



UNIVERSIDAD POPULAR AUTÓNOMA DEL ESTADO DE PUEBLA

**Minería de datos educativa aplicada en el
descubrimiento de patrones que inciden
en la deserción escolar universitaria**

Tesis que presenta

José Carlos Camino Hampshire

Matrícula: 17980002

Para obtener el grado de

**Maestro en Ciencia de Datos e
Inteligencia de Negocios**

Directora de Tesis

Dra. Argelia Berenice Urbina Nájera

Marzo, 2020



UPAEP – Secretaría General

Dirección General de Apoyos Académicos

Dirección del Centro de Recursos para el Aprendizaje y la Investigación.

Biblioteca Central - **Karol Wojtyla**

Tesis Digitales Restricciones de uso:

DERECHOS RESERVADOS ©

PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de textos, imágenes, gráficas, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente de donde la obtuvo mencionando el autor o autores involucrados en el documento.

Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



Hoja de liberación

Ciudad de Puebla, a 27 de febrero de 2020

Dra. Ma. del Rocío Guadalupe Morales Salgado

Directora de Programa Educativo

Maestría en Ciencia de Datos e Inteligencia de Negocios

UPAEP-Universidad

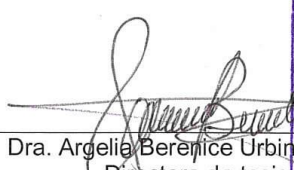
PRESENTE

Asunto: Aprobación para la digitalización
de tesis

Los que suscriben la presente, en calidad de comité asignado a la **tesis** del Lic. **José Carlos Camino Hampshire**, estudiante de la **Maestría en Ciencia de Datos e Inteligencia de Negocios** con número de matrícula **17980002**, le informamos que hemos revisado la **tesis** titulada **Minería de datos educativa aplicada en el descubrimiento de patrones que inciden en la deserción escolar universitaria**.

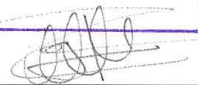
Como resultado de esta revisión consideramos que la **tesis** cumple con los requisitos necesarios para ser defendida con el propósito de obtener el grado de Maestro en **Ciencia de Datos e Inteligencia de Negocios**, por lo que aprobamos que se realice la digitalización formal correspondiente y por ende la programación del examen de grado.

La saludamos afectuosamente, anticipando nuestro agradecimiento por la atención a la presente.


Dra. Argelia Berenice Urbina Nájera
Directora de tesis


Cordialmente,
U UPAEP
Decanato de
Posgrados
Ingenierías
2


Dra. Martha Leticia Gaeta
González
Asesora


Dra. María del Rocío Guadalupe
Morales Salgado
Asesora

U UPAEP
POSGRADOS
17 SUR 901
Barrio de Santiago
Puebla, Pue. México
C.P. 72410

Tel.: 01 (222) 229 9400
Fax: 01 (222) 2325251
01 800 224 2200
www.upaep.mx



Agradecimientos

Este proyecto de Tesis y en general, de la maestría en ciencia de datos e inteligencia de negocios en su totalidad, son el reflejo del hombre que mis padres han forjado.

Agradezco a mi padre, José Carlos Camino López, por enseñarme con el ejemplo que el trabajo duro rinde siempre frutos, que hay que compartir con los demás lo que se tiene, no lo que nos sobra, y esto se refiere a lo material, al afecto y en este caso al conocimiento. Gracias por siempre confiar en mí y por ser uno de los pilares más importantes de mi vida para ser el hombre que soy, gracias por cuidarme, guiarme y estar siempre ahí cuando te necesité, y ahora, por hacerlo desde el cielo.

Agradezco a mi madre, Virginia Hampshire, por ser uno de los motores de mi vida, que siempre me ha impulsado para ser un mejor estudiante, un mejor hijo y una mejor persona. Ella siempre confió en que yo podría alcanzar esta meta y me motivó para lograrlo.

Agradezco a mis hermanos, Karla y Giancarlo Camino, pues siempre han estado de uno y otro modo a mi lado apoyando en lo que necesite. También agradezco a mis sobrinos Salvador y Emilio pues ellos me impulsan a ser un ejemplo para ellos.

Agradezco a todos mis familiares, en especial a mi Abuela Virginia de Hampshire, que siempre me brindan su confianza y amor incondicional.

Agradezco a mi esposa, Viridiana Hernández Quintana, por su amor y apoyo incondicional, por soportar mis largas jornadas, por ser mi *role model* de profesionalismo y perseverancia.

Agradezco también a mi empresa Accenture, por brindarme la flexibilidad para tomar clase, por su apoyo compartiendo casos de negocio, mejores prácticas, datos, para compartirlos con mis compañeros de clase y profesores, enriqueciendo las conversaciones con situaciones que se viven en el mundo empresarial.

Agradezco a toda la Comunidad Universitaria por habilitar los recursos necesarios para adquirir los conocimientos del área de ciencia de datos e inteligencia de negocios requeridos para alcanzar el grado de maestro. Pero en especial quiero agradecer a mi profesora, compañera de publicaciones y directora de tesis, la Doctora Argelia Berenice Urbina Nájera, por su profesionalismo, guía, liderazgo, solidaridad, conocimiento, entre otras muchas virtudes que ha demostrado y compartido conmigo, de principio a fin, a lo largo de este programa académico y del desarrollo de este proyecto de tesis. Sin su apoyo nada de esto hubiera sido posible.

“El conocimiento es poder, la información es libertadora, la educación es la premisa del progreso, en toda sociedad, en toda familia”

(Kofi Annan)



INDICE

CAPÍTULO I: PROPUESTA Y ORGANIZACIÓN	8
1.1 PLANTEAMIENTO DEL PROBLEMA	8
1.2 JUSTIFICACIÓN	10
1.3 OBJETIVO GENERAL	12
1.4 ALCANCES Y LIMITACIONES	12
1.5 VIABILIDAD DE LA INVESTIGACIÓN	12
1.6 RESULTADOS ESPERADOS	12
1.7 CONTRIBUCIONES ORIGINALES	13
CAPÍTULO II: MARCO TEÓRICO.....	14
2.1 ANTECEDENTES DE LA DESERCIÓN ESCOLAR UNIVERSITARIA	14
2.2 ESTUDIO SOBRE LAS CAUSAS DE LA DESERCIÓN ESCOLAR UNIVERSITARIA DESDE EL AÑO 2000 A LA FECHA	15
2.3 MECANISMOS UTILIZADOS EN LOS DOS ÚLTIMOS SEXENIOS (2006-2012, 2012-2018) PARA PREVENIR LA DESERCIÓN ESCOLAR UNIVERSITARIA.....	16
2.4 MINERÍA DE DATOS EDUCATIVA.....	22
2.5 ÁRBOLES DE DECISIÓN.....	22
2.6 MÁQUINAS DE SOPORTE VECTORIAL	24
2.7 PROCESO PARA EL DESCUBRIMIENTO DE CONOCIMIENTO	25
CAPÍTULO III: ESTADO DEL ARTE.....	27
3.1 DESERCIÓN ESCOLAR UNIVERSITARIA	27
3.2 MINERÍA DE DATOS EDUCATIVA EN LA DESERCIÓN UNIVERSITARIA	35
CAPÍTULO IV: METODOLOGÍA	38
4.1 DISEÑO DE LA INVESTIGACIÓN.....	38
4.2 SELECCIÓN DE LA POBLACIÓN Y MUESTRA.....	38
4.3 RECOLECCIÓN DE DATOS	39
4.3.1 Selección del Instrumento	39
4.3.2 Validez del Instrumento	40
4.3.3 Aplicación del Instrumento	40
4.4 PREPARACIÓN DE DATOS	40
4.5 APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EDUCATIVA PARA ANÁLISIS DE DATOS.....	41
CAPÍTULO V: RESULTADOS	43
5.1 SELECCIÓN DE ATRIBUTOS.....	43
5.2 ÁRBOLES DE DECISIÓN.....	43
5.3 MÁQUINA DE SOPORTE VECTORIAL	46
CAPÍTULO VI: ANÁLISIS Y DISCUSIÓN.....	48
CONCLUSIONES Y RECOMENDACIONES.....	51
REFERENCIAS.....	53

Resumen

Dada la relevancia de la deserción escolar universitaria para la sociedad, desde hace años se han realizado diferentes estudios que abordan la problemática desde el punto de vista cualitativo y en algunos casos con técnicas cuantitativas básicas. En años recientes, el uso de técnicas de minería de datos educativa ha cobrado relevancia por su capacidad para de forma cuantitativa, predecir el desempeño educativo, determinar perfiles de comportamiento, generar modelos de retención, fracaso escolar, entre otros. En este trabajo se muestra la aplicación del proceso de descubrimiento de conocimiento a un conjunto de datos de 500 estudiantes de Instituciones de Educación Superior públicas y privadas. Se utiliza el algoritmo selección de atributos para identificar los factores más importantes que inciden en la decisión de desertar; también, se utilizan máquinas de soporte vectorial y árboles de decisión (92.6% de exactitud) para clasificar a los estudiantes de acuerdo con sus atributos y definir patrones que consiguen alertar una inminente deserción que puedan derivar en acciones antes de que esta suceda. Se encontró que los tres factores más relevantes que un estudiante considera para tomar la decisión de desertar son la falta de asesorías, la falta de un ambiente estudiantil adecuado y la falta de seguimiento académico, mientras que un patrón para desertar incluye factores como: ambiente estudiantil, apoyos financieros insuficientes, experiencia de una situación incómoda, el servicio, entre otros. Con esto se confirma lo establecido por la literatura donde se puede afirmar que la deserción escolar universitaria es un fenómeno multifactorial.

Palabras clave: asesorías, ambiente estudiantil, árboles de decisión, seguimiento académico, máquinas de soporte vectorial, selección de atributos.

Abstract

Given the relevance of university dropouts for society, for years different studies have been carried out that address the problem from a qualitative point of view and in some cases with basic quantitative techniques. In recent years, the use of educational data mining techniques has become relevant because of their ability to quantitatively predict educational performance, determine behavioral profiles, generate retention models, and school failure, among others. This paper shows the application of the knowledge discovery process to a data set of 500 students from public and private Higher Education Institutions. The algorithm of attribute selection is used to identify the most important factors that influence the decision to drop out; also, vector support machines and decision trees are used (92.6% accuracy) to classify students according to their attributes and define patterns that manage to alert an imminent dropout that can lead to actions before it happens. It was found that the three most relevant factors that a student considers in making the decision to drop out are the lack of counseling, the lack of an adequate student environment, and the lack of academic follow-up, while a pattern of dropping out includes factors such as: student environment, insufficient financial support, experience of an uncomfortable situation, and service, among others. This confirms what has been established in the literature where it can be stated that dropping out of university is a multifactorial phenomenon.

Key words: academic follow-up, student environment, decision trees, support vector machines, feature selection, machine learning.

Introducción

México tiene la proporción más baja entre los países de la OECD (Por sus siglas en inglés *Organisation for Economic Co-operation and Development*) de adultos entre 25 y 64 años con un título de educación superior (17%) cifra inferior al promedio de la organización que es de 37% (OECD, 2019). Por ello, la deserción escolar toma gran relevancia como un problema de dimensiones considerables en México y específicamente en el estado de Puebla, el cual ocupa el sexto lugar nacional en este tema (González-Guzmán, 2017), de ahí que, en este estudio, dentro de otras cosas, se analizan las estrategias planteadas por el gobierno en los dos sexenios anteriores y en el actual en busca de reducir los índices de deserción en el sistema educativo mexicano.

De acuerdo con la Real Academia Española de la Lengua (2013) la deserción escolar es la acción de separarse o abandonar obligaciones en cuanto a compromisos escolares. Las diferentes aproximaciones al tema permiten comprender que existe una matriz amplia de factores de tipo psicológico, cultural, económico, social, institucional y pedagógico que explican la deserción universitaria.

Recientemente, el uso de técnicas de minería de datos educativa ha cobrado relevancia en el análisis de diversos aspectos educativos como la deserción escolar; su aplicación tiene como objetivo seguir la huella digital de los estudiantes y descubrir de manera oportuna un cambio en el comportamiento vinculado a aspectos académicos que puedan predecir, por ejemplo, una inminente deserción o abandono escolar. De este modo, el objetivo de este estudio es hacer uso de la minería de datos educativa mediante la aplicación de los algoritmos selección de atributos, árboles de decisión y máquinas de soporte vectorial, para identificar los principales factores que influyen en la deserción universitaria, a la vez, que se encuentran patrones para prevenirla, usando un conjunto de datos de 500 registros con 56 atributos divididos en 9 categorías, obtenidos a través de una encuesta administrada a estudiantes inscritos en Instituciones de Educación Superior (IES) públicas y privadas en el estado de Puebla, México.

Para ello, en el capítulo I de este documento se presenta la problemática de la deserción escolar en las IES y cómo esta se abordará en el resto del documento.

En el capítulo II se muestra el marco teórico de la tesis donde se analizan las estrategias propuestas por el gobierno mexicano de los dos sexenios anteriores y el actual a través de sus programas sectoriales de educación, enseguida se realiza una breve revisión de la literatura acerca de la deserción universitaria planteada desde dos aristas: metodologías tradicionales y aplicación de minería de datos educativa. Se define a la minería de datos educativa, su proceso y los algoritmos a utilizar resaltando su capacidad para identificar patrones nuevos y no triviales para resolver y mejorar procesos educativos para terminar con la descripción del proceso de descubrimiento de conocimiento en base de datos.

