiene el algoritmo. Sin embargo, cabe destacar que la primera prueba de reglas de asociación en la figura 21, las reglas obtienen métricas aceptables lo que resulta que los patrones encontrados contienen información relevante, de igual manera se observa que las reglas no son las mismas y a partir de esta información se pueden exponer los patrones detectados.

4.2 Interpretación de resultados.

En el apartado anterior se ejecutaron las pruebas realizadas con las técnicas de clasificación, agrupación y reglas de asociación mediante los algoritmos *J48*, *Apriori* y *SimpleKMeans* respectivamente. Los resultados se encuentran descritos en el orden en que se fueron ejecutando las técnicas de minería empleadas.

4.2.1 Clasificación Árboles de Decisión.

De acuerdo con las tres pruebas ejecutadas anteriormente para la técnica de árboles de decisión, se determinó que la variable que sí generaba información valiosa y confiable fueron las variables Tipo_institución, Género, Edad, Trabaja, Beca, Internet,

P_1, P_2, P_3 y P_6 como variables de apoyo y como variable principal P_3 ¿Por problemas derivados de la contingencia tuve que darme de baja del ciclo escolar? a continuación, se presentan los modelos de árboles obtenidos en Weka con la variable P_3 como pregunta raíz.

De acuerdo con la información generada con el algoritmo J48 en Weka en la siguiente figura 27 se presentan las métricas generadas y el modelo de árbol de decisión. Se observa que el 72% de las instancias fueron clasificadas correctamente lo que da un total de 6307 registros del Dataset, de los cuales el puntaje mayor se obtuvo en la clase “b” en donde los alumnos están totalmente en desacuerdo que por problemas derivados de la contingencia tuvieron que darse de baja del ciclo escolar, mientras que en la clase “d” pertenecen los estudiantes que están ni de acuerdo ni en desacuerdo que por problemas derivados de la contingencia tuvieron que darse de baja del ciclo escolar de acuerdo con la variable P_3 como el nodo raíz empleado en la ejecución del algoritmo.

Figura 27

Resumen de Evaluación del Modelo de Árbol de Decisión en Weka con Algoritmo J48.

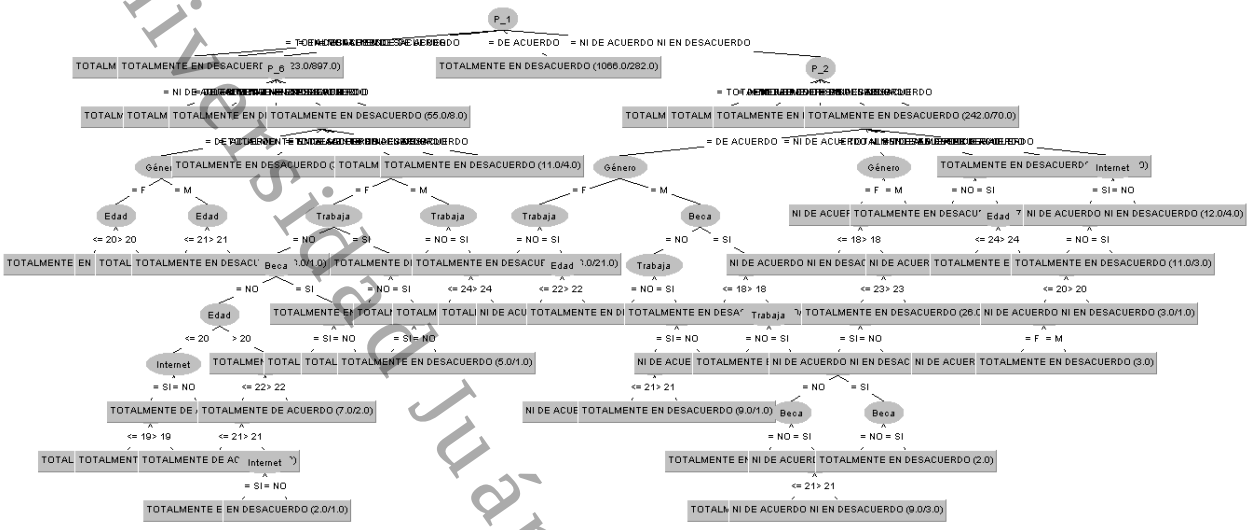
```
=== Summary ===

Correctly Classified Instances        6307               72.0059 %
Incorrectly Classified Instances      2452               27.9941 %
Kappa statistic                          0.18
Mean absolute error                      0.1767
Root mean squared error                  0.2996
Relative absolute error                 91.5657 %
Root relative squared error             96.4762 %
Total Number of Instances             8759

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.001    0.000    0.250      0.001    0.003      0.014    0.604     0.104     EN DESACUERDO
               0.967    0.840    0.730      0.967    0.832      0.227    0.594     0.754     TOTALMENTE EN DESACUERDO
               0.232    0.010    0.573      0.232    0.331      0.344    0.681     0.238     TOTALMENTE DE ACUERDO
               0.237    0.022    0.612      0.237    0.341      0.332    0.685     0.324     NI DE ACUERDO NI EN DESACUERDO
               0.003    0.001    0.167      0.003    0.005      0.016    0.580     0.061     DE ACUERDO
Weighted Avg.  0.720    0.592    0.645      0.720    0.644      0.221    0.610     0.592

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
    1  655    3   16    1 |    a = EN DESACUERDO
    2 5939   64  131    4 |    b = TOTALMENTE EN DESACUERDO
    0  346  106    4    0 |    c = TOTALMENTE DE ACUERDO
    0  834    5  260    0 |    d = NI DE ACUERDO NI EN DESACUERDO
    1  365    7   14    1 |    e = DE ACUERDO
```

Nota: Elaboración propia.

En la figura 28 se presentan el árbol de decisión generado de acuerdo con los parámetros establecidos en el *software* con un total de 45 nodos y 61 hojas a continuación, se presenta el diagrama de árbol completo.

Figura 28

Árbol de Decisión Con algoritmo J48 en Weka.



Nota: Elaboración propia.

De acuerdo con los resultados en el diagrama de árbol, se procedió a realizar la siguiente tabla en donde se muestran las decisiones más representables y así hacer una mejor interpretación de los resultados. Para la presentación de los patrones encontrados en el árbol de decisión se procedió a representar las decisiones de una forma más corta por lo cual la escala tipo *Likert* se renombró de la siguiente manera:

1=Totalmente de acuerdo

2=De acuerdo

3=Ni de acuerdo ni en desacuerdo

4=En desacuerdo

5=Totalmente en desacuerdo

En la tabla 10 se describen los 10 patrones principales encontrados con esa nomenclatura con el fin que no fueran tan extensos.

Tabla 10

Decisiones más Representativas en Weka Mediante el Árbol de Decisión.