En el capítulo III se presenta el estado del arte donde se resume la literatura seleccionada desde el año 2000 (16 artículos) relacionada con la identificación de factores que inciden en la deserción escolar mediante métodos tradicionales. Del mismo modo se presenta revisión de literatura (35 artículos) donde se utiliza minería de datos educativa; identificando a las máquinas de soporte vectorial como el de mejor desempeño para clasificar conjuntos de datos relacionados con procesos educativos, el de menor



desempeño es regresión lineal y el más utilizado es árboles de decisión; sólo en 3 artículos se mostró evidencia de utilizar selección de atributos.

En el capítulo IV describe la metodología utilizada en este estudio la cual se fundamenta en el proceso de descubrimiento de conocimiento en bases de datos (Selección de la población y muestra, recolección de datos, preparación de datos y aplicación de técnicas de minería de datos educativa).

En el capítulo V se presentan los resultados donde se identifican los factores y patrones que inciden en una deserción escolar inminente, logrando concluir en acciones que favorecen una prevención oportuna de la misma. El estudio muestra que después de aplicar el algoritmo de selección de atributos se identifican 27 factores principales y que las técnicas de minería de datos educativa empleadas clasifican mejor una vez aplicada esta comparado con el desempeño utilizando el conjunto de datos completo.

En el capítulo VI se presenta una comparativa de los resultados del estudio con los de la literatura revisada encontrando diversas semejanzas y divergencias. Y finalmente, se presentan las conclusiones y recomendaciones a partir de los hallazgos encontrados.

Capítulo I: Propuesta y Organización

En esta sección se presenta la problemática de deserción escolar en Instituciones de Educación Superior que se aborda en este trabajo de Tesis; la SEP (2015) define a la deserción escolar como el número o porcentaje de estudiantes que abandonan las actividades escolares antes de terminar algún grado o nivel educativo, originada por diversos factores intrínsecos (personales) y extrínsecos (familiares, personales, sociales, económicos). Se presentan cifras de matrícula por nivel y se hace énfasis en la educación superior del estado de Puebla.

En este apartado se describe cómo se abordará la problemática a través de la identificación de patrones y factores más significativos que inciden en la deserción escolar apoyados de técnicas de minería de datos educativa con herramientas de uso libre.

1.1 Planteamiento del Problema

De acuerdo con la Real Academia Española de la Lengua (2013) la deserción escolar es la acción de separarse o abandonar obligaciones en cuanto a compromisos escolares. Mientras que la Secretaría de Educación Pública [SEP] (2015) lo expresa como el número o porcentaje de estudiantes que abandonan las actividades escolares antes de terminar algún grado o nivel educativo, originada por diversos factores intrínsecos (personales) y extrínsecos (Familiares, personales, sociales, económicos).

En México, como en otros países desarrollados o en vías de desarrollo, la deserción escolar es un problema que enfrentan todas las universidades. De acuerdo con la SEP (2015), tres de cada diez jóvenes entre 18 y 24 años tendrían espacio en el nivel de la Educación Superior. En el ciclo 2015-2016 sólo había 5 mil 343 escuelas de Educación Superior en un país de alrededor de 120 millones de habitantes según Ulloa Arellano (2017). En dicho periodo, había matriculados 3 millones 648 mil 945 alumnos en las diferentes modalidades de educación superior; de los cuales, 2.57 millones acuden a universidades e instituciones de educación superior públicas; mientras que 1.069 millones lo hacen en instituciones o universidades privadas (Hernández, 2017).

A estas cifras se suman 595 mil 456 estudiantes del nivel superior que están matriculados en modalidades “no escolarizadas”, de los cuales 504 mil 643 lo están en el nivel licenciatura y 90 mil 813 lo están en nivel de posgrado. En esta modalidad de enseñanza, la mayor oferta es privada, pues 331 mil 316 alumnos estudian en instituciones y universidades privadas y 264 mil 140 lo hacen en instituciones públicas, cuyo porcentaje de abandono escolar es mayor durante el primer año.

Según datos de la SEP (2016), Puebla se colocó en el sexto lugar a nivel nacional en el tema de abandono escolar, pues dos de cada diez alumnos que se inscribieron desertaron. El mayor nivel de deserción escolar fue en educación superior (10.5%), seguida de media superior (9.3%), secundaria (3.1%) y en primaria (0.2%). En otras palabras, de los 2 millones 026 mil 294 alumnos inscritos en dicho periodo, desertaron un 23.1%, lo que equivale a 468 mil 074 niños y jóvenes. Estas cifras consideran tanto a las escuelas públicas como privadas. Mientras que en el ciclo escolar 2016-2017 sumaron 219 mil 529 estudiantes en abandonar sus estudios en el nivel superior.

Específicamente en el estado de Puebla, de acuerdo con el Sistema Nacional de Información Estadística Educativa (2019), en su última actualización realizada en septiembre de 2017, el número de alumnos inscritos en IES es de 230,788 en el ciclo escolar 2016-2017, con un índice de deserción del 9.7%.

Por su parte, el Gobierno del Estado de Puebla a través de su Plan sectorial 2011-2017 propone estrategias y líneas de acción para combatir la deserción escolar en todos los niveles educativos. Particularmente, en la estrategia 1.7 su línea de acción es “Ofrecer mecanismos de apoyo oportuno a escuelas, grupos y alumnos con mayores deficiencias académicas, de acuerdo con los resultados en sus evaluaciones estandarizadas”.

Alineado a lo anterior, Instituciones gubernamentales como el Instituto Nacional para la Evaluación de la Educación (INEE) o la SEP Puebla presentan directrices para combatir la deserción escolar en el nivel básico y medio superior, así como directrices generales para todos los niveles educativos. Algunas directrices que el gobierno federal está siguiendo según El Informador (2019), están orientadas al otorgamiento de becas (10 millones para educación básica, media superior y superior) y la creación del programa “Universidades para el Bienestar Benito Juárez García”, donde se destinaron mil millones de pesos en el Presupuesto de Egresos de la Federación 2019 para la construcción de 100 universidades ubicadas en municipios de alta y muy alta marginación, que brinden servicios de educación superior gratuitos y de calidad a jóvenes con bachillerato terminado que no tienen posibilidades de ingresar a las instituciones de educación superior actualmente existentes. La meta del Programa es incorporar a un total de 300 mil estudiantes a la educación de nivel superior a lo largo de los seis años de gobierno.

El Gobierno Federal a través de la SEP (2013) en su Programa Sectorial de Educación 2013-2018, proponen estrategias y líneas de acción para combatir la deserción escolar en todos los niveles educativos. Particularmente, en la estrategia 1.6 “Utilizar la información derivada de las evaluaciones en la toma de decisiones para mejorar la calidad de la educación y evitar el abandono escolar”; en la estrategia 3.2 “Impulsar nuevas acciones educativas para prevenir y disminuir el abandono escolar en la educación media superior y superior”, destacando las líneas de acción 3.2.1 “Diseñar y utilizar herramientas para identificar tempranamente el riesgo de abandono en la educación media superior” y 3.2.2 “Capacitar a directores, docentes y tutores en el uso de herramientas que ayuden a evitar el abandono escolar en la educación media superior”.

Derivado de esto se crea un macroproyecto financiado por el Fondo de Investigación UPAEP, Convocatoria 2018, denominado “Mecanismos para reducir el índice de deserción escolar universitaria aplicando aprendizaje automático: Caso IES Puebla”, liderado por la Dra. Argelia Berenice Urbina Nájera y con la colaboración de un equipo interinstitucional adscritos a las universidades UPAEP, Universidad Tecnológica de la Mixteca y Universidad IEU. El cual se encuentra dividido en dos fases que se describen a continuación:

Primera fase

Objetivo: identificación de los patrones que inciden en la decisión de desertar en IES de Puebla, Utilizando una base de datos de una IES Privada en Puebla con más de 10,500 instancias y 12 atributos utilizando algoritmos de aprendizaje automático supervisado como árbol de decisión, máquina de soporte vectorial, Naïve Bayes, bosques aleatorios y selección de atributos

Segunda fase

Objetivo: Aplicar encuestas para descubrir el patrón de comportamiento que influye en los índices de deserción escolar de alumnos de educación superior del estado de Puebla

Es preciso mencionar que el proyecto de tesis planteado se enmarca en la segunda fase y, por tanto, busca ofrecer un mecanismo apoyado por minería de datos educativa que ayude a identificar analíticamente los patrones y variables que más impactan en la deserción escolar universitaria (licenciaturas e ingenierías).

1.2 Justificación

En las estadísticas de la SEP (2016), relacionadas con la educación a nivel superior, se reporta que el 10.5% de deserción escolar ocurre en el primer año de estudios universitarios.

Adicional a esto, Mateos Ayala (2018) identifica algunas consecuencias o efectos que se derivan de la deserción escolar dentro de las cuales destacan:

- La deserción escolar tiene un gran impacto en el capital humano, ya que afecta negativamente los procesos sociales, económicos y políticos en cuanto a la proyección de desarrollo del país.
- Generan elevados costos sociales y privados porque derivan en la captación de una fuerza de trabajo menos calificada.
- Para el Estado resulta costoso ya que este se ve en la obligación de realizar y financiar programas sociales.
- Incremento de las desigualdades sociales, lo que genera la falta de integración entre los individuos.
- La falta de educación impide que hombres y mujeres tengan la posibilidad de salir de la pobreza por no tener las capacidades mentales ni físicas para aspirar un trabajo estable y bien remunerado.

En el mismo tenor, Benitez (2012) sostiene que la deserción universitaria deja secuelas en los estudiantes, instituciones, sociedad y gobierno, mismas que se listan a continuación.

- Secuelas en el Estudiante:
 - Pérdida de la inversión hecha en material de estudio (libros, apuntes, fotocopias, útiles, viáticos, etc.).
 - Interrupción de su desarrollo personal y profesional.
 - Pérdida del acceso a un mejor nivel de vida.
 - Dificultad para insertarse en el mercado laboral y también para acceder a los puestos mejor remunerados.
 - Pérdida de contacto directo con otros jóvenes de igual edad con los que comparte intereses, aspiraciones, sueños, ideales, valores.
 - Descenso de los niveles de autoestima y confianza.
 - Perturbación de su comportamiento general y de la dinámica de todas sus relaciones (laboral, sentimental y familiar).
 - Trastornos físicos ligados con la frustración y el fracaso (insomnio, pérdida de apetito, dolores crónicos).

- Repercusiones para las Instituciones:
 - Pérdida importante de capital humano.
 - Malgasto y desperdicio de su presupuesto.
 - Desaprovechamiento de sus instalaciones e infraestructura.
 - Condicionamiento de su futuro porque la persistencia del problema las obliga a hacer más gastos y esfuerzos en el corto plazo.
 - Fomento de una sensación de temor y desconfianza hacia la institución.
 - Todas estas secuelas obligan a la institución a redoblar su labor y atención sobre los alumnos aún en carrera para compensar su pérdida.
 - Aumento del distanciamiento entre la enseñanza superior y la enseñanza media (se echan culpas mutuas).
 - Pérdida de tiempo y recursos al tener que repensar las estructuras académicas y pedagógicas más de lo conveniente y necesario.
 - Distanciamiento del aparato productivo que requiere mano de obra profesional.
- Efectos en la Sociedad y Gobierno:
 - Malgasto o desperdicio de recursos económicos irrecuperables y muy necesarios en otras áreas (salud, trabajo, defensa).
 - Pérdida de mano de obra profesional y especializada, lo que puede desencadenar problemas para atraer inversiones o encarar proyectos productivos.
 - Crecimiento de la cantidad de jóvenes calificados como “NiNi” (Ni estudia, Ni trabaja).
 - Aumento de los niveles de trabajo informal o en negro.
 - Decaimiento del valor de la educación como motor del progreso y el bienestar personal.
 - Aumento en los niveles de violencia social pues las frustraciones que sienten los jóvenes se trasladan con ellos a otros sectores de la comunidad.

De acuerdo con El Financiero (2017), una persona con licenciatura aumenta considerablemente su salario, siendo 80% superior al de una persona que sólo concluyó el bachillerato, además de reducir el riesgo de estar empleado en el sector informal en un 51%. En este mismo sentido, Molano (2017) calcula que el salario de una persona con licenciatura es alrededor del 75% superior al de una persona con bachillerato, y una persona con posgrado gana en promedio 194% más que una persona de bachillerato.

Una forma de abordar esta problemática es a través del campo de la minería de datos educativa, por lo que en la presente investigación se pretende ayudar a identificar los patrones o variables más significativas que influyen en los índices de deserción escolar, por medio de mecanismos de análisis aplicados a grandes conjuntos de datos a través de herramientas gratuitas. Para esto, se administraría una encuesta a estudiantes de IES, la cual ayudaría a identificar los patrones y variables latentes que inciden en la deserción universitaria. Consecuentemente, las IES públicas y privadas podrían utilizar los mecanismos derivados de la presente investigación para coadyuvar a reducir los índices de deserción escolar de manera significativa y colaborar en mejorar aspectos sociales, económicos y políticos de la entidad a la que pertenecen.

1.3 Objetivo General

Descubrir el patrón de comportamiento que influye en los índices de deserción escolar de alumnos de educación superior del estado de Puebla, por medio de algoritmos de minería de datos educativa.

Objetivos Específicos

- Aplicar el proceso de descubrimiento del conocimiento a los datos con el fin de leerlos adecuadamente en la herramienta de ciencia de datos aplicando suavizado, agregación, generalización y normalización.
- Identificar los factores que influyen en los índices de deserción en IES del estado de Puebla a través de la selección de atributos y árboles de decisión
- Predecir comportamientos futuros mediante la identificación de dependencias entre los factores que impactan la deserción escolar.