No.	Clase	Variable	Decisión
1	5	{P_2=1}	P_1=3
2	2	{P_2=3}, {P_6=2}, {Género=M}, {Beca=NO}, {Trabaja=SI}, {Internet=NO}	P_1=3
3	5	{P_2=3}, {P_6=2}, {Genero=M}, {Beca=SI}, {Edad>18}, {Trabaja=SI}	P_1=3
5	5	{P_2= 3}, {P_6=3}, {Genero=F}, {Edad>18}, {Internet=SI}, {Trabaja=SI} {Beca=SI}	P_1=3
6	5	{P_6=1}, {P_2=1}, {Genero=M}, {Trabaja=SI}, {Edad>24}	P_1=1
7	5	{P_6=1}, {P_2=1}, {Genero=F}, {Trabaja=No}, {Beca=SI}, {Internet=SI}	P_1=1
8	5	{P_6=1}, {P_2=1}, {Genero=F}, {Trabaja=NO}, {Beca=NO}, {Edad==20} {Internet=SI}	P_1=1
9	1	{P_6=1}, {P_2=1}, {Genero=F}, {Trabaja=NO}, {Beca=NO}, {Edad<=20}, {internet=NO}	P_1=1

Nota: Elaboración propia.

De acuerdo con las decisiones encontradas en los resultados clasificados se detectaron los siguientes patrones:

La clase “b” como se mostró en la matriz de confusión de la figura 27 representa a los alumnos que están totalmente en descuerdo que por problemas derivados de la contingencia tuvieron que darse de baja del ciclo escolar haciendo un total de 8139 estudiantes de educación superior, los patrones enlistados con esta clase están escritos con el número cinco de la tabla 10.

Para comenzar se encontraron seis patrones en el cual están totalmente en desacuerdo en que tuvieron que darse de baja del ciclo escolar por problemas derivados de la contingencia, aunque estuvieron ni de acuerdo ni en desacuerdo en que durante la

pandemia aprendieron los mismo que en clases presenciales por lo cual existe un balance entre el aprendizaje que obtuvieron en clases virtuales que en presenciales.

En el patrón número tres identifica que los profesores les ofrecieron alternativas de trabajo si no contaban con los recursos como internet o computadora para realizar sus actividades o tareas y aunque el cambio de modalidad afectó su entorno inmediato como amigos o familiares, si el alumno contaba con beca y aun así trabajaba estaba en total desacuerdo en darse de baja ciclo escolar.

Cabe destacar que los patrones enlistados con el número cinco cuentan con las mismas características, aunque el alumno trabajara, tuviera beca, contara o no con una conexión a internet o la decisión seguía siendo la misma en no desertar del ciclo escolar a pesar de los problemas generados por la contingencia.

Por otro lado, se encuentran la clase “c” a como se puede ver en la matriz de confusión de la figura 27 con un total de 185 alumnos representa a los estudiantes que estuvieron de acuerdo en darse de baja del ciclo escolar por problemas de la contingencia, los patrones enlistados con esta clase se encuentran escritos con el número dos en la tabla 10.

De acuerdo con el patrón número dos, predomina el género masculino en donde el alumno no contaba con una beca, no tenía internet en su casa y trabajaba pues al añadirle los problemas derivados de la contingencia el alumno tenía que darse de baja de ciclo escolar añadiéndole que ese grupo de alumnos no estaban de acuerdo ni en

desacuerdo que los profesores les ofrecieron alternativas de trabajo si no contaban con los recursos como el internet o computadoras para realizar sus actividades.

4.2.2 Agrupamiento.

Para el caso del método de agrupamiento se verificaron las métricas mediante la matriz de confusión para corroborar la calidad de las agrupaciones obtenidas. Se realizó la prueba con dos agrupaciones con las variables Edad, P_1, P_2, P_3, P_4 y P_5 como variables de apoyo y como clase principal Internet: ¿Cuenta con internet?, las cuales se presentan de la siguiente manera: 3531 en el clúster cero y 5228 en el clúster uno. Cabe destacar que 4291 instancias fueron clasificadas incorrectamente siendo igual al 48.98% (ver figura 29).

Figura 29

Resumen de evaluación del modelo de Agrupación con el Algoritmo SimpleKMeans en WEKA.

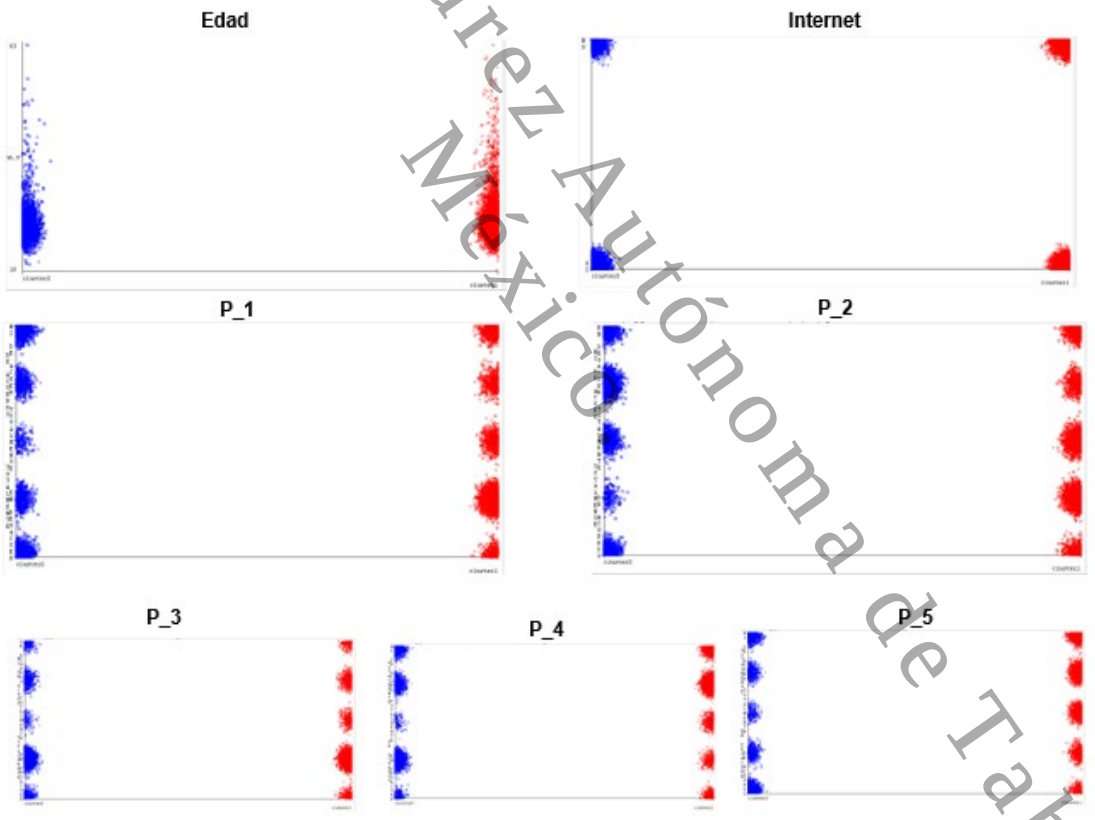
```
Clustered Instances

0       3531 ( 40%)
1       5228 ( 60%)


Class attribute: Internet
Classes to Clusters:

    0    1  <-- assigned to cluster
 2235 3172 | SI
 1296 2056 | NO

Cluster 0 <-- NO
Cluster 1 <-- SI

Incorrectly clustered instances :       4291.0   48.9896 %
```

Nota: Elaboración propia.

En la figura 30 se pueden ver las gráficas realizadas con la variable clúster en el eje “X” y en el eje “Y” se emplearon las variables Edad, Internet, P_1, P_2, P_3, P_4 y P_5. El clúster 0 se encuentra identificado en las gráficas de color azul mientras que el clúster 1 de color rojo. De acuerdo con la información de las agrupaciones y variables usadas.

Figura 30

Gráficos de Dispersión en Weka con Algoritmo SimpleKMeans.



Nota: Elaboración propia.

De acuerdo con la información presentada en la figura 30, se aprecia un balance de población entre los dos clústeres y se pude apreciar en la variable internet en el clúster rojo que los alumnos que sí contaban con internet estuvieron en descuerdo que debían de darse de baja del ciclo escolar, mientras que, en el clúster azul, aunque no tenían internet se encontraban en un balance de no desertar del ciclo escolar. De acuerdo con los gráficos generados en Weka se procedió a realizar la interpretación de los patrones resultados de los dos clústeres ejecutados.

Patrones Identificados en el Clúster Cero.

De acuerdo con los diagramas ejecutados en el clúster cero se aprecia una cantidad de alumnos de los cuales la mayoría contaba con acceso a internet, de los alumnos que sí contaban con internet durante la contingencia están en desacuerdo que durante la contingencia aprendieron lo mismo en clases virtuales que en presenciales mientras que los que no contaban con el acceso a internet mostraron un balance en que aprendieron lo mismo durante la contingencia.

Por otro lado, existe un balance entre los grupos formados en el cual consideran que los profesores les ofrecieron diversas alternativas de trabajo si no contaban con los recursos necesarios para llevar las clases virtuales como internet o computadora para realizar sus actividades.

Aunque los estudiantes contaran o no con una conexión a internet consideran un total desacuerdo en que por problemas derivados de la contingencia tuvieron que darse se baja del ciclo escolar, por lo que se deduce que los profesores les dieron alternativas

de trabajo, de igual forma están de acuerdo en que las formas de evaluación fueron las adecuadas y aunque no contaran con acceso a internet estuvieron de acuerdo en que pueden aprobar todas sus asignaturas del ciclo escolar que cursaban.

En esta primera agrupación existe un balance en donde los estudiantes se mostraron neutrales a no darse de baja del ciclo escolar, aunque no contaran con una conexión a internet y aun así poder aprobar todas sus asignaturas y que aprendieron lo mismo que en clases presenciales.

Patrones Identificados en el Clúster Uno.

En este grupo nombrado clúster de los alumnos que no cuentan con internet se encuentran totalmente en desacuerdo en que aprendieron lo mismo en clases virtuales que en las clases presenciales, de igual forma existe una inclinación a que los profesores no les ofrecieron alternativas de trabajo si no contaban con los recursos necesarios como una conexión a internet o computadora.

En el caso contrario para los que sí contaban con internet se muestra una inclinación a que los profesores sí les ofrecieron alternativas de trabajo.

Consideran en no darse de baja del ciclo escolar, aunque contaran o no con el acceso a internet y totalmente de acuerdo en que a pesar de la contingencia consideraban que podían aprobar todas sus asignaturas del ciclo escolar y finalmente para esta agrupación los estudiantes están totalmente de acuerdo en que las formas de evaluación durante la contingencia fueron las adecuadas.

En los patrones de este grupo se identificó principalmente que el internet no fue un factor que orilló a los estudiantes para continuar el ciclo escolar por los problemas derivados de la contingencia y los estudiantes estuvieron más seguros que podían continuar sus estudios virtualmente, aunque el patrón apunta en que aprendieron menos virtualmente.

4.2.3 Reglas de Asociación.

Mediante la técnica de reglas de asociación con las variables Tipi_institución, Trabaja, Internet, P_1, P_2, P_3, P_4 y P_5 se obtuvieron 1752 instancias con un soporte mínimo del 20%, una confianza del 97% y un total de 16 ciclos requeridos para ejecutar el algoritmo, las reglas generadas por el algoritmo fueron hechas con un total de cuatro ítems como tamaño máximo y uno como tamaño mínimo (ver figura 31).

Figura 31

Resumen de Evaluación del Modelo de Reglas de Asociación con el Algoritmo Apriori.

```
Apriori
=======

Minimum support: 0.2 (1752 instances)
Minimum metric <confidence>: 0.97
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 17

Size of set of large itemsets L(2): 28

Size of set of large itemsets L(3): 17

Size of set of large itemsets L(4): 5
```

Nota: Elaboración propia.

Cabe destacar que solo se le solicitaron las primeras 20 reglas al algoritmo, de igual manera el parámetro *lift* tiene un puntaje mayor a uno en todas las reglas lo que se traduce en que todos los patrones identificados contienen información confiable y relevante (ver figura 32).