1.4 Alcances y Limitaciones

Esta investigación se enfoca en el estudio de la deserción escolar en el nivel superior del estado de Puebla ya que concentra el mayor número de universidades del País al tener 486 de las cuales 388 son privadas, mismo que se encuentra en el ranking de los 10 estados en tener el mayor índice de deserción en este nivel educativo. En este tenor, se pretende realizar el estudio de campo en el 1.7% de las universidades para recolectar datos de aproximadamente 500 estudiantes a fin de crear mecanismos que ayuden a prevenir de manera temprana una deserción escolar en el nivel superior, particularmente en el primer año de estudios de jóvenes universitarios.

1.5 Viabilidad de la investigación

Esta investigación es viable dado que se cuenta con recursos humanos y materiales necesarios para llevarla a cabo, además de tener relación con Instituciones de Educación Superior del Estado de Puebla en las cuales se aplicarán las encuestas que serán la fuente de datos para la aplicación de los métodos de minería de datos educativa.

En esta investigación se aplicarán 500 encuestas en diferentes universidades del Estado de Puebla, lo que permitirá generar resultados estadísticamente representativos (97.5% de nivel de confianza y 5% de error muestral admisible) sobre los factores que inciden en la deserción universitaria.

1.6 Resultados Esperados

El resultado que se espera obtener de esta investigación es una clasificación de los factores que inciden en la deserción escolar de Instituciones de Educación Superior del Estado de Puebla, además de la posibilidad de identificar patrones (combinación e interrelación entre factores) que determinen o pronostiquen si un estudiante es propenso a desertar en el primer año de estudio.



1.7 Contribuciones Originales

Se han identificado diversos trabajos donde se han realizado estudios analíticos sobre la deserción escolar de Instituciones de Educación Superior, pero esta investigación se refiere a estudiantes del Estado de Puebla, estrato poblacional del cual no se encuentra referencia bibliográfica y que puede ser una pauta para ciudades en donde la población estudiantil comparta las mismas características.

Capítulo II: Marco Teórico

En este capítulo se describen los antecedentes de la deserción escolar universitaria partiendo de su definición hacia la concepción de los factores que la provocan; así como, los mecanismos utilizados en los dos últimos sexenios para poder disminuir los índices a nivel global. Pudiendo determinar que el gobierno mexicano se interesa principalmente por el factor de cobertura (acceso a la educación) y al factor financiero a través de otorgar becas y apoyos, en todos los casos tomando en consideración los factores sociales enfocados en los sectores de la población más marginados y rezagados.

2.1 Antecedentes de la deserción escolar universitaria

De acuerdo con la Real Academia Española de la Lengua (2013) la deserción escolar es la acción de separarse o abandonar obligaciones en cuanto a compromisos escolares. Mientras que la SEP (2015) lo expresa como el número o porcentaje de estudiantes que abandonan las actividades escolares antes de terminar algún grado o nivel educativo, originada por diversos factores intrínsecos (personales) y extrínsecos (Familiares, personales, sociales, económicos).

De acuerdo con Páramo y Correa (1999), muchos autores confunden la deserción escolar con fenómenos inherentes del entorno académico como la mortalidad estudiantil, ausentismo y retiro forzoso. Ellos destacan que la deserción entendida tanto el abandono definitivo de las aulas de clase como el abandono de la formación académica es decisión personal del sujeto influenciado por circunstancias intrnas o externas a este, y no obedece a un retiro escolar forzoso como por ejemplo a través de una expulsión. Los autores señalan que las universidades tienen la obligación de establecer los mecanismos académicos, administrativos y de ajuste a la vida universitaria del estudiante para que supere en debida forma las dificultades de los programas académicos y que culminen satisfactoriamente sus estudios universitarios. Los autores distinguen los siguientes tipos de deserción:

- Total: Abandono definitivo de la formación académica individual.
- Discriminada por causas: De acuerdo con la causa de la decisión.
- Por facultad: Cambio de facultad o carrera.
- Por programa: Cambio de programa en la misma facultad.
- A primer semestre de carrera: Por inadecuada adaptación a la vida universitaria.

En el mismo tenor, Silva Laya (2011) establece que la deserción escolar universitaria se intensifica en mayor medida en el primer año de estudio, representando entre 20% y 30%, y puede llegar hasta el 60% de los casos en las distintas universidades. En su estudio se destacan factores que inciden en incrementar la deserción en el primer año como lo son la adaptación a la vida universitaria (integración sociocultural, mayor nivel de exigencia académica, relación con profesores) el contexto personal previo a ingresar a la universidad y las propias aspiraciones del estudiante (qué tan relevante es para él concluir una carrera universitaria). Las diferentes aproximaciones al tema del primer año universitario permiten comprender que existe una matriz amplia de factores de tipo psicológico, cultural, económico, social, institucional y pedagógico que explican la deserción universitaria. La autora también establece que los jóvenes con un origen social pobre y marginal, y que además tienen un bajo capital escolar previo, acarrean menores

posibilidades tanto de inscribirse en la universidad como de permanecer en ella. Silva Laya destaca que la atención a los estudiantes del primer año es fundamental para favorecer trayectorias exitosas, comenzando con el reconocimiento institucional de la importancia de este periodo para instrumentar mecanismos de apoyo a la integración académica y social, así como estrategias que transformen la experiencia en el aula para estimular un mayor compromiso del joven con sus estudios y el entorno académico.

2.2 Estudio sobre las causas de la deserción escolar universitaria desde el año 2000 a la fecha

Si bien el abandono escolar no obedece a una sola causa, sí hay una razón que origina la decisión de desertar. Para conocer sus causas, en este apartado se presentan 16 trabajos relacionados seleccionados bajo los siguientes criterios: a partir del año 2000 a la fecha, empleo de métodos estadísticos (enfoques mixtos, enfoques cualitativos, enfoques cuantitativos), centrados en la educación superior o educación terciaria (licenciatura, maestría y doctorado) para algunos países como México, Japón, Canadá, Estados Unidos y Nueva Zelanda. La tabla 1 muestra la lista de factores que se han identificado por medio de métodos estadísticos; los factores están ordenados alfabéticamente.

Factores	Autores
Creencias	Carvajal Olaya & Trejos Carpintero (2016)
Distancia y transporte (cobertura)	López Vélez (2004); Lavado & Gallegos (2005); Sandoval (2001); Abarca & Sánchez (2005); Lugo (2013)
Factores académicos	Ruíz (2009); Erazo, et. al. (2013); Carranza (2000); Londoño Ardila (2013); Rode, Bjornoy y Sogaard (2013); Lugo (2013); Fozdar, Kumar, & Kannan (2006); Vries, León Arenas, Romero Muñoz, & Hernández Saldana (2011); Carvajal Olaya & Trejos Carpintero (2016)
Factores económicos	Carranza (2000); López Vélez (2004); Ruíz (2009); Lavado, Gallegos (2005); Erazo, et al. (2013); Sandoval (2001); Cabrera, Bethencourt, Pérez & Alfonso (2006); Londoño Ardila (2013); Rode, Bjornoy y Sogaard (2013); Lugo (2013); Carvajal Olaya & Trejos Carpintero (2016); Cortés (2019)
Factores personales/familiares	Erazo, et al. (2013); Sandoval (2001); Cabrera, Bethencourt, Pérez & Afonso (2006); Londoño Ardila (2012); Rode, Bjornoy & Sogaard (2013); Lugo (2013); Fozdar, Kumar, & Kannan (2006); Carvajal Olaya & Trejos Carpintero (2016); Cortés (2019)
Factores sociales	López Vélez (2004); Ruíz (2009); Cabrera, Bethencourt, Pérez & Afonso (2006); Londoño Ardila (2013); Rode, Bjornoy & Sogaard (2013); Lugo (2013); Carvajal Olaya & Trejos Carpintero (2016); Cortés (2019)
Falta de orientación	Carranza (2000); Abarca & Sanchez (2005); Rode, Bjornoy & Sogaard (2013);
Institucionales	De los Santos (2004); Ruíz (2009); Abarca y Sánchez (2005); Cabrera, Bethencourt, Pérez & Afonso (2006); Londoño Ardila (2013); Rode, Bjornoy & Sogaard (2013); Fozdar, Kumar, & Kannan (2006); Carvajal Olaya & Trejos Carpintero (2016); Cortés (2019)
Seguridad	López Vélez (2004)
Selección/Vocación por la carrera	Abarca & Sanchez (2005); Lugo (2013); Vries, León Arenas, Romero Muñoz, & Hernández Saldana (2011)
Situación de salud	Erazo, et. Al. (2013); Lugo (2013)
Situación laboral	Carranza (2000); Ruíz (2009); Rode, Bjornoy & Sogaard (2013); Vries, León Arenas, Romero Muñoz, & Hernández Saldana (2011)

Tabla 1. Factores de la deserción universitaria identificados en 16 artículos

Fuente: Urbina-Nájera, et al. (2019)

Como se detalla en la figura 1 los factores con mayor frecuencia se encuentran los económicos (12), seguidos por los académicos, personales e institucionales (9) y factores sociales (8); las cifras no implican que, en los 16 trabajos revisados, estos los consideren como principales causas.

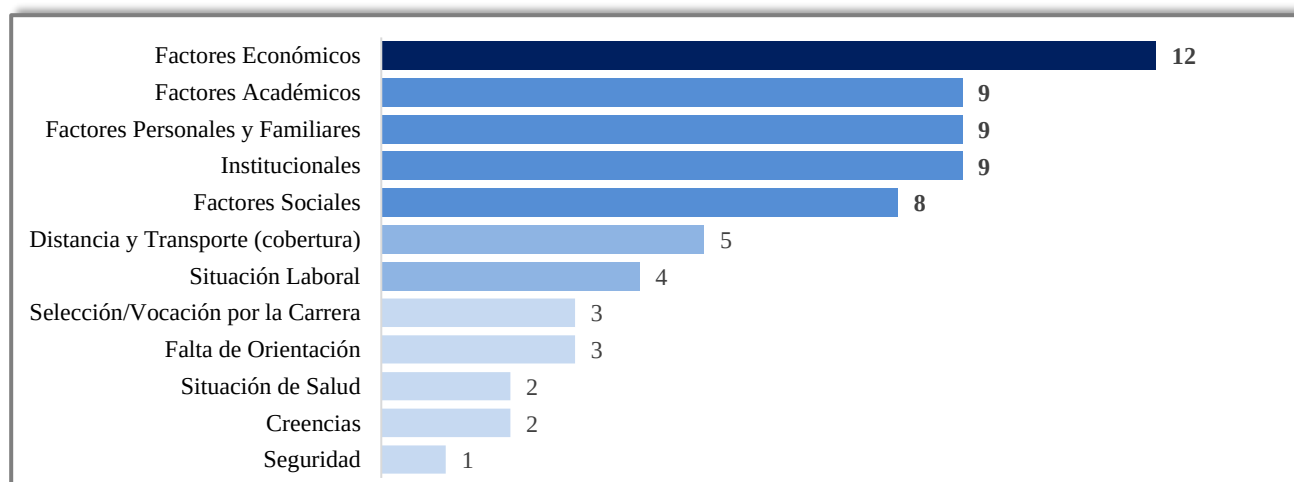


Figura 1. Frecuencia de aparición de los factores de deserción en artículos de referencia
 Fuente: Elaboración propia con información de Urbina-Nájera, et al. (2019)

Dentro de los factores económicos se han asociado, por mencionar algunos, la falta de recursos para pago de colegiaturas, falta de apoyos financieros, pérdida de empleo del padre o tutor, subestimación de costos. En la agrupación de los factores académicos se ubican aquellos relacionados a la asesoría, tutoría, acompañamiento u orientación insuficiente o inadecuada, bajo rendimiento académico personal, incumplimiento de reglamento académico, nivel académico de la universidad demasiado alto/bajo, oferta de asignaturas u horarios insuficientes. Dentro de los factores personales se encuentran los rasgos de personalidad del estudiante diferencias aquellos que completan sus estudios de aquellos que no, entre estos rasgos, están las intenciones, las normas subjetivas, las conductas de logro, el autoconcepto académico, la percepción de dificultad de los estudios y el nivel de aspiraciones, además de problemas dentro de su familia, separación de padres, entre otros. En la agrupación de factores institucionales se ubican las características de la institución universitaria, los servicios estudiantiles que ofrece, los indicadores de docencia, enseñanza y calidad, la infraestructura y las experiencias de los estudiantes en el aula. Los factores sociales abarcan situaciones como la integración del estudiante a la vida universitaria en cuanto a la relación con profesores, compañeros, y en general con los sistemas académicos y sociales de la universidad.

2.3 Mecanismos utilizados en los dos últimos sexenios (2006-2012, 2012-2018) para prevenir la deserción escolar universitaria

De acuerdo con la Organización para la Cooperación y el Desarrollo Económico (OCDE, 2019), México tiene la proporción más baja entre sus países miembro de adultos entre 25 y 64 años de edad en contar con un título universitario con tan solo 17%, siendo 37% el promedio a nivel global; en la región se ven cifras también por debajo del promedio como Argentina con el 21% y Chile, Colombia, Costa Rica con

el 23%. Por otro lado, la diversidad regional de México tiene un gran potencial, por lo que mantener una distribución geográfica equilibrada de las instituciones de educación superior ha sido tradicionalmente una de las prioridades de las políticas públicas y ha mejorado el acceso a la educación. No obstante, la ausencia de un mecanismo de planificación eficaz ha generado tensiones alrededor de la asignación de financiamiento y la capacidad de respuesta de la educación superior ante las necesidades regionales y locales. Por todo esto el gobierno debe implementar acciones que aborden la problemática de acceso y permanencia a la educación superior, además de garantizar una de calidad y pertinencia, para ello en cada sexenio se define un plan sectorial de educación donde se definen las estrategias y acciones a realizar en este rubro.