Figura 32

Reglas de Asociación con Algoritmo Apriori.

```
Best rules found:

 1. Trabaja=NO Internet=NO 2465 ==> Tipo_institucion=PUBLICA 2450    <conf:(0.99)> lift:(1.02) lev:(0.01) [45] conv:(3.76)
 2. Internet=NO P_3=TOTALMENTE EN DESACUERDO 2249 ==> Tipo_institucion=PUBLICA 2223    <conf:(0.99)> lift:(1.01) lev:(0) [28] conv:(2.04)
 3. Internet=NO 3352 ==> Tipo_institucion=PUBLICA 3312    <conf:(0.99)> lift:(1.01) lev:(0) [41] conv:(2)
 4. Trabaja=NO P_5=NI DE ACUERDO NI EN DESACUERDO 2051 ==> Tipo_institucion=PUBLICA 2020    <conf:(0.98)> lift:(1.01) lev:(0) [19] conv:(1.57)
 5. P_4=NI DE ACUERDO NI EN DESACUERDO 1819 ==> Tipo_institucion=PUBLICA 1788    <conf:(0.98)> lift:(1.01) lev:(0) [13] conv:(1.39)
 6. Trabaja=NO P_1=TOTALMENTE EN DESACUERDO P_3=TOTALMENTE EN DESACUERDO 1910 ==> Tipo_institucion=PUBLICA 1876    <conf:(0.98)> lift:(1.01) lev:(0) [12] conv:(1.33)
 7. Trabaja=NO P_1=TOTALMENTE EN DESACUERDO 2537 ==> Tipo_institucion=PUBLICA 2489    <conf:(0.98)> lift:(1.01) lev:(0) [13] conv:(1.26)
 8. P_5=NI DE ACUERDO NI EN DESACUERDO 2636 ==> Tipo_institucion=PUBLICA 2585    <conf:(0.98)> lift:(1.01) lev:(0) [13] conv:(1.24)
 9. Trabaja=NO P_3=TOTALMENTE EN DESACUERDO 4868 ==> Tipo_institucion=PUBLICA 4771    <conf:(0.98)> lift:(1) lev:(0) [21] conv:(1.21)
10. Trabaja=NO 6720 ==> Tipo_institucion=PUBLICA 6582    <conf:(0.98)> lift:(1) lev:(0) [26] conv:(1.18)
11. P_4=DE ACUERDO 2038 ==> Tipo_institucion=PUBLICA 1996    <conf:(0.98)> lift:(1) lev:(0) [7] conv:(1.16)
12. P_2=TOTALMENTE EN DESACUERDO 2032 ==> Tipo_institucion=PUBLICA 1987    <conf:(0.98)> lift:(1) lev:(0) [4] conv:(1.08)
13. P_2=NI DE ACUERDO NI EN DESACUERDO 2074 ==> Tipo_institucion=PUBLICA 2028    <conf:(0.98)> lift:(1) lev:(0) [4] conv:(1.08)
14. P_1=TOTALMENTE EN DESACUERDO 3323 ==> Tipo_institucion=PUBLICA 3246    <conf:(0.98)> lift:(1) lev:(0) [4] conv:(1.04)
15. P_1=TOTALMENTE EN DESACUERDO P_3=TOTALMENTE EN DESACUERDO 2426 ==> Tipo_institucion=PUBLICA 2368    <conf:(0.98)> lift:(1) lev:(0) [1] conv:(1)
16. P_3=TOTALMENTE EN DESACUERDO 6140 ==> Tipo_institucion=PUBLICA 5990    <conf:(0.98)> lift:(1) lev:(0) [0] conv:(0.99)
17. Trabaja=NO P_3=TOTALMENTE EN DESACUERDO P_4=TOTALMENTE DE ACUERDO 2494 ==> Tipo_institucion=PUBLICA 2433    <conf:(0.98)> lift:(1) lev:(-0) [0] conv:(0.98)
18. P_1=NI DE ACUERDO NI EN DESACUERDO 2013 ==> Tipo_institucion=PUBLICA 1962    <conf:(0.97)> lift:(1) lev:(-0) [-1] conv:(0.95)
19. Trabaja=NO P_4=TOTALMENTE DE ACUERDO 2900 ==> Tipo_institucion=PUBLICA 2823    <conf:(0.97)> lift:(1) lev:(-0) [-6] conv:(0.91)
20. P_5=DE ACUERDO 1937 ==> Tipo_institucion=PUBLICA 1885    <conf:(0.97)> lift:(1) lev:(-0) [-4] conv:(0.89)
```

Nota: Elaboración propia.

De acuerdo con las reglas 20 reglas generadas en la herramienta Weka se identificaron cinco mejores reglas para ser consideradas e interpretadas por la información que aportan.

1. {Internet=NO, P_3=Totalmente en desacuerdo, Tipo_institucion=PUBLICA}

2223 *ítems* cuentan con esta regla los cuales son estudiantes que no cuentan con internet y aun así están totalmente en desacuerdo en que tuvieron que darse de baja del ciclo escolar por problemas derivado de la contingencia perteneciendo a una universidad pública de Tabasco.

2. {Trabaja=NO, P_1=Totalmente en desacuerdo, P_3=Totalmente en desacuerdo, Tipo_institucion=Pública}

1876 *ítems* cuentan con esta regla las cuales afirman que los alumnos que no trabajan estuvieron totalmente en desacuerdo en que durante la contingencia aprendieron lo mismo en las clases virtuales que en las presenciales y que también no se vieron en la idea de darse de baja del ciclo escolar por los problemas de la contingencia cabe destacar que todos estos pertenecen a universidad pública de Tabasco.

3. {Trabaja=NO, P_3 Totalmente en desacuerdo, Tipo_institución=pública}

Esta regla cuenta con 4771 *ítems* los cuales son estudiantes que no trabajan y están totalmente de desacuerdo en darse de baja del ciclo escolar por problemas de la contingencia todos pertenecen a una universidad pública de Tabasco.

4. {P_2=Ni de acuerdo ni en desacuerdo, Tipo_institución=pública}

Esta regla cuenta con 2028 *ítems* los cuales son estudiantes de instituciones públicas de Tabasco que se encuentran en un balance de que los profesores les ofrecieron alternativas de trabajo si no contaban con los recursos necesarios como internet, o computadora para realizar sus actividades.

5. {Trabaja=NO, P_3=Totalmente en desacuerdo, P_4=Totalmente de acuerdo, Tipo_institucion=Pública}

Esta regla cuenta con 2494 *ítems* de estudiantes de universidades públicas de Tabasco, los cuales a pesar de la contingencia consideran que pueden aprobar todas sus

asignaturas del ciclo escolar y no darse de baja cabe destacar que esta regla contiene alumnos que no trabajan.

Para finalizar se resume que las reglas encontradas coinciden con patrones identificados con la técnica de árbol de decisión y agrupación en donde, aunque no contaran con acceso a internet, no desertaban del ciclo escolar que estaban cursando y consideran que podían aprobar sus asignaturas, aunque se encontraban en un balance que durante la contingencia aprendieron lo mismo que en clases presenciales, por lo que el desempeño académico se vio reflejado en un aprovechamiento académico estable durante la pandemia. De igual forma, los profesores sí les ofrecieron alternativas de trabajo si no contaban con acceso a internet o computadoras para realizar sus actividades, por lo que consideraban que sí podían aprobar sus asignaturas pues los patrones identifican que la forma de evaluar a los estudiantes fueron las adecuadas y se mantuvo en un margen estable.

Capítulo 5. Conclusiones, discusión, recomendaciones y trabajos futuros

5.1 Conclusiones

Mediante el desarrollo de la investigación se llevaron a cabo las diferentes etapas de este proyecto, en el capítulo de resultados se mostró la información obtenida mediante los patrones encontrados en el Dataset.

El objetivo general del proyecto de investigación fue Identificar patrones de aprovechamiento, deserción y reprobación que afectaron a los estudiantes de educación superior de Tabasco en el contexto de la pandemia por Covid 19 empleando técnicas de minería de datos. Con base en el objetivo general se establecieron tres objetivos específicos los cuales están relacionados con las técnicas de minería de datos: arboles de decisión, agrupación y reglas de asociación. De igual manera, como último objetivo específico se estableció analizar los resultados de patrones obtenidos de estudiantes de educación superior en pandemia para la representación y exposición del conocimiento obtenido.

Se lograron cumplir todos los objetivos específicos con la herramienta *Weka.* Para la clasificación se ejecutó la técnica de árboles de decisión mediante el algoritmo *J48*, para la técnica de agrupación se empleó el algoritmo *SimpleKMeans* y finalmente la técnica de reglas de asociación con el algoritmo *Apriori* a la cual en los parámetros se le pidió que ejecutara las mejores 20 reglas. Para cada técnica mencionada anteriormente se ejecutaron tres pruebas. Cabe señalar que en todas las técnicas realizadas se hallaron

coincidencias y con todos los resultados obtenidos se fueron interpretando los patrones para identificar el conocimiento, por lo que el objetivo general fue alcanzado y cumplido.

De acuerdo con los resultados obtenidos en la herramienta *Weka* al emplear la técnica de árboles de decisión mediante el algoritmo *J48,* los patrones obtenidos apuntan a que existen seis reglas resultantes en las cuales estuvieron totalmente en desacuerdo en que tuvieron que darse de baja del ciclo escolar por problemas derivados de la contingencia, aunque estuvieron ni de acuerdo ni en desacuerdo en que durante la pandemia aprendieron los mismo que en clases presenciales por lo cual existe un balance entre el aprendizaje que obtuvieron en clases virtuales que en presenciales.

De igual forma se identificó que los profesores les ofrecieron alternativas de trabajo si no contaban con los recursos como internet o computadora para realizar sus actividades o tareas y aunque el cambio de modalidad afectó su entorno inmediato como amigos o familiares, si el alumno contaba con beca y aun así trabajaba estaba en total desacuerdo en darse de baja ciclo escolar.

Otra conclusión obtenida es que, aunque el alumno trabajara o tuviera beca contara o no con una conexión a internet la decisión seguía siendo la misma en no desertar del ciclo escolar a pesar de los problemas generados por la contingencia.

Por otro lado, mediante la técnica de agrupación con el algoritmo *SimpleKMeans,* los patrones identificados aprecian una cantidad de alumnos de los cuales la mayoría contaba con acceso a internet, de los alumnos que sí contaban con internet durante la contingencia, están en desacuerdo que durante la contingencia aprendieron lo mismo en

clases virtuales que en presenciales, mientras que los que no contaban con el acceso a internet mostraron un equilibrio en que aprendieron lo mismo durante la contingencia.

Por otro lado, existe un balance entre los grupos formados en los que consideran que los profesores les ofrecieron diversas alternativas de trabajo si no contaban con los recursos necesarios para llevar las clases virtuales como internet o computadora para realizar sus actividades.

Aunque los estudiantes contaran o no con una conexión a internet, consideraron un total desacuerdo en que por problemas derivados de la contingencia tuvieron que darse de baja del ciclo escolar, por lo que se deduce que los profesores les dieron alternativas de trabajo. De igual forma estuvieron de acuerdo en que las formas de evaluación fueron las adecuadas y aunque no contaron con acceso a internet estuvieron de acuerdo en que pudieron aprobar todas sus asignaturas del ciclo escolar que cursaban en ese momento.

Finalmente, la técnica de reglas de asociación se ejecutó con el algoritmo *A priori.* Los patrones identificados apuntan a que aquellos estudiantes que no contaban con *internet* estaban totalmente en desacuerdo en darse de baja del ciclo escolar por problemas derivados de la contingencia perteneciendo a una universidad pública de Tabasco.

De igual forma los alumnos que no trabajan estuvieron totalmente en desacuerdo en que durante la contingencia aprendieron lo mismo en las clases virtuales que en las presenciales y que aun así al no obtener ese aprendizaje deseado, no se vieron en la

idea de darse de baja del ciclo escolar por los problemas de la contingencia. Cabe destacar que todos estos pertenecen a universidad pública de Tabasco.

De acuerdo con los patrones encontrados en las tres técnicas de minería de datos ejecutadas se puede exponer que el desempeño académico de los estudiantes de educación superior en Tabasco en pandemia no fue muy trastocado, se mantuvieron en un ambiente distinto pues su aprovechamiento, deserción o reprobación no se vio tan afectado en los estudiantes aunque no contaran con los recursos como lo fue la conexión a internet, computadora, su aprendizaje obtenido o la aprobación de las asignaturas a pesar de que las formas de evaluación no fueron las más adecuadas, lo cual puede deberse a la inclusión repentina de un cambio de modalidad, de la presencial a la virtual.

En la tabla 11 se muestran los principales patrones de que fueron identificados mediante las técnicas de árbol de decisión, agrupación y reglas de asociación.

Tabla 11

Patrones Principales Identificados

Variable	Patrones identificados
Aprovechamiento	La falta de internet no fue un factor para que los alumnos desertaran del ciclo escolar. Su aprovechamiento académico fue balanceado pues los patrones apuntan en que aprendieron los mismo en clases virtuales y presenciales. Los profesores ofrecieron alternativas de trabajo si no contaban con los recursos como internet o computadora para realizar tareas y aunque el cambio de modalidad

	afectó su entorno inmediato como amigos o familiares, si el alumno contaba con beca y aun así trabajaba no pensaban en darse de baja del ciclo escolar. Fueron pocos los problemas derivados de la contingencia en educación superior en Tabasco con respecto a su Aprovechamiento, deserción y reprobación.

Nota: Elaboración propia.

De acuerdo con los principales patrones identificados en Tabasco, el internet no fue una amenaza para los estudiantes de educación superior, aunque existe población con pocas posibilidades económicas, la clave estuvo en las alternativas de aprendizaje ofrecidas por los profesores de las IES, lo que les ayudó a no tomar la decisión de desertar o reprobar, de tal manera que pudieran seguir con sus estudios de educación superior a distancia.

Se concluye que el implementar técnicas de minería de datos con un Dataset con un volumen considerable de registros, diseñado exprofeso para explicar las variables de aprovechamiento, deserción y reprobación; se pudo obtener un conocimiento relevante, respaldado en evidencia empírica, y se logró la obtención de los patrones de estudiantes.

5.2 Discusión

En este apartado se discuten los hallazgos de la investigación, con los de otras investigaciones; la discusión se centra en algunos aspectos principales identificados al ejecutar las técnicas de minería de datos los cuales fueron el acceso a internet, el

aprendizaje obtenido en y alternativas de trabajo que se les dieron a los alumnos en pandemia.

Existen muchas investigaciones abordando temas de desempeño académico que se realizaron con *Datasets* que contenían registros y variables diferentes, también la ejecución de herramientas y técnicas de minería de datos distintos, pero sobre todo con un objetivo en particular y la misma finalidad de poder identificar patrones y reglas que pudieran explicar la percepción de los estudiantes con respecto a su desempeño en pandemia.

Tal es que caso de López (2022) entre los resultados de su estudio señala que, en la mayoría de los casos, los estudiantes contaban con la infraestructura tecnológica y competencias digitales para continuar con los procesos de aprendizaje de manera virtual. Estos resultados son poco coincidentes con los de la investigación de esta tesis, ya que, para el contexto de Tabasco, la mayoría de los alumnos no contaban con los recursos para llevar las clases a distancia.