Dentro de los planes sectoriales de educación en los sexenios del Lic. Felipe Calderón Hinojosa y el Lic. Enrique Peña Nieto, se ha abordado el tema de la deserción escolar.

Específicamente en el Programa Sectorial de Educación 2007 – 2012 publicado por la SEP (2007) se establece que como parte de los retos que enfrenta México se encuentra que el sistema educativo presenta serias deficiencias, con altos índices de reprobación y deserción de los alumnos, y bajos niveles de aprovechamiento. El objetivo 2 dentro del programa busca una mayor igualdad de oportunidades educativas, de género, entre regiones y grupos sociales como indígenas, inmigrantes y emigrantes, personas con necesidades educativas especiales. Para lograrlo, es necesaria la ampliación de la cobertura, el apoyo al ingreso y la permanencia de los estudiantes en la escuela, el combate al rezago educativo y mejoras sustanciales a la calidad y la pertinencia. Dentro de las estrategias y líneas de acción enfocadas en combatir la deserción escolar o fomentar la permanencia se encuentran:

- 1.11.- Establecer servicios de orientación educativa, tutorías y atención a las necesidades de los alumnos, principalmente de aquéllos que están en riesgo de abandono o fracaso escolar.
- 1.15.- Fomentar la operación de programas de apoyo y atención diferenciada a los estudiantes, para favorecer su formación integral y mejorar su permanencia, egreso y titulación oportuna.
 - Contribuir al impulso de programas de tutoría y de acompañamiento académico de los estudiantes a lo largo de la trayectoria escolar para mejorar con oportunidad su aprendizaje y rendimiento académico.
 - Promover programas para que los alumnos terminen sus estudios en los tiempos previstos en los programas académicos, para incrementar los índices de titulación.
 - Apoyar la realización de programas de regularización para estudiantes de nuevo ingreso con deficiencias académicas, así como de programas orientados a desarrollar hábitos y habilidades de estudio.
 - Promover actividades de integración de los estudiantes de nuevo ingreso a la vida social, académica y cultural de las instituciones de educación superior, y fortalecer los programas de apoyo a la formación integral de los estudiantes.
 - Favorecer la simplificación de los procedimientos administrativos y de los trámites necesarios para la titulación, registro de título y expedición de cédula profesional.
 - Apoyar la instrumentación de acciones para facilitar la transición de la educación superior al empleo o, en su caso, al posgrado.

- Fomentar la realización de estudios para conocer mejor las características, necesidades, circunstancias y expectativas de los estudiantes, para desarrollar políticas de atención.
- 2.9.- Aumentar la cobertura de la educación media superior, dando prioridad a las entidades federativas con mayor rezago y demanda social.
 - Incrementar la cobertura y diversificar la oferta educativa, otorgando oportunidades de acceso y permanencia a mujeres y hombres.

Dentro del Programa Sectorial de Educación 2013 – 2018 publicado por la SEP (2013) en el sexenio del Lic. Enrique Peña Nieto se señala que la tarea educativa debe beneficiar a todos los grupos de la población, es por ello que el aumento de cobertura necesariamente debe ser complementado con acciones que atiendan las particularidades de grupos vulnerables que por razones diversas enfrentan barreras que impiden el acceso y la permanencia en la educación. También se propone establecer una comunicación estrecha con los padres de familia para que apoyen en proceso formativo de sus hijos, poner en juego estrategias de acompañamiento y apoyo a estudiantes, y reducir los factores de riesgo que amenazan la permanencia de los jóvenes en la escuela. Dentro de las estrategias y líneas de acción enfocadas en combatir la deserción escolar o fomentar la permanencia se encuentran:

- 1.6.- Utilizar la información derivada de las evaluaciones en la toma de decisiones para mejorar la calidad de la educación y evitar el abandono escolar.
 - Colaborar con el INEE para que la evaluación del sistema educativo pondere equilibradamente a todos sus componentes.
 - Fomentar junto con las entidades federativas la cultura del uso de las evaluaciones para la mejora del logro educativo.
 - Promover el uso de los resultados de las evaluaciones de logro para detectar y apoyar a alumnos con dificultades o rezagos en sus aprendizajes.
 - Proveer apoyos para favorecer la aceptación y facilitar el correcto análisis de las evaluaciones en las supervisiones y al interior de las escuelas.
 - Facilitar el establecimiento de redes de escuelas para compartir los aprendizajes obtenidos en las evaluaciones.
- 3.2.- Impulsar nuevas acciones educativas para prevenir y disminuir el abandono escolar en la educación media superior y superior.
 - Diseñar y utilizar herramientas para identificar tempranamente el riesgo de abandono en la educación.
 - Capacitar a directores, docentes y tutores en el uso de herramientas que ayuden a evitar el abandono escolar en la educación.
 - Promover programas de inducción e integración de estudiantes de nuevo ingreso al entorno escolar en la educación.
 - Impulsar programas de tutorías académicas, acompañamiento y cursos remediales de apoyo a los estudiantes de educación.
 - Fortalecer las modalidades de becas y otros estímulos dirigidos a favorecer la retención de los jóvenes en riesgo de abandono escolar.

- Realizar acciones para prevenir conductas de riesgo (violencia, adicciones, salud sexual y reproductiva, entre otros) que suelen desembocar en el truncamiento de la trayectoria educativa.
- Ajustar mecanismos académicos y administrativos que faciliten el tránsito entre modalidades y servicios educativos, y eviten discontinuidades en el trayecto escolar.
- Promover ambientes escolares libres de violencia, acoso y abuso sexual y favorecer una educación integral de los jóvenes.
- Propiciar el aprecio por las diversas expresiones de la cultura, la práctica del deporte y la sana convivencia, libre de discriminación y violencia.
- 3.6.- Promover la eliminación de barreras que limitan el acceso y la permanencia en la educación de grupos vulnerables.
 - Mejorar los mecanismos para identificar y atender oportunamente a las poblaciones excluidas del Sistema Educativo Nacional o en mayor riesgo de abandono.
 - Ampliar y mejorar los programas de becas destinados a atacar el abandono escolar en todos los tipos y modalidades de educación.
 - Dotar de útiles escolares a los alumnos de educación básica para apoyar su proceso educativo.
 - Fortalecer la efectividad de la educación para grupos vulnerables mediante el trabajo estrecho con las familias de la población atendida.

A su vez, el Lic. Andres Manuel López Obrador a través de su Plan Nacional de Desarrollo 2019 – 2024 publicado por la Secretaría de Gobernación (2019), establece una serie de objetivos, estrategias y líneas de acción respecto a la educación.

- 2.1.- Brindar atención prioritaria a grupos históricamente discriminados mediante acciones que permitan reducir las brechas de desigualdad sociales y territoriales.
 - Diseñar, desarrollar e implementar una política pública integral, multisectorial dirigida a la Primera Infancia, basada en atención especializada con énfasis en salud, nutrición, educación y cuidados.
- 2.2.- Garantizar el derecho a la educación laica, gratuita, incluyente, pertinente y de calidad en todos los tipos, niveles y modalidades del Sistema Educativo Nacional y para todas las personas.
 - Asegurar el acceso y permanencia en la educación, ofreciendo oportunidades educativas que tengan como eje principal el interés superior de las niñas, niños, adolescentes, priorizando a las mujeres, los pueblos indígenas y a los grupos históricamente discriminados.
 - Elevar la calidad y pertinencia de la educación en todos los tipos, niveles y modalidades del Sistema Educativo Nacional, considerando la accesibilidad universal y partiendo de las necesidades primordiales de la población y de la comunidad.
 - Mejorar la infraestructura básica y equipamiento de los espacios educativos en todos los tipos, niveles y modalidades del Sistema Educativo Nacional, generando condiciones adecuadas, de accesibilidad e incluyentes para el desarrollo integral de las actividades académicas y escolares.

En este sentido de acuerdo con el Informador (2019), también se desarrollan acciones orientadas al otorgamiento de becas (10 millones para educación básica, media superior y superior) y la creación del programa “Universidades para el Bienestar Benito Juárez García”, donde se destinaron mil millones de pesos en el Presupuesto de Egresos de la Federación 2019 para la construcción de 100 universidades ubicadas en municipios de alta y muy alta marginación, que brinden servicios de educación superior gratuitos y de calidad a jóvenes con bachillerato terminado que no tienen posibilidades de ingresar a las instituciones de educación superior actualmente existentes. La meta del Programa es incorporar a un total de 300 mil estudiantes a la educación de nivel superior a lo largo de los seis años de gobierno.

Discusión y Análisis

De acuerdo con datos de la OCDE, México cuenta con la proporción de adultos (25-64 años) con título universitario (17%), muy por debajo del promedio de los países miembro que es de 37%. Es por ello por lo que el gobierno debe proponer estrategias que deriven en políticas públicas para garantizar el acceso y permanencia en la educación, principalmente la educación superior, pues es bien sabido que una población bien preparada habilita un mayor crecimiento del país, además de forma individual, permite acceder a salarios mejor remunerados que los que puede aspirar una persona con estudios de nivel medio superior o inferior. De ahí que se presenten las políticas públicas propuestas por los dos sexenios anteriores.

El programa sectorial de educación propuesto en el sexenio del Lic. Felipe Calderón Hinojosa busca combatir principalmente los factores sociales de desigualdad de género y origen étnico, lo que conlleva a proponer acciones que también faciliten el acceso a la educación en términos de distancia a la escuela en zonas de mayor rezago y las propias creencias de los jóvenes y sus familias. Otros factores abordados en el programa sectorial 2007 – 2012 son el facilitar trámites administrativos (factores institucionales), factores académicos y de orientación vocacional a través de tutorías, así como el apoyo de los estudiantes para integrarse de forma satisfactoria a la vida social estudiantil.

El plan sectorial de educación 2013 – 2018 durante el sexenio del Lic. Enrique Peña Nieto también reconoce que existen diversos factores que inciden en la deserción escolar, aquí se pretende abordar principalmente el acceso a la educación de zonas marginadas, respetando las particularidades propias de las regiones, es decir, enfocando esfuerzos en factores de cobertura y creencias; también propone estrategias de apoyo y acompañamiento a estudiantes e incremento de becas y entrega de útiles, abordando los factores académicos y financieros; por último, propone crear un ambiente libre de violencia y riesgos como las drogas y abuso sexual, además de la propuesta de generar programas de inducción que permitan a los jóvenes de nuevo ingreso incorporarse satisfactoriamente a la vida estudiantil, esto atiende factores sociales y de seguridad. Un tema para destacar es que en este plan sectorial de educación se propone dar mayor énfasis a las evaluaciones académicas, tanto de estudiantes como de profesores, y generar herramientas que permitan analizar los resultados a detalle e identificar aquellos factores que más inciden en una educación de mala calidad o en su defecto, el abandono de esta.

En el caso del Lic. Andrés Manuel López Obrador, no se ha publicado un plan sectorial de educación, sin embargo, se proponen ciertas estrategias en su Plan Nacional de Desarrollo 2019 – 2024, en el cual

se atiende principalmente el acceso a la educación y la desigualdad social de grupos marginados en los que se incluye a las mujeres, niños, indígenas, entre otros, así como atender de manera eficiente temas de salud, nutrición y cuidados especiales, para lo cual se propone la construcción de nuevas escuelas y el correcto equipamiento de estas y las ya existentes, abordando con esto factores sociales, de cobertura, infraestructura y salud. Por otro lado, se genera una estrategia que busca otorgar becas y construir 100 universidades, con lo que se abordan factores financieros y nuevamente de cobertura.

Con lo anterior podemos ver que el gobierno mexicano se interesa principalmente por el factor de cobertura (acceso a la educación) y al factor financiero a través de otorgar becas y apoyos, en todos los casos tomando en consideración los factores sociales enfocados en los sectores de la población más marginados y rezagados. Los dos sexenios anteriores también ponen gran énfasis en apoyar a los estudiantes de nuevo ingreso en su proceso de incorporación al ambiente estudiantil y durante su estancia a través de tutorías, cursos y evaluaciones atendiendo de esta manera los factores académicos.

Uno de los hallazgos de esta revisión es que el gobierno mexicano desarrolla estrategias principalmente para la educación obligatoria o de manera general para todo el sistema educativo, direccionando estrategias de manera específica a la educación superior en pocas ocasiones.

Si bien el gobierno a lo largo de los años ha diseñado estrategias de alto nivel para garantizar el acceso a la educación y combatir la deserción escolar, estas no han generado un gran impacto en las cifras, pues mantenemos el último lugar de la OCDE de adultos con título universitario, además que la educación media superior, si bien se identifica una tendencia a la baja, la deserción escolar no ha reducido sus niveles significativamente tal como se muestra en la figura 2.