Gutiérrez y Meza (2021) señalan que, si un alumno tiene problemas económicos, y falta de tiempo, la probabilidad de desertar es la más alta. Estas aseveraciones contrastan con los hallazgos de la tesis, ya que las respuestas de los patrones obtenidos refieren que, aunque el alumno trabajara, tuviera beca contara o no con una conexión a internet, la decisión seguía siendo la misma; no desertar del ciclo escolar a pesar de los problemas generados por la contingencia.

Miguel (2020) y Ramírez-Melo *et al.*(2022) apuntan que si un alumno accedía a las clases mediante un dispositivo de acceso público o de renta es porque estaba en riesgo de reprobar en cambio, a que se conectara por medio de una laptop o computadora de escritorio disminuían sus probabilidades de ser vulnerable, del mismo modo señalan que algunas de las inconformidades de los estudiantes ante este cambio de modalidad radicaron en la mala comunicación con los profesores, que las clases se basaran en cargas de tareas, sin explicación previa o retroalimentación, en algunos casos y sin duda, la conectividad representaba un problema.

En contraste con lo anterior en esta investigación se identificaron patrones en los que el acceso a internet no fue un problema que orillara a los alumnos a reprobar o darse de baja del ciclo escolar porque los profesores les ofrecieron alternativas de trabajo si no contaban con los recursos necesarios (conexión a internet o computadora) para realizar sus tareas o actividades. Por consiguiente, puede decirse que fueron perspectivas y ambientes diferentes las que causaron un desempeño académico distinto.

Ortega-Encinas *et ál*., (2022) refieren que la educación virtual o a distancia posee características que la diferencian en gran medida de la educación presencial muchos de los estudiantes conceden un carácter más práctico a sus objetivos de aprendizaje. De igual forma, refieren que algunos estudiantes aprenden más en la práctica, pero también eso va dependiendo de la materia o disciplina de estudio. Para el contexto de esta investigación, se comprobó mediante la técnica de clasificación con árboles de decisión, la existencia de un patrón en el cual los alumnos apuntan que durante la pandemia

aprendieron los mismo que en clases presenciales, de igual forma consideraron que las formas de evaluación fueron las adecuadas.

5.3 Recomendaciones

Derivado de la experiencia generada durante el desarrollo de esta investigación, se presentan recomendaciones sobre la herramienta utilizada en este desarrollo y el Dataset utilizado.

De acuerdo con el modelo *FURPS* la herramienta *Weka* fue la herramienta con el mejor puntaje ya que tiene funciones que apoyan a personas que tienen poca experiencia en el uso de herramientas tecnológicas, pues tiene cargada las técnicas y los algoritmos para ser ejecutados con el *Dataset*.

Algunas de las limitaciones de la herramienta Weka son que, al visualizar los diagramas de árbol, no se puede observar adecuadamente sí consta de muchas variables, por lo que hay que recurrir a podar el árbol, lo que representa una eliminación de variables, lo que viene en detrimento de la calidad al modelo.

Es de gran importancia documentar el proceso de análisis utilizando Weka, incluyendo los pasos de preprocesamiento de datos, selección de atributos, construcción y evaluación de modelos, lo cual ayuda a comunicar claramente los métodos y resultados.

Otro punto importante por considerar es asegurarse de que el Dataset sea relevante para la investigación y los objetivos, éste debe contener datos que puedan responder a las preguntas que se plantean en el estudio.

Por último, con relación al *Dataset*, se recomienda realizar preguntas cerradas porque así no hay tanto desfase al realizar la limpieza y transformación del *Dataset* lo que permitirá mejorar la calidad de los datos sin tener valores faltantes, errores o inconsistencias que puedan afectar la integridad del análisis.

5.4 Trabajos futuros

A continuación, se mencionan algunas directrices sobre las que se puede seguir indagando sobre este tema de investigación:

- El Dataset utilizado en este trabajo fue realizado por medio de una encuesta aplicada en pandemia, por lo que sería recomendable realizar un nuevo análisis con un Dataset en tiempo actual para identificar nuevos patrones.
- Utilizar y comparar herramientas de minería de datos con programación como Python y lenguaje R para identificar quién tiene mejor nivel de confiabilidad.
- En este trabajo para la identificación de patrones se utilizaron técnicas de clasificación, agrupación y reglas de asociación con los algoritmos *J48, SimpleKMeans y Apriori* respectivamente, sin embargo, utilizar nuevas técnicas con diferentes algoritmos podrían ser empleados y comparar resultados con los de este estudio.
- Esta investigación se exploró desde las variables de aprovechamiento, deserción y reprobación de estudiantes de educación superior de Tabasco, por lo que sería novedoso abordar otras variables dentro de las IES para resolver problemáticas.

- Desarrollar un modelo predictivo utilizando técnicas de minería de datos para identificar patrones y factores de riesgo que puedan predecir la deserción estudiantil en IES de Tabasco, con la finalidad de intervenir tempranamente y diseñar estrategias de retención efectivas.
- Utiliza técnicas de análisis de trayectorias académicas para identificar patrones comunes entre estudiantes que tienen éxito académico y aquellos que enfrentan dificultades. Por ejemplo, examinando factores como el desempeño en cursos previos, el nivel socioeconómico, la ubicación geográfica, entre otros, para comprender mejor qué impulsa el éxito o el fracaso académico.

Alojamiento de la Tesis en el Repositorio Institucional	
Título de Tesis:	Aprovechamiento, deserción y reprobación de estudiantes universitarios de Tabasco desde la minería de datos.
Autor(a) o autores(ras) de la Tesis:	Luis Manuel Juárez López
ORCID:	https://orcid.org/0009-0002-9192-4796
Resumen de la Tesis:	El desempeño escolar que hubo en tiempos de pandemia debido al Covid-19 fue un reto en el ambiente educativo para tomar las clases a distancia puesto que no todos los alumnos contaban con los medios de comunicación necesarios para recibirlas. Sin embargo, las Tecnologías de la Información fueron el pilar fundamental para llevar a cabo las clases con las diferentes herramientas y plataformas en línea. Para efectos de esta investigación se aplicaron técnicas de minería de datos con un enfoque cuantitativo las cuales fueron clasificación, agrupación y reglas de asociación. el *Dataset* se obtuvo al realizar una encuesta integrada por seis *ítems,* la cual se efectuó en el año 2020 por medio de cuestionarios realizados en formularios de *Google* enviados a estudiantes que cursaban el ciclo enero-junio 2020 de los diferentes programas de licenciatura y posgrado que se imparten en Tabasco, México. Derivado de lo anterior, se identificaron las variables de aprovechamiento académico, deserción y reprobación en tiempos de pandemia mediante técnicas de minería de datos.
Palabras claves de la Tesis:	Educación, Minería de datos, Pandemia.
Referencias citadas:	Se muestra a partir de la página 105.

Referencias

Aquino, S. P. (2021). Baja o deserción en tiempos de la pandemia del COVID-19. Experiencia de un estudiante de pregrado. En G.C. Medina, Aquino. S.P. y López, M. (Coord.). *La tecnología educativa en tiempos de pandemia* (pp.190-208). Gradus editora. https://es.graduseditora.com/_files/ugd/c7d661_16fb591baa034f00a995f9868e9aa92b.pdf

Azuara, W. G. (2023). Identificación de patrones de conducta, asociadas con las barreras para el uso de innovaciones Tecnológicas en estudiantes de educación superior. Aplicando técnicas de minería de datos. [Tesis de Maestría].

Beltrán, L. (2022). *Construcción de modelos predictivos de la deserción universitaria utilizando minería de datos, caso de estudio: CETYS Universidad campus ensenada* [Tesis doctoral, Universidad Popular Autónoma de Puebla]. Repositorio Institucional. https://repositorio.cetys.mx/handle/60000/1509

Callejas-Cuervo, M. y Alarcón-Aldana, A. C. (2017). Modelos de calidad del software, un estado del arte. *Entramado*, *13*(1), 236–250. https://doi.org/10.18041/entramado.2017v13n1.25125

Carracedo, P. y Terrádez M. (2016). El proceso de descubrimiento de conocimiento a partir de datos. *Universitat Oberta de Catalunya (UOC),* 1-66. http://hdl.handle.net/10609/138187

Cendejas-Valdez, J. L., Acuña-López, M. A., Cortes-Morales, G. y Bolaños-Jiménez, G. (2017). El uso de modelos y metodologías de minería de datos para la inteligencia de negocios. *Revista de Sistemas Computacionales y TIC'S, 3*(8). 54-63. https://www.ecorfan.org/spain/researchjournals/Sistemas_Computacionales_y_TICs/vol3num8/Revista_de_Sistemas_Computacionales_y_TIC%60S_V3_N8.pdf

Constanzo, M.A. (2014). Comparación de modelos de calidad, factores y métricas. *Informe Científico Técnico UNPA*, *6*(1). 1-36. 10.22305/ict-unpa.v6i1.89

Corzo, C. (s.f.). Deserción escolar. https://www.uaeh.edu.mx/scige/boletin/prepa3/n8/p1.html#r1

Chalpartar, L. T. M., Fernández, A. M., Betancourth, S. y Gómez, Y. A. (2022). Deserción en la población estudiantil universitaria durante la pandemia, una mirada cualitativa. *Revista Virtual Universidad Católica Del Norte,* (66). 37–62. https://doi.org/10.35575/rvucn.n66a3

Diaz, D. y Ruiz, A. (2018). Reprobación escolar en el nivel medio superior y su relación con el autoconcepto en la adolescencia. *Revista Latinoamericana De Estudios Educativos, 48*(2), 125-142. https://doi.org/10.48102/rlee.2018.48.2.49

Díaz, D. y Ruiz, A. (2018). Reprobación escolar en el nivel medio superior y su relación con el autoconcepto en la adolescencia. *Revista Latinoamericana de Estudios Educativos, 48*(2). 125-142. https://doi.org/10.48102/rlee.2018.48.2.49

Dicovskiy, L. M. y Pedroza, M. E. (2018). Minería de datos, una innovación de los métodos cuantitativos de investigación, en la medición del rendimiento académico

universitario. *Revista Científica De FAREM-Estelí,* (24), 143–152. https://doi.org/10.5377/farem.v0i24.5557

Edel, R. (2003). El rendimiento académico: concepto, investigación y desarrollo. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 1*(2), 1-15. http://hdl.handle.net/10486/660693

Garbanzo, G. M. (2012). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación, 31*(1), 43–63. https://doi.org/10.15517/revedu.v31i1.1252

García, J. F., Chávez, J. A., Contreras, J. R., Martínez, L. y Pineda, L. C. (2022). Retos y perspectivas de la educación universitaria intercultural en tiempos de pandemia. Una visión docente. *Revista iberoamericana para la investigación y el desarrollo educativo, 13*(25), 1-29. https://doi.org/10.23913/ride.v13i25.1314

Gutiérrez, G. A. y Meza, M. (2021). Modelo de árbol de decisiones para determinar los factores de deserción de estudiantes en una institución de educación superior mexicana. *Revista de Gestión Empresarial y Sustentabilidad, 7*(1), 19-30. https://rges.umich.mx/index.php/rges/article/view/67

Guzmán, Z. K. (2020). Identificación de patrones de conducta, asociadas con procrastinación y autorregulación académica en estudiantes de educación superior, aplicando técnicas de minería de datos. [Tesis de Maestría].

Hernández, J., Ramirez, J. y Ferri, C. (2004). *Introducción a la minería de datos.* Universidad politécnica de valencia. https://www.amazon.com.mx/Introduccion-Mineria-Datos-Hernandez/dp/8420540919

Izar, J. M., Ynzunza, C. B., y López, H. (2011). Factores que afectan el desempeño académico de los estudiantes de nivel superior en Rioverde, San Luis Potosí, México. *Revista de Investigación Educativa*, (12), 1-18. https://doi.org/10.25009/cpue.v0i12.50

Jaramillo, A. y Paz-Arias, H. P. (2015). Aplicación de técnicas de minería de datos para determinar las interacciones de los estudiantes en un entorno virtual de aprendizaje. *Revista Tecnológica - ESPOL, 28*(1). 64-90. http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/351

Ley federal de protección de datos personales en posesión de los particulares. (2010) *Diario Oficial de la Federación.* 05-07-2010.1-18. https://www.diputados.gob.mx/LeyesBiblio/pdf/LFPDPPP.pdf

Lizana, M.I. (2020). Ventajas de R como herramienta para el Análisis y Visualización de datos en Ciencias Sociales. *Revista Científica de la UCSA, 7*(2), 97-111. https://doi.org/10.18004/ucsa/2409-8752/2020.007.02.097

López, M. D. y Contreras, A. (2022). El impacto de la pandemia por covid-19 en estudiantes mexicanos de educación media superior. *Revista iberoamericana para la investigación y el desarrollo educativo, 12*(24), 1-27. https://doi.org/10.23913/ride.v12i24.1141

Martínez, J. J. (2017). Análisis de minería de datos distribuida con Weka Parallel en computadoras con múltiples procesadores físicos y lógicos. *Economía y Administración (E&A), 6*(2), 155–166. https://doi.org/10.5377/eya.v6i2.4307

Martínez, L. y Zamudio, B. F. (2021). Tendencias de la Educación Superior en el Estado de Tabasco ante la pandemia por el COVID 19, en un estudio de caso. En V. Estrada, S. Mora, M. Pilar. [Coord.]. *Estudios sobre cultura y desigualdad en las regiones.* (pp. 1-20). Universidad Nacional Autónoma de México, Instituto de Investigaciones Económicas y Asociación Mexicana de Ciencias para el Desarrollo Regional. http://ru.iiec.unam.mx/id/eprint/5592