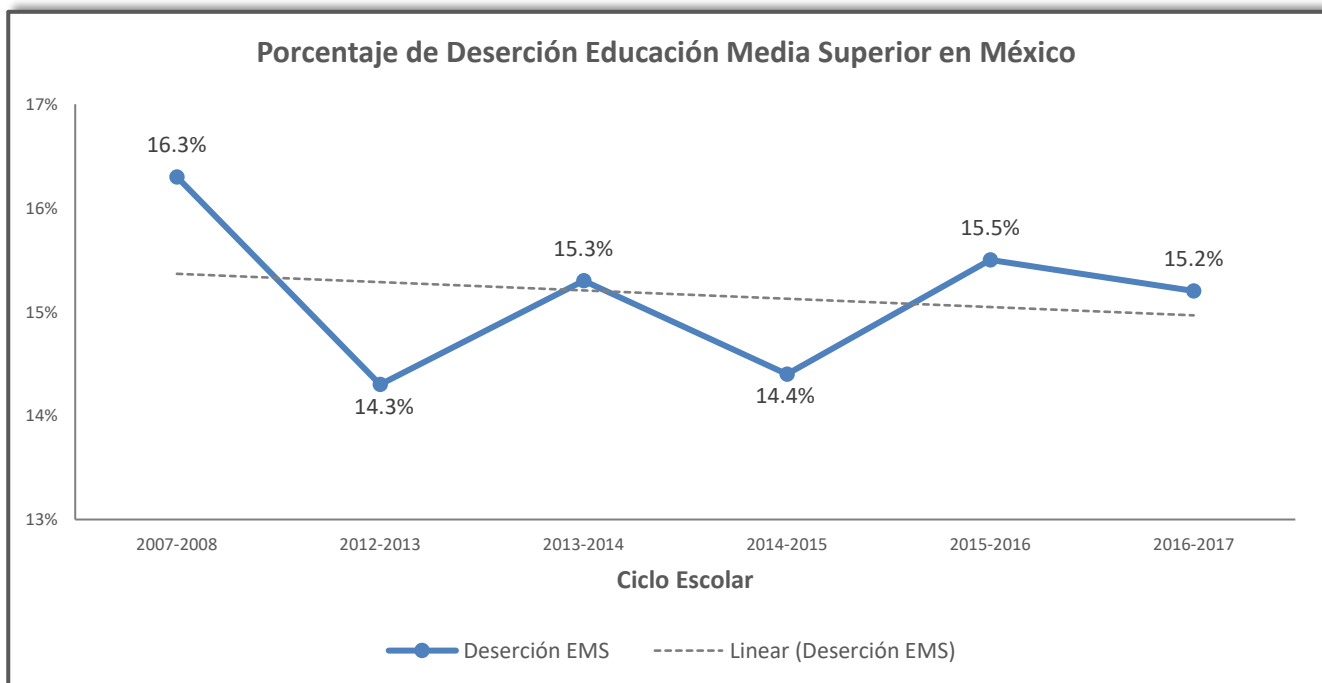


Figura 2. Porcentaje de deserción escolar en educación media superior en México
Fuente: Elaboración propia con datos de INEE (2019)

2.4 Minería de datos educativa

Dominguez, Álvarez, & Gil-Jaurena (2016) hablan de como el *Big Data*, entendido como la capacidad de gestionar grandes volúmenes de datos digitales a través de gran capacidad de almacenamiento y procesamiento, una vez inmiscuido en prácticas sociales tiene una conceptualización más allá de un simple artefacto tecnológico, es decir, es necesario dotar a la gestión de datos masivos de un contexto que ayude a interpretar su impacto en la sociedad. La capacidad de acceder directamente a gran cantidad de datos procedentes de la actividad social genera tensiones en la definición de los fines, en los procedimientos y, por extensión, en la propia consideración de la ciencia social basada en esos datos.

La educación es un ámbito propicio para experimentar con el proceso masivo de datos y su análisis, abriendo la posibilidad para mejorar el aprendizaje y apoyar la investigación sobre los procesos de enseñanza. En este sentido, la investigación educativa ha incorporando de manera habitual técnicas como el modelado de estudiantes, los sistemas de recomendación, la minería de datos o el análisis de la trazas de procesos educativos.

La minería de datos educativa se vale de diferentes técnicas para descubrir los patrones ocultos que se encuentran en los conjuntos de datos educativos, en este sentido los más utilizados son:

- Estadística: Se refiere a los reportes presentes en los sistemas de gestión de aprendizaje donde generalmente se presentan estadísticas descriptivas.
- Visualización de Información: Representaciones gráficas de la información que habilita una sencilla interpretación y ágil toma de decisiones comparado con información mostrada en texto plano o tablas.
- Análisis de Redes Sociales: A medida que las redes sociales se vuelven importantes para apoyar aprendizaje en red, las herramientas que permiten administrar, visualizar y analizar estas redes han ganado popularidad. Al representar visualmente una red social, se podrían establecer conexiones interesantes viendo y explorando en una forma fácil.
- Minería de Datos: Entendida como parte del proceso de descubrimiento de conocimiento en bases de datos (KDD, *Knowledge Discovery in Databases*), donde se encuentran técnicas de aprendizaje supervisado (clasificación y predicción, algunos se detallan en la siguiente sección), aprendizaje no supervisado (agrupamiento o *Clustering*) y reglas de asociación.

2.5 Árboles de Decisión

Los árboles de decisión se ubican dentro de una rama del aprendizaje computacional denominada aprendizaje simbólico, en la que también se encuentran los modelos de reglas de decisión, estrechamente relacionados con los árboles. El aprendizaje mediante árboles de decisión es una técnica que permite analizar decisiones secuenciales basadas en el uso de resultados y probabilidades asociadas. Mitchell (2000) lo define como “Un método de aproximación de una función objetivo de valores discretos en el cual la función objetivo es representada mediante un árbol de decisión”. Dicha representación se denota por un nodo de decisión, un nodo de probabilidad y una rama. El nodo de decisión representado por un

cuadrado, indica que una decisión necesita tomarse en ese punto del proceso. El nodo de probabilidad representado por un rectángulo redondeado, indica que en ese punto del proceso ocurre un evento aleatorio. Finalmente, la rama muestra los distintos caminos que se pueden emprender cuando se toma una decisión o bien ocurre algún evento aleatorio, este es representado por una línea.

En cuanto a la medición del desempeño de clasificación del árbol de decisión se comienza por definir la matriz de confusión, también llamada matriz de predicción o de clasificación, es una herramienta de visualización que se emplea para obtener información sobre las clasificaciones reales y predicciones realizadas por un sistema de clasificación (Bird, Klein, & Loper, 2009). Entonces, la matriz de confusión es una tabla donde cada celda $[i,j]$ indica cómo se clasificó alguna instancia con respecto a su clase preestablecida (Tabla 3). Los casos bien clasificados se encuentran en las entradas de la diagonal (en este caso las celdas $[a,d]$ de la Tabla 3) pues los grupos pronosticados y reales son los mismos; los elementos fuera de la diagonal se encuentran mal clasificados (Witten & Frank, 2005; Bird, Klein, & Loper, 2009).

		Predicción	
		Negativo	Positivo
Actual	Negativo	a	b
	Positivo	c	d

Tabla 3. Matriz de confusión cuando se tienen dos posibles resultados de clasificación: Negativo y Positivo.

En otras palabras, la matriz de confusión presenta cuatro posibles resultados obtenidos de una sola predicción para un caso de dos clases con clases "1" ("sí") y "0" ("no"), que son: verdaderos positivos (VP), negativos verdaderos (NV), falsos positivos (FP) y falsos negativos (FN). Un falso positivo es cuando el resultado se clasifica incorrectamente como "sí" (o "positivo"), cuando en realidad es "no" (o "negativo"). Un falso negativo es cuando el resultado se clasifica incorrectamente como negativo cuando en realidad es positivo. Los verdaderos positivos y verdaderos negativos son, evidentemente, las clasificaciones correctas (Freitas, 2002; Bird, Klein, & Loper, 2009). En la tabla 1, a representa el número de predicciones correctas cuando una instancia es negativa (NV); b es el número de predicciones incorrectas cuando una instancia es positiva (FP); c es el número de predicciones incorrectas cuando una instancia es negativa (FN); d es el número de predicciones correctas cuando una instancia es positiva (VP).

Por otro lado, las métricas frecuentemente usadas que se obtienen de la matriz de confusión son: Exactitud, Precisión, Recuperación, Medida F, cuyos conceptos se describen a continuación:

- **Exactitud (*Accuracy*):** Es la proporción del número total de predicciones que son correctas. Se determina aplicando la ecuación (AC). En otras palabras, la exactitud mide la fracción de casos en la diagonal de la matriz de confusión, se define con la ecuación $AC = \frac{a+d}{a+b+c+d}$
- **Precisión (*Precision*):** Algunas veces llamada consistencia o confianza. Es la proporción de la predicción de los casos positivos correctos, se calcula aplicando la ecuación $P = \frac{d}{b+d}$.

- Recuperación (*Recall*): También llamado verdadero positivo, completitud o sensibilidad. Se calcula empleando la ecuación $R = \frac{a}{p}$. Donde p se puede interpretar como el número de casos que deberían haber sido recuperados en función de algunos criterios de búsqueda.
- Medida F (*F-Measure*): Es una medida que combina la precisión con la recuperación para dar una puntuación única. Se define con la ecuación $MF = \frac{2 \times P \times R}{P+R}$ (Freitas, 2002; Witten & Frank, 2005; Bird, Klein, & Loper, 2009 y M. P. van der Aalst, 2011).

2.6 Máquinas de Soporte Vectorial

Tal como lo señalan Jimenez Moscovitz & Rengifo (2010) las Máquinas de Soporte Vectorial (SVM por sus siglas en inglés *Support Vector Machines*) son una técnica de aprendizaje automático que permite resolver problemas de clasificación y regresión de manera eficiente incluso con un número limitado de datos, con lo que se ha posicionado por encima de otras técnicas como las redes neuronales. Las SVM se basan en la Teoría del Aprendizaje Estadístico desarrollada por V. Vapnik y A. Chervonenkis en 1992.

Según explica Betancourt (2005) las SVM mapean las instancias en un espacio de características de una dimensión mayor a los puntos de entrada e identifican un hiperplano que maximice el margen entre clases a través de resolver un problema de programación cuadrática a través de la incorporación de multiplicadores de Lagrange. Las SVM utilizan algunos puntos de entrada denominados vectores de soporte para ser multiplicados por funciones de *Kernel*, de aquí que en algunas ocasiones se le denomine a esta técnica Máquinas de *Kernel*.

Existen casos linealmente separables y casos no linealmente separables, en este sentido en MathWorks (2019) se presentan las funciones de *Kernel* más utilizadas en las SVM (véase la tabla 4).

Tipo de SVM	Kernel
Función de base radial (RBF) o Gaussiana	$K(x_1, x_2) = \exp(-\ x_1 - x_2\ ^2 / 2\sigma^2)$
Lineal	$K(x_1, x_2) = x_1^T x_2$
Polinómica	$K(x_1, x_2) = (x_1^T x_2 + 1)$
Sigmoid	$K(x_1, x_2) = \tanh(\beta_0 x_1^T x_2 + \beta_1)$

Tabla 4. *Kernels* mas utilizados en las Máquinas de Soporte Vectorial
 Fuente: MathWorks (2019)

Al igual que los árboles de decisión, el desempeño de clasificación de las SVM se mide a través de la Exactitud, Precisión, Recuperación y Medida F.

De acuerdo con Betancourt (2005) las fortalezas de las SVM son entrenamiento de instancias relativamente sencillo, se escalan bien en espacios dimensionales altos, el compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente (por definición las SVM buscan minimizar el error de clasificación). Como debilidad establece que se necesita una “buena” función *kernel*, es decir, se necesitan metodologías eficientes para sintonizar los parámetros de inicialización de la SVM.

2.7 Proceso para el descubrimiento de conocimiento

El descubrimiento de conocimiento en bases de datos (KDD, *Knowledge Discovery in Databases*), se concentra principalmente en las siguientes tareas: colección de datos, preprocesamiento de los datos, selección de atributos y aplicación de algoritmos de aprendizaje computacional. Enseguida se describe cada una de estas tareas:

- **Colección de datos.** Los datos son toda aquella información con la que una computadora trabaja y procesa para obtener un resultado (ejemplo: texto, imágenes, documentos, etc.). Hoy en día, existe una gran cantidad de datos en diferentes formatos y diferentes bases de datos, tales como pagos, costos, activos, *kardex* de estudiantes, listas de asistencia, entre otros. El proceso de recolección de datos incluye todos los pasos necesarios para obtener los datos deseados en un formato digital. Los métodos de recolección de datos incluyen la adquisición y almacenamiento de las nuevas observaciones, consultar bases de datos existentes de acuerdo con el problema, y si es necesario realizar cualquier combinación de datos (Han, Kamber, & Pei, 2011). En el caso particular de este trabajo, los datos han sido recolectados en un formato electrónico, el proceso para ello se describe en las siguientes secciones.
- **Preprocesamiento de los datos.** Consiste en manipular, enriquecer, reducir o transformar los datos originales para ser accesibles, posteriormente, con mayor facilidad (Han, Kamber, & Pei, 2011). La codificación de los datos consiste en asignar valores de los campos de datos que describen las propiedades de cada instancia, pudiendo ser numéricos o nominales. Los atributos numéricos, a veces llamados continuos, toman números reales o enteros. Por otro lado, los atributos nominales pueden tomar valores de un conjunto finito establecido de antemano. Es posible transformar atributos numéricos a nominales y viceversa.
- **Posteriormente, los datos ya codificados pueden tener valores incorrectos o inclusive la ausencia de algunos.** Entonces, deben pasar por un proceso de limpieza que consiste para el primer caso, simplemente en eliminarse, mientras que para los valores faltantes se puede hacer interpolación usando los datos existentes (Witten & Frank, 2005). Luego, la fase de transformación implica combinar datos que residen en diferentes fuentes para proporcionar una visión unificada de estos, de manera que los datos se conviertan de un formato fuente a un formato destino en función de la herramienta a utilizar y poder ser cargados en ella sin inconvenientes de lectura.
- **Selección de atributos.** Es frecuente que se tenga un gran número de atributos para cada instancia en un conjunto de datos, sin embargo, puede ser que no todos sean relevantes para caracterizar al

objeto. De hecho, si se utilizan todos los atributos pueden, en muchos casos, causar un problema (Witten, Frank, & Hall, 2011). En otras palabras, el gran número de atributos representa un espacio de alta dimensión, por lo que es necesario llevar a cabo una reducción de la dimensionalidad, seleccionando sólo aquellos pocos atributos relevantes para el problema. Este pequeño conjunto de atributos debe conservar la mayor cantidad de información posible que describa a los ejemplos (Bishop, 2007).

- Aplicación de algoritmos de aprendizaje computacional. El aprendizaje computacional (AC) es la rama de la Inteligencia Artificial que se dedica al estudio de los agentes/programas que aprenden o evolucionan basados en su experiencia, para realizar una tarea determinada cada vez mejor (Mitchell, 2000). La aplicación de algoritmos de aprendizaje computacional es la última fase del proceso para el descubrimiento del conocimiento, cuyo objetivo principal es utilizar la evidencia conocida para poder crear una hipótesis y poder dar una respuesta a nuevas situaciones desconocidas, es decir, partiendo de este hecho se deben seleccionar aquellos algoritmos que den respuesta a diversas situaciones. En el caso de este trabajo se han considerado aplicar selección de atributos para identificar a los atributos más relevantes y árboles de decisión para determinar aquellos patrones que identifiquen el comportamiento que ocurre cuando un individuo es propenso a abandonar sus estudios universitarios; ambos algoritmos son descritos en las siguientes secciones.

Capítulo III: Estado del Arte

En este capítulo se aborda la revisión de la literatura acerca de la deserción escolar, así como de los factores que inciden en esta. De igual manera que se mencionan trabajos relacionados sobre la aplicación de la minería de datos como una solución para disminuir los índices de deserción en diversos contextos educativos.

3.1 Deserción escolar universitaria

Garza Carranza (2000) propone algunas soluciones a la problemática del evidente porcentaje de “deserción” y casos de pérdida académica en el Instituto Tecnológico de Nuevo Laredo. Expone el 31 % de deserciones escolares entre los años de 1996 y 1998 dato proporcionado por el Departamento de Servicios Escolares del propio instituto, lo cual denota una investigación de índole cuantitativa y propuestas cualitativas de abordaje y solución. La investigación identificó factores producidos por la institución únicamente, al igual que, se planteó la meta de disminuir el porcentaje de deserción y reprobación del 3% al 4%, entre los años de 1995 al 2000. Los resultados obtenidos son: La deserción escolar en el Instituto Tecnológico es una problemática latente desde su fundación, la población desertora presenta una serie de factores sociales y psicológicos de índole familiar y falta de apoyo motivacional por parte del Instituto. Según el estudio, la mayor deserción estudiantil se evidencia en el cambio del estudiante de secundaria a estudios superiores. El factor económico se presenta en el aumento del empleo remunerado entre los estudiantes, se denota un abandono en la comunidad estudiantil por la falta de recursos económicos para solventar su educación. El estudio describe la desarticulación entre los programas de estudio en bachillerato y los del ITNL, lo que genera un alto índice de pérdida académica en los primeros semestres, la investigación refiere que, la educación secundaria no cumple con la formación que requieren los programas de educación superior impartidos por el Instituto Tecnológico. A modo de solución a la problemática planteada se establece 1. Implementar programas de nivelación académica entre los niveles de secundaria y superior; 2. Programas de orientación administrativos y vocacionales dirigidos a los alumnos nuevos y antiguos con horarios flexibles; 3. Generar apoyos de instituciones públicas para el financiamiento de la educación superior a alumnos de bajos recursos económicos.

Navarro Sandoval (2001) presenta un acercamiento cualitativo y cuantitativo a modo de informe de investigación de campo, se enfoca en las principales causas de la inasistencia, abandono y aumento en la no incorporación al sistema escolar mexicano. Los datos estadísticos han sido consolidados por el XII Censo General de Población y Vivienda del año 2000. Los enfoques problemáticos en la educación mexicana son temas básicos en el panorama educativo gubernamental. Los últimos períodos presidenciales se han detenido en el fenómeno de manera cualitativa, sin proporcionar herramientas de cambio, o posibilidades verdaderas de estrategias contra el abandono y deserción escolar entre los jóvenes mexicanos. La autora establece que variables como, la marginación escolar vista y analizada como un alejamiento del sistema escolar, da como resultado condiciones sociales y económicas que llevan a la desigualdad educativa, donde no se concibe la educación como un derecho constitucional, sino como una proeza del destino para comunidades marginadas. La autora señala que, a mayor nivel

socioeconómico, el nivel de asistencia escolar de los grupos poblacionales es mayor, especialmente en jóvenes entre los 15 a 19 años es similar a países como Alemania, Francia y Japón. Existen factores psico-sociales como son: la motivación, las aspiraciones, y las características culturales de la región, que median y repercuten en el entorno y determinan los logros académicos de los jóvenes. La autora establece que es posible apoyar con estrategias de retorno a la escuela por parte gobierno y la SEP para que apoyen a jóvenes en formación temprana. También destaca que es necesario realizar estudios en el ámbito del hogar y en las instituciones escolares para identificar más factores que incidan en el abandono escolar.

De Los Santos (2004) presenta una investigación cualitativa y el uso metodológico de una exhaustiva revisión documental especializada, concluye temas importantes sobre la deserción escolar a nivel universitario. Clarifica el fenómeno de la “eficiencia terminal”, demostrando el interés y el gran número de estudios académicos existentes sobre este tema pedagógico. La definición de “eficiencia terminal” está fuertemente constituida bajo la doctrina de los alcances ideológicos de la producción industrial del siglo XX, y una fuerte caracterización militar por sus significantes homólogos como son: traición, abandono y huida, cuando se refiere al individuo/alumno en concreto. El autor analiza y compara el número de alumnos que se inscriben por primera vez en una carrera profesional formando una determinada generación, frente a los que logran egresar de la misma generación, y haber aprobado las asignaturas correspondientes a la carrera (Camarena, Chávez, & Gómez, 1984). La problemática de la “deserción” y con la conceptualización de “eficiencia terminal”, se demuestra como fenómeno común y demostrable en los contextos escolares tanto en la preparatoria como en la educación superior (Muñoz Izquierdo, 1979). Un problema concreto con tres visiones significativas: la primera el orden individual (sujeto que ejerce), familiar (contexto y psique cognitiva), y social e institucional (condiciones de ejecución), factores referidos al desempeño escolar de los alumnos y al desempeño de la institución en que están inscritos. La pregunta investigativa que subyace a las conclusiones es: ¿Cómo afrontar el fenómeno de la deserción al interior de las instituciones de educación superior?, la solución del autor es: ninguna estrategia de cambio y mediación debe de estar solitaria de un contexto, cada institución universitaria debe seleccionar sus acciones propias y adoptar varios planes de acción, nunca una única alternativa. Las alternativas que propone el autor son: un cambio en la asesoría académica y en momentos iniciales, orientación y servicios estudiantiles tempranos en la vida del estudiante, una adaptación al ámbito universitario desde la secundaria.

Vélez & López Jiménez (2004) realizan un análisis a manera de informe de investigación sobre las principales causas de deserción y mortalidad de los estudiantes de nivel superior, señalando conceptos como: “estrategias prácticas”, las que buscan ser homologadas y propuestas para ser implementadas en instituciones educativas, con la problemática de la deserción universitaria, estrategias que se implementaron en escenarios verdaderos y muy concretos. El escenario práctico fue la Universidad de La Sabana, Facultad de Comunicación Social y Periodismo, sede Bogotá, Colombia; el tema principal dio como resultado conclusiones como: las universidades buscan modelos netamente comerciales, orientados a una visión puramente administrativa, y abandonando la tarea social de la educación superior la cual es generar soluciones a los problemas de un país y su entorno social por medio de su capital humano. Los autores usaron una metodología cuantitativa / comparativa, y generaron una visión crítica en trabajos académicos aplicados en Estados Unidos, España, Argentina y Chile, los autores afirman que

la deserción estudiantil es un problema común en estos cinco contextos y en diferentes características del modelo educativo. Como temas concluyentes se afirma que la deserción universitaria es un elemento inherente y consecuente al proceso de educación en la educación superior (profesionalización), a su vez, postula que no sólo la responsabilidad es del estudiante poco motivado con un rendimiento estudiantil bajo, aunado a la poca orientación vocacional en déficit desde la preparatoria, sino también, a una falta de políticas educativas adecuadas, ingresos masivos a programas desarticulados con las realidades de países en formación y/o siguiendo modelos externos y ajenos a realidades de su entorno, es decir, la incompatibilidad en términos de contradicciones y discordancias entre lo que se enseña frente a lo que sirve a un país.

Abarca Rodriguez & Sanchez Vindas (2005) realizan una investigación de metodología exploratoria - demostrativa, desarrollada en la Universidad de Costa Rica. El tiempo de cobertura fue de cinco años específicos (1993-1998), con especial atención a la cohorte del año 1997. Se determinó como población objetivo la estudiantil desertora, tomando como bases tópicos analíticos: las posibles causas e influencias para que el estudiante abandonara sus estudios superiores. Como instrumento de investigación se usó la modalidad de entrevista cerrada, más de 450 entrevistas aplicadas a estudiantes mediante un cuestionario validado por los investigadores; y 30 entrevistas a expertos académicos de las distintas áreas administrativas y con alto categoría en la universidad. Los resultados fueron: 1. Los estudiantes presentaron un escaso conocimiento sobre las carreras en las que quedaron matriculados; 2. Los estudiantes no lograron ingresar a la carrera de su preferencia, convirtiéndose en la mayor variable que provoca el alejamiento de la catedra universitaria, esta razón alienta la determinación de abandono en la mayoría de casos, de la misma manera, los estudiantes abandonan la institución universitaria, pero se incorporan a otras opciones de educación sin abandonar el sistema universitario. Como conclusión no determinante, pero si existente, el factor económico no es un determinante en la decisión de deserción estudiantil. Los autores cuestionan para futuras investigaciones si será conveniente y urgente reinventar los mecanismos de ingreso a la carrera universitaria, además de dar prioridad a aquellos estudiantes que realmente muestren el interés de terminar sus estudios de educación superior.

Lavado & Gallegos (2005) realizan un informe de investigación, el cual analiza el fenómeno de la deserción escolar en el ciclo inicial y secundario en el Perú. La población objetivo fue la infantil entre los 9 a 16 años, en diferentes tipologías de riesgos o circunstancias de abandono de la escuela. El enfoque se presenta como explicativo – exploratorio. Lo contundente en los resultados resalta las limitaciones económicas encontradas en contextos rurales y urbanos. El estudio resalta con demostraciones paramétricas y no-paramétricas que analizan factores como: la década de los noventa en Perú se observó una mejora en matrícula escolar, pero perduró el atraso escolar y la deserción, afectando el nivel educativo de la sociedad durante las siguientes dos décadas. Del mismo modo, factores como el analfabetismo y la desescolarización de los años noventa del siglo XX, principalmente en adultos y madres cabeza de hogar; la pobreza en el campo y las ciudades, arrojan hallazgos a señalar como que la secundaria es el punto más crítico de deserción escolar. Al igual, el informe evidencia la falta en la oferta de escuelas, del mismo modo, la lejanía e inseguridad social de las mismas, es un factor que influye en la decisión de enviar a las niñas (población vulnerable, con un alto índice de embarazo adolescente en Sudamérica) a la escuela en el Perú. Los autores enfatizan en factores socio-geográficos de las escuelas,

demarcando que, los niños y niñas de las zonas urbanas son los que presentan una mayor probabilidad de deserción. En síntesis, el estudio es enfático en señalar la inexistencia en políticas educativas acordes con las necesidades del Perú lo que promueve y fortalece la problemática de deserción escolar. Los autores realizan un llamado a generar directrices de seguimiento al perfil del alumno y desarrollo de un currículo (factores como el multilingüismo- lenguas nativas en vías de extinción) que responda a las necesidades de todo un país.

Cabrera, Bethencourt, Alvarez, & González (2006) señalan la problemática española del abandono de los estudios universitarios, apuntan a una alarmante cifra del 50% en algunas carreras. Es un estudio inductivo con perfil mixto entre cuantitativo y cualitativo, de enfoque demostrativo - crítico. Hace el señalamiento de diferentes experiencias de estudios de universidades regionales y algunas europeas, mostrando los resultados de investigaciones sobre la deserción escolar, que ha alcanzado límites sin precedentes en la historia de la universidad española. Presentan diferentes causalidades y enfoques de la misma problemática. Demuestra la emergencia de cambiar el viejo modelo de universidad medieval (*universitas humanique*), a una universidad abierta, social, democrática y presta a las necesidades específicas del alumnado. La investigación de los autores demuestra una cantidad de estrategias prácticas, reconocidas y ratificadas en el éxito que han alcanzado, luego de realizar cambios profundos durante los últimos 15 años en España, soluciones a nombrar son: cursos de extensión, otras prácticas formativas interdisciplinarias, importancia de las prácticas deportivas, culturales, informativas, educación para todas las edades, una universidad con cursos de formación permanente, todas estas estrategias corroboradas en la práctica. El informe señala precisiones tan agudas como: el cimiento más importante de la universidad es el bienestar del alumnado, nunca olvidar que el abandono o deserción es el indicador de la calidad institucional, y hace el llamado a comprender la responsabilidad perpetua y perdurable de la escuela: es colocar todos los medios necesarios para atender y ayudar al alumno, sea la dificultad que sea.

Fozdar, Kumar, & Kannan (2006) presentan un estudio sobre las principales razones del abandono estudiantil en el Programa de Licenciatura en Ciencias de la Universidad Nacional Abierta Indira Gandhi. La primera, la distancia geográfica que el estudiante debe desplazarse. La segunda, el factor económico, dado el alto costo en las colegiaturas y la falta de subsidios estatales. La tercera, la falta de edificios – facultades con la infraestructura necesaria para albergar a los estudiantes. La cuarta, la ineficiente y poca información acerca de los programas educativos y las carreras universitarias que demandan los estudiantes. La no finalización o denominada abandono escolar, conduce a pérdidas financieras tanto para los estudiantes de como para la institución educativa. Los autores señalan que “la deserción estudiantil”, ha producido problemas psicológicos en algunos estudiantes que sienten que han fallado personalmente y a sus comunidades. Las sugerencias que presenta el estudio son: La Universidad y su facultad de Ciencias deben generar estrategias de apoyo al estudiante por medio de encuestas y entrevistas con ellos para poder evaluar el rendimiento de los alumnos y abordar los problemas relacionados con el programa desde los primeros años. El profesorado debe estar en contacto con los alumnos por medios tecnológicos, ya que, los presenciales son escasos y muy difíciles, denominados “aprendizaje a distancia”. Los docentes deberán implementar más sesiones de asesoramiento y estrategias de interacciones grupales para generar sentido de comunidad estudiantil, y actividades que el estudiante pueda adelantar en su comunidad. Los autores apuntan que son los estudiantes de dicho programa los

principales interesados en generar un sistema de apoyo para las futuras generaciones de estudiantes en la facultad de Ciencias.

Ruiz Calderón (2009) explica y analiza la problemática de la deserción en la Educación Superior en Nicaragua desde la Universidad de las Regiones Autónomas de la Costa Caribe Nicaragüense (URACCAN). La contribución de la autora está dirigida a la elaboración de propuestas en tres niveles reales del problema: El institucional, personal, y social. Las preguntas generadoras fueron, ¿Cuáles son los índices de deserción y tasa de titulación de URACCAN Las Minas en el periodo 2001-2007?, ¿Qué factores personales y/o familiares están asociados a la deserción? y ¿Cuáles son los desacuerdos que tienen los desertores frente a la academia, y el costo real de la deserción para la Universidad Autónoma Atlántico Norte? La autora presenta un análisis cuantitativo – demostrativo, la finalidad es realizar una propuesta de análisis de la problemática de la deserción estudiantil superior. La deserción universitaria se constituye como un problema serio de carácter nacional. Representa una pérdida de recursos humanos, técnicos y de infraestructura, los cuales se cuantifican en altos costos económicos. La autora encontró un desgaste, una desmoralización y abatimiento en el recurso humano administrativo que atiende los casos de deserción estudiantil, por otro lado, el capital humano de un país pierde porque no se constituye a futuro como uno capacitado para el desarrollo del país, en este caso Nicaragua. Son importantes las recomendaciones de la autora con respecto a disminuir la deserción en la educación superior con estrategias aplicadas a la juventud en los recintos escolares de básica a secundaria, por otra parte, Ruiz Calderón insta en el fortalecimiento de los programas de divulgación y orientación vocacional en los jóvenes. Al igual que, mejorar la calidad en la divulgación de los contenidos de las carreras profesionales, y el mejoramiento de infraestructura física de las instituciones en sus diferentes espacios académicos.

La terminología de “deserción estudiantil” y la preocupación a nivel universitario es el tema central de la investigación de De Vires, León Arenas, Romero Muñoz, & Hernández Saldaña (2011). Demuestran una tasa del 50% de abandono en la población que no logra concluir sus estudios, pero ¿Qué sucede si el estudiante se traslada de carrera y se mantiene en el sistema de educación superior?, con esta perspectiva los autores tratan de demostrar que según desde donde se mire la problemática, esta cambia, no se llama “desertor” al estudiante que se mantiene en el sistema, sino, aquel que abandona de manera total la educación superior. La metodología fue cualitativa con análisis documental y uso de entrevistas semiestructuradas. Los autores señalan en gran medida la incompatibilidad de estudios y trabajo remunerado, para solventar los costos de manutención y vida universitaria, otro factor que señalan es la falta de vocación vista como compromiso estudiantil en niveles superiores. Estos dos últimos factores, el económico aunado con el vocacional genera un alto índice en la reprobación y la pérdida de estudiantes en los diferentes programas educativos. De igual modo, las instituciones educativas tienen que privilegiar la eficiencia terminal o la tasa de titulación, estas influyen en el prestigio, la acreditación y el financiamiento de estas. El estudio es enfático en señalar y criticar una contradicción en las políticas de una universidad pública latinoamericana, la cual establece apoyar a los jóvenes de bajos recursos económicos con “colegiaturas bajas y becas”, pero la contradicción radica en la exigencia de tiempo completo, eso acarrea que los estudiantes de bajos ingresos se “retiren” o sean “retirados” no por pobreza, sino, porque resulta imposible combinar los estudios con un trabajo para manutención. En conclusión, el principal problema, en términos del estudio, es “la decepción” que siente el estudiante con el pobre apoyo

de la carrera que él escogió, y por falta de flexibilidad de la universidad para ayudar a su razón de ser: el estudiante.

La temática central que presentan Erazo, Valle, Rojo, Campillo, & Ruiz (2013) corresponde a la preocupación mundial de la deserción estudiantil a nivel superior, su abordaje parte de las condiciones estructurales de las universidades mexicanas, en las perspectivas académicas y administrativas, al igual que, las características sociales e intelectuales de la población estudiantil, en el estudio de caso la población objetivo fue: los estudiantes de las áreas de ciencias biológicas, químicas y ciencias de la salud de la UNAM. El análisis se realizó en el año 2012 y 2013 con el objetivo de identificar los motivos de abandono, en niveles iniciales, medios y terminales del ciclo escolar profesional; La metodología aplicada fue cualitativa con grupos focales. Los resultados se obtuvieron en términos de “los motivos principales de abandono”. El estudio postula las sugerencias de tipo práctico y medible para disminuir el abandono estudiantil. El motivo principal de abandono fue descrito desde el punto de vista económico, la falta de recursos se enfatiza en la necesidad de trabajar de los estudiantes para solventar los costos de carrera, seguido por los problemas de salud del estudiante. Los motivos secundarios fueron la falta de orientación vocacional y deficiente preparación académica desde la secundaria. Las sugerencias relevantes de la investigación se encuentran desde la orientación profesional a partir del bachillerato, seguido por mejorar los servicios administrativos en la universidad para optimizar la información en temáticas como: cambios de carrera, cambios a sistemas de educación alternativos como la educación abierta y a distancia. Los autores finalizan a modo de conclusión: primero, la decisión de abandonar es “multifactorial” aunque predomina el de tipo económico, determinado por la compra de materiales de trabajo y libros para las materias, segundo, razones académicas debido a la falta de conocimiento del contenido de cada carrera, desinformación que desmotiva al estudiantado convirtiéndolo también en factor de abandono.

Londoño Ardila (2013) señala y describe un número determinado de factores de riesgo en los ámbitos personales, académicos, institucionales y socioeconómicos relacionados con “la deserción estudiantil” ubicado espacialmente en la Corporación Universitaria Lasallista, Departamento de Caldas, Colombia. Como muestra poblacional total fueron 281 estudiantes activos y que ejercen sus estudios en los diferentes programas académicos en el año 2010. Como población objetivo central fueron 31 estudiantes que desertaron durante el año 2009. El método investigativo fue de carácter descriptivo, utilizado metodología cuantitativa en frecuencias, medidas y elaboración de tendencias que describían factores conductuales. Los factores de riesgo mostraron que “la deserción académica” está basada en denominadores socioeconómicos. El fenómeno de la deserción escolar según el autor está vinculado con las inestables condiciones económicas del estudiante. Es deficiente el tratamiento administrativo desde la institución universitaria frente a las situaciones de deserción estudiantil. Al mismo tiempo que, en el ámbito psicosocial, el autor plantea como variable la baja escolaridad de los padres, y cómo esta influye en “la deserción académica”, se establece una tendencia en los estudiantes retirados, cuyos padres los motivan a valorar más el trabajo remunerado que la capacitación profesional. Otro factor que señala el autor es la percepción negativa del nivel académico de los docentes por parte de los estudiantes que desertaron de la institución. Como conclusión el autor señala la obligación de las instituciones de educación superior en analizar y mejorar la modalidad de contratación de sus docentes, pues,

vinculaciones por tiempos y salarios limitados dificulta la sostenibilidad de actividades curriculares de calidad en beneficio de la comunidad académica.

En la siguiente disertación académica, Lugo (2013) presenta la metodología cualitativa con técnicas de revisión documental y de carácter analítico. Propone una reflexión sobre la dimensión de la deserción estudiantil en el ámbito venezolano, luego de realizar una revisión bibliográfica de autores que han tratado el tema en Latinoamérica, la autora señala que dicha problemática de la deserción afecta no sólo a niños y jóvenes, sino, también representa una pérdida económica para el Estado. El eje conceptual de la autora es la poca motivación que presenta el estudiante frente a su carrera, debido al desconocimiento de su finalidad, su exigencia y su campo de acción profesional, es decir, el alumno ingresa a la educación superior en Venezuela sin tener una información completa de la carrera escogida. La autora continúa señalando cómo los problemas de rendimiento académico en la universidad son asociados a los malos hábitos de estudio adquiridos en una pobre educación secundaria. También apunta que el evidente “abandono o fracaso” escolar que afronta Venezuela redundo en una indudable desigualdad social, económica y cultural. La afirmación de la autora señala que luego de dicha deserción los ex – estudiantes muestran una incapacidad para la socialización, una falencia en volver a adquirir una formación académica/cognitiva que les posibilite el acceso a un mejor empleo y poder participar en el proceso productivo del país. Su teoría es enfática, dichos estudiantes que han pasado por la deserción académica llegan a presentar una incapacidad, una tendencia a la marginación económica y social, y al inicio de un nuevo ciclo de desigualdades sociales. En una cadena de eventos, lo anterior conlleva a deficiencias en el bienestar material de los niños y adolescentes de estratos que no tendrán la oportunidad de ir a la universidad y alcanzar la capacitación superior.

Malene, Bjornoy, Sommersel, & Sogaard Larsen (2013) presentan en esta publicación un extracto de un informe técnico llamado “Fenómenos de deserción en las universidades: ¿Qué es el abandono? ¿Por qué ocurre la deserción? ¿Qué pueden hacer las universidades para prevenirlo o reducirlo?”. El objetivo del presente fragmento de investigación es mostrar los fenómenos de abandono en las universidades europeas en una forma breve y precisa. Entre tanto, distingue entre el abandono voluntario que realiza el estudiante a otra carrera, y el abandono involuntario que lleva al abandono formal de la educación superior por completo. La deserción formal se adjudica a factores preuniversitarios como: lugar de origen, situación social, y rendimiento académico en la secundaria. Variables como la transferencia voluntaria del estudiante a otra institución educativa son factores que la universidad no puede atender por falta de planeación. Los autores afirman que los dos principales momentos de abandono son: el inicial en los primeros años que es diferente y menos complejo que el abandono posterior, en el transcurso de la carrera. Pocos estudios han podido analizar los posibles determinantes del abandono universitario dentro de una perspectiva del tiempo como variable en las universidades europeas. Sin embargo, la deserción académica en etapas universitarias esta mediada por factores preuniversitarios, es decir, el enfoque académico en la escuela secundaria; y en condiciones externas a la universidad (situación económica, búsqueda de empleo) se destacan para la deserción posterior, en etapas terminales del ciclo académico. Un factor de mucho valor y preponderante es el desempleo. Los efectos perjudiciales de la deserción estudiantil en comparación con la tasa de graduación son preocupantes y han generado estudios sobre su repercusión en las tasas de empleo formal en los países de la unión europea.

Carvajal Olaya & Trejos Carpintero (2016) realizan una investigación de tipo análisis documental extraída del repositorio ALFA GUIA, compuesto por las 120 investigaciones avanzadas sobre educación en Latinoamérica, y cuya línea temática señala los tipos, factores y perfiles asociados al abandono y deserción académica. Los autores presentan en su análisis las probables causas de “deserción estudiantil” en Latinoamérica como en algunos países europeos; basados en las perspectivas analíticas del Estructuralismo Genético del sociólogo Pierre Bourdieu. El análisis del estudio fija primero, que el crecimiento de la clase media en los países genera el aumento en la demanda de un mejor sistema educativo en los países analizados; segundo, dichos sistemas de educación latinoamericanos no ofertan una educación de calidad capaz de garantizar el tránsito estudiantil hasta terminar el proceso de graduación del estudiante; tercero, la falta de reestructuración educativa de los sistemas nacionales genera la deserción estudiantil. En conclusión, la investigación señala que el éxito en el rendimiento académico depende de la clase social a la que pertenece el estudiante, es decir, el sistema educativo favorece a aquellos estudiantes que traen mejores “capitales” (económico, cultural, social) desde el aula de clase en su formación inicial y secundaria. “Capitales” concepto visto como un “valor” básico en una educación de calidad, dichas calidades son acumuladas a lo largo de vida del estudiante y dependen de la clase social en la que haya nacido. Por tanto, el individuo cuidará su “capital”, su saber y tenderá a conservarlo y mejorarlo, el cual lo ayudará a escalar en las estructuras sociales de su país, de manera contraria, un estudiante sin un “capital” que no sea apreciado por él, tenderá a desfallecer y se retira del sistema sin importar la pérdida del valor y la caída en la estructura social planteada por su entorno.

Cortés, Álvarez, Llanos, & Castillo (2019) realizan su estudio basado en una metodología cualitativa de orden mixto, con un fuerte análisis documental de experiencias de múltiples universidades, en lo concerniente a las razones y variables asociadas a los altos índices de deserción estudiantil (entiéndase como: abandono), encontrada en universidades de la República de Chile. Los resultados explican cinco factores relacionados con las causas de la deserción: primero, el económico, referido a altos costos en las matrículas y altos intereses en la financiación de créditos, la situación se complica al presentarse el abandono estudiantil y permanecer la deuda adquirida con el sistema bancario chileno; segundo, el factor social ubicado en la falta de personal con los perfiles requeridos; tercero, la proliferación de nuevas instituciones educativas universidades con muy poco tiempo en la esfera educativa del país, con una diversidad de currículos sin rigurosidad comprobada, así mismo, con altos costos de sus colegiaturas; cuarto, el progresivo aumento del desempleo en las personas que realizaron dicho abandono; quinto, la falta de mano de obra calificada. La situación de la deserción universitaria en Latinoamérica es una problemática del hemisferio, países que sufren los altos costos económicos y sociales debido a un fenómeno que no respeta fronteras geográficas ni étnicas. La deserción estudiantil es un fenómeno sociocultural que necesita mucha más investigación en Latinoamérica, especialmente en la educación superior. Como gran apartado, la investigación propone a los futuros investigadores realicen sus trabajos formales con la participación de múltiples actores sociales como: gobierno, empresas privadas y organismos internacionales, sin dejarle únicamente la responsabilidad a la academia como promotora de soluciones.

3.2 Minería de datos educativa en la deserción universitaria

Después del año 2000 se han empleado otras técnicas científicas para identificar aquellas causas que originan la deserción escolar no solamente la universitaria, también la ocurrida en secundaria y preparatoria. Particularmente, el uso de técnicas de inteligencia artificial ha cobrado gran importancia, pues hay estudios donde se describe la aplicación de algoritmos de aprendizaje computacional para dar solución a alguna situación en el contexto educativo. En este sentido, se han identificado los algoritmos comúnmente utilizados: *Näive Bayes*, redes neuronales, vecinos más cercanos, regresión lineal y logística, máquinas de soporte vectorial, selección de atributos y árbol de decisión.

En la Tabla 5 se presentan trabajos de 2003 a 2019 relacionados con la aplicación del algoritmo árboles de decisión en la solución de situaciones educativas en el nivel superior. Siendo 82.87% el mayor porcentaje de predicción reportado en Yamao, Saavedra, Campos Pérez, & Hurtado (2018).

Situaciones Educativas	Autores
Predicción del desempeño	Kabra & Bichkar (2011); Vijayalakshmi & Kumar (2011); Barbosa Manhães, Serra da Cruz, & Zimbrão (2014); Al-Barrak & Al-Razgan (2016); Agaoglu (2016); Chiheb, Boumahdi, Bouarfa, & Boukraa (2017)
Captación de matrícula en IES particulares	Estrada-Danell, Zamarripa-Franco, Zúñiga-Garay, & Martínez-Trejo (2016)
Éxito académico	Morales Carrillo & Parraga-Alava (2018)
Predicción de la deserción escolar universitaria	Kotsiantis, Pierrakeas, & Pintelas (2003); Dekker, Pechenizki, & Vleeshouwers (2009); Márquez, Cano, Romero, & Ventura (2012); Yukselturk, Ozekes, & Kılıç Türel (2014); Abu-Oda & El-Halees (2015); Márquez-Vera et al. (2016); Sivakumar, Venkataraman, & Selvaraj (2016); Yamao, Saavedra, Campos Pérez & Hurtado (2018)
Modelos predictivos de la deserción escolar universitaria	Aulck, Velagapudi, Blumenstock, & West (2017); Rodríguez-Maya, Lara-Álvarez, May-Tzuc, & Suárez-Carranza (2017);
Retención	Raju & Schumacker (2015); Delen (2011); Kumar Yadav, Bharadwaj, & Pal (2012);
Perfiles de comportamiento	Guevara, et al. (2019)
Fracaso escolar	Márquez, Cano, Romero, & Ventura (2012)
Estudiantes en riesgo	Er (2012)
Disminución de la tasa de abandono	Pal (2012)

Tabla 5. Situaciones educativas resueltas aplicando árboles de decisión de 2003 a 2019.

Fuente: Urbina Nájera & Camino Hampshire (2019)

Después de la revisión presentada en la Tabla 5, se observa que solamente tres trabajos han aplicado el algoritmo “selección de atributos” para identificar aquellos factores que inciden en la deserción escolar. En el primero, utilizaron un conjunto de datos con 77 atributos (características) de 670 jóvenes entre 15-18 años; aplicaron 10 métodos para obtener el ranking de los factores predominantes, considerando solo aquellos que tuviesen una frecuencia mayor o igual a 2, seleccionando solamente 15 atributos de los 77 atributos totales entre los que destacan, según su importancia: calificación en diversas áreas del conocimiento (ocho atributos), nivel de motivación, puntaje obtenido en educación secundaria, edad, número de hermanos, grupo, hábitos de fumar y promedio del examen de admisión (Márquez-Vera, Romero Morales, & Ventura Soto, 2012).

En el segundo, emplearon un conjunto de datos con 20 atributos de 1,966 jóvenes; aplicando selección de atributos obtuvieron los factores más importantes en la deserción: Género, aquellos relacionados con el total de créditos por periodo, créditos aprobados por periodo, promedio general y promedio en matemáticas (Alkhasawneh & Hargraves, 2014).

Y en el tercero, usaron un conjunto de datos con 60 atributos de 419 jóvenes de preparatoria y a partir del algoritmo selección de atributos obtuvieron los mejores atributos: promedio en la secundaria, grupo, número de estudiantes en el grupo, edad, asistencia, nivel educativo de la madre, distancia, consumo regular de alcohol, hábitos de fumar, sanciones administrativas, lugar utilizado para estudiar, nivel de motivación, calificación en matemáticas, ciencias sociales y humanidades (Márquez-Vera, et al., 2016). Cabe destacar que tanto en este artículo como en los presentados por Kotsiantis, Pierrakeas, & Pintelas (2003) y Barbosa Manhães, Serra da Cruz, & Zimbrão (2014), además de utilizar árbol de decisión, también se utiliza máquinas de soporte vectorial para robustecer sus estudios (estos no utilizan selección de atributos).

Barbosa Manhães, Serra da Cruz, & Zimbrão (2014) destacan la alta tasa de deserción que ocurren en las universidades públicas de Brasil y cómo esta afecta al gobierno, a la sociedad, a las empresas y a los propios estudiantes, pues se destinan recursos no recuperables y se presenta la falta de recursos con las habilidades necesarias para el mercado laboral. De ahí el desarrollar este experimento en la universidad pública más grande de Brasil, la Universidad Federal de Río de Janeiro, el cual cuenta con 14,000 instancias con información académica de los estudiantes extraída del sistema de gestión académico de la universidad, dejando fuera aquellas variables no académicas como lo pueden ser las demográficas, financieras, sociales; además de contar con tres clases que son: alumno desertor, actualmente registrado, graduado; estas variables pasan previamente por un proceso de transformación y categorización. En cuanto a la aplicación de algoritmos de minería de datos educativa, el árbol de decisión tuvo la segunda precisión más baja del experimento con 82.77% solo superando a *Näive Bayes* de 79.59%, mientras que SVM fue el algoritmo que mejor clasificó con 87.39% de precisión. Cabe destacar que para los autores de este experimento es muy relevante el tiempo de ejecución, dando como resultado que el árbol de decisión tardara 0.89 segundos en ejecutar, solo por encima de *Näive Bayes* con tan solo 0.13 segundos, mientras que SVM es el segundo algoritmo que más tarda con 15.85 segundos detrás únicamente de redes neuronales con 4,607.21 segundos.

Kotsiantis, Pierrakeas & Pintelas (2003) desarrollan un estudio que aborda la problemática de altos índices de deserción en los programas en línea de la universidad Abierta Helénica, por lo que se busca identificar patrones que les permitan identificar de forma temprana si un estudiante es propenso a desertar, para lo cual se aplican seis diferentes algoritmos de aprendizaje automático supervisado para clasificar los casos específicamente en un módulo del curso de Informática. El conjunto de datos considera variables demográficas, familiares, académicas y laborales. La prueba se divide en 5 experimentos donde al principio se entrenan los algoritmos únicamente con la información demográfica, posteriormente se incluye también la asistencia a su primera asesoría, luego la calificación de la primera actividad escrita, la segunda asesoría y por último la segunda actividad escrita. El resultado que arroja el experimento es que el árbol de decisión clasifica con una precisión desde 63.02% hasta 83.75% y las máquinas de soporte vectorial desde 58.09% hasta 82.96%.



Finalmente, con estos hallazgos, se puede determinar que el algoritmo árboles de decisión es el más utilizado para resolver situaciones en el contexto educativo cuyo porcentaje mayor de predicción reportado es de 83.75%; que el algoritmo con mejor desempeño al predecir algún evento educativo es máquinas de soporte vectorial con un porcentaje del 98.1% y el algoritmo con el menor desempeño para predecir una situación educativa es regresión lineal con un porcentaje del 66.59%. Estos aciertos, permiten en este trabajo, una oportunidad para identificar aquellos factores más importantes que intervienen en la deserción escolar universitaria a partir de un conjunto de datos de 56 atributos usando selección de atributos, y también, se vislumbra un reto de obtener un mejor desempeño en la aplicación del algoritmo árboles de decisión.

Capítulo IV: Metodología

En este capítulo se describe el proceso seguido para la realización de los objetivos de esta investigación. Partiendo desde el diseño de la investigación hasta culminar con la descripción del proceso para el descubrimiento del conocimiento que conlleva a la aplicación de las técnicas de minería de datos educativa mencionadas en capítulos anteriores.

4.1 Diseño de la Investigación

Esta investigación toma como base el Proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD, *Knowledge Discovery in Databases*) para la aplicación de técnicas de minería de datos educativa a un conjunto de datos de construcción propia. De este modo se aplican los algoritmos selección de atributos y árboles de decisión, para identificar los principales factores que influyen en la deserción universitaria, a la vez, que se encuentran patrones para prevenirla, usando 500 registros obtenidos a través de una encuesta administrada a estudiantes inscritos en Instituciones de Educación Superior (IES) públicas y privadas en el Estado de Puebla, México. La metodología utilizada se muestra a continuación en la figura 3.



Figura 3. Metodología para determinar las características más relevantes en la deserción universitaria.

Fuente: Elaboración propia.

4.2 Selección de la Población y Muestra

La población se determina considerando 230,788 estudiantes de educación superior del Estado de Puebla, México de acuerdo con datos del Sistema Nacional de Información Estadística Educativa [SNIE] (2019). El cálculo de la muestra fue realizado con un nivel de confianza del 97.5% y un error de muestreo admisible del 5%. Se ha utilizado la Fórmula 1 de muestreo aleatorio simple por proporciones de poblaciones finitas propuesta en Morillas (2014).

$$n = \frac{N Z_{1-\alpha/2}^2 p q}{(N-1)\varepsilon^2 + Z_{1-\alpha/2}^2 p q} \quad \text{Fórmula 1}$$

Donde: n = tamaño de la muestra; N = Tamaño de la Población; Z = Nivel de confianza; p = Probabilidad de éxito o proporción esperada, 0.5 cuando se desconoce el valor; q = Probabilidad de fracaso; ε = Error

admisible. Por tanto, el valor de la muestra se observa en la Ecuación 1, de manera que se ha truncado a 500 encuestados.

$$n = \frac{(230,788)(2.24)^2(0.5)(0.5)}{(230,788-1)(0.05)^2 + (2.24)^2(0.5)(0.5)} = 500.7 \text{ Ecuación 1}$$

El instrumento se aplicó en diversas universidades públicas y privadas del estado de Puebla para solicitar el llenado en línea del instrumento. Obteniendo una población de 300 estudiantes de IES públicas y 200 estudiantes de IES privadas. De los cuales, 311 son mujeres (176 de IES públicas y 135 IES privadas) y 188 son hombres (124 de IES públicas y 65 de IES privadas).

4.3 Recolección de Datos

4.3.1 Selección del Instrumento

El instrumento utilizado para la recolección de datos fue adaptado de la <Encuesta internacional sobre abandono en la educación superior> del proyecto Alfa Guía, 2014 (Valle Gómez, Eslava Gómez, Manzano Patiño, & García Minjares, 2014). En la Figura 4 se presenta la descripción de dicha adaptación.

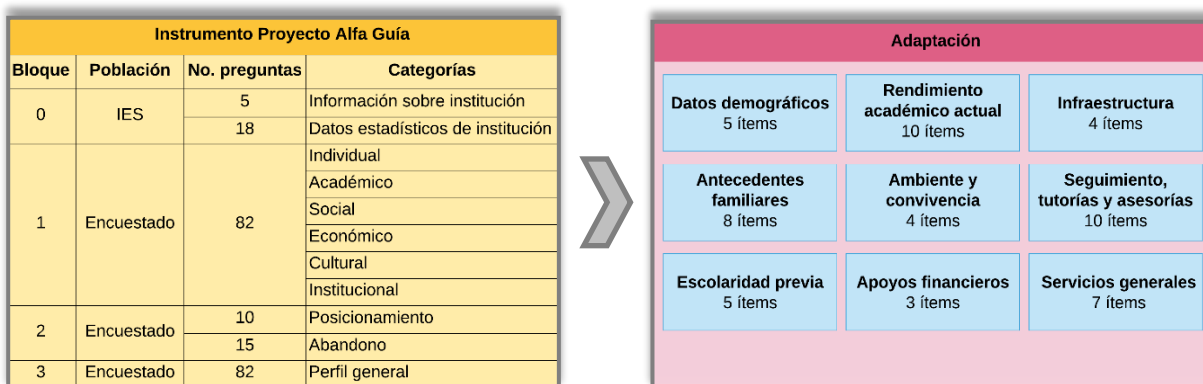