Meza, M.P. y Gutiérrez, G.A. (2021). Modelo Basado en Árbol de Decisiones para Determinar los Factores de Deserción de Estudiantes en una Institución de Educación Superior Mexicana. *Revista de Gestión Empresarial y Sustentabilidad, 7*(1), 19-30. https://rges.umich.mx/index.php/rges/article/view/67

Miguel, J. A. (2020). La educación superior en tiempos de pandemia: una visión desde dentro del proceso formativo. *Revista Latinoamericana De Estudios Educativos, 50*, 13-40. https://doi.org/10.48102/rlee.2020.50.ESPECIAL.95

Miranda, M. A, y Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. *Formación universitaria, 10*(3), 61-68. https://dx.doi.org/10.4067/S0718-50062017000300007

Organización de las Naciones Unidas para la Educación la Ciencia y la Cultura. [UNESCO]. (2020). *La educación en tiempos de Covid-19.* Cepal.

https://www.cepal.org/es/publicaciones/45904-la-educacion-tiempos-la-pandemiacovid-19

Ortega-Encinas, L. U., López, J. S., Sortillón, P. E., Gamiño, D. T. y Cheu, E. (2022). Impacto en el rendimiento escolar bajo condiciones de pandemia SARS-COV2. *Revista De Investigación Académica Sin Frontera: División De Ciencias Económicas Y Sociales*, *15*(37). https://doi.org/10.46589/rdiasf.vi37.429

Osorio, M. C., Mejía L.H. y Navarro, J. A. (2012). Perfil del alumno de éxito en el aprovechamiento escolar de la asignatura de física general: Caso del Plantel Ignacio Ramírez Calzada. *Espacios Públicos, 15*(35), 134-15. https://espaciospublicos.uaemex.mx/article/view/19733

Pérez-Gutiérrez, B. R. (2020). Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico. *Revista UIS Ingenierías, 19*(1), 193–204. https://doi.org/10.18273/revuin.v19n1-2020018

Portilla, J. G., Gamboa, E. J., Chan, M. R., López, C. H. y Martin, J. A. (2022). Análisis de los índices de reprobación en la carrera de ITICS utilizando técnicas de inteligencia artificial y minería de datos en el tecnológico nacional de México campus Conkal. *South Florida Journal of Development, 3*(6), 7268–7278. https://doi.org/10.46932/sfjdv3n6-069

Ramírez-Melo, L., Delgado-Ávila, E. R. y Montufar-Benítez, M. A. (2022). Aplicación de técnicas de minería de datos para la caracterización de estudiantes bajo el efecto

de la COVID-19. *Padi Boletin científico de ciencias básicas en ingeniería del ICBI, 10*(2), 75-81. https://doi.org/10.29057/icbi.v10iEspecial2.8669

Ramón, P. García, V. Aquino, S. P. y Silva, M. P. (2023). La deserción escolar de estudiantes universitarios perspectivas y propuestas desde los propios actores. https://pculturales.ujat.mx/FilesPublicaciones/files398/la%20desercion%20escolar%20de%20estudiantes%20universitarios.pdf

Reyes, J. F. y García, R. (2005). El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierías,* *8*(26), 37-47. https://ingenierias.uanl.mx/anteriores/26/pdfs/26_el_proceso.pdf

Reyes-Nava, A., Gil-Antonio, I. y Antonio-Velázquez, J. A. (2021). Identificación de factores de riesgo que causan la deserción de alumnos que estudian a distancia por causa del COVID19 usando técnicas de minería de datos. En A. Ledesma (Coord.), Ciencias de la Ingeniería y Tecnología (pp. 62-72). Handbooks-TX-©ECORFAN-México. https://www.ecorfan.org/handbooks/Handbooks_Ciencias_de_la_Ingenieria_y_Tecnologia_TX/Handbooks_Ciencias_de_la_Ingenieria_y_Tecnologia_TX_4.pdf

Riquelme, J.C., Ruiz, R. y Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, *10*(29). 11-18. http://hdl.handle.net/11441/43290

Ruiz-Ramírez, R., García-Cué, J. L. y Pérez-Olvera, M. A. (2014). Causas y consecuencias de la deserción escolar en el bachillerato: Caso Universidad Autónoma de Sinaloa. *Ra Ximhai, 10*(5), 51-74. 10.35197/rx.10.03.e1.2014.04.rr

Saucedo, M., Herrera-Sánchez, S. del C., Díaz, J. J., Bautista, S., y Salinas, H. A. (2015). Indicadores de reprobación: Facultad de Ciencias Educativas (UNACAR). *RIDE Revista Iberoamericana Para La Investigación Y El Desarrollo Educativo, 5*(9), 96 - 106. https://www.ride.org.mx/index.php/RIDE/article/view/7

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional* (pp. 63-86). Ediciones Universidad Cooperativa de Colombia. http://dx.doi.org/10.16925/978958x7600490

Troche, A. (2014). Aplicación de la minería de datos sobre bases de datos transaccionales. *Fides et Ratio - Revista de Difusión cultural y científica de la Universidad La Salle en Bolivia, 7*(7), 58-66. http://www.scielo.org.bo/pdf/rfer/v7n7/v7n7_a05.pdf

Universidad de Alcalá. (s.f.). R y Python en Data Science. https://www.master-data-scientist.com/r-python-data-science/#:~:text=Python%20es%20tambi%C3%A9n%20un%20lenguaje,usa%20tipado%20multiplataforma%20y%20din%C3%A1mico.

Vega-Rodríguez, L. Y. y Botero-Suaza, L. E. (2020). Formación profesional inmersa en la transformación digital con el fin de mejorar la calidad en la educación. *Cultura, Educación y Sociedad, 12*(1), 37–46. https://doi.org/10.17981/culteduso c.12.1.2021.03

Velásquez, L y Hitpass, B. (2014). El nivel de Actividad en el Proceso Educativo como Indicador de Riesgo de Deserción Estudiantil medido en tiempo real con apoyo de tecnología BAM. 10.13140/2.1.3217.8880

Villalobos, E. B., Cornejo, M. C. y Rivera, M. M. (2021). Estrategias virtuales de pandemia para abatir los índices de reprobación en asignaturas de Ciencias Básicas. *Revista Mapa, 5*(24). 71-91 https://revistamapa.org/index.php/es/article/view/289

Yanes, M., Silva, M. P. y Payró, P. (2022). Barreras afrontadas por profesores ante innovaciones tecnológicas por COVID-19: una visión desde la minería de datos. *Revista Dilemas Contemporáneos Educación Política y Valores,* 1-15. https://doi.org/10.46377/dilemas.v10i1.3312

Zarria, C., Arce, C., y Lam, J. (2016). Estudio de variables que influyen en la deserción de estudiantes universitarios de primer año, mediante minería de datos. *Ciencia Amazónica (Iquitos), 6*(1), 73-84. https://doi.org/10.22386/ca.v6i1.110

Glosario

F

FURPS: *Funcionality, Usability, Reliability, Performance & Suportability*

I

IES: Instituciones de Educación Superior

K

KDD: *Knowledge Discovery in Databases*

U

UNESCO: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura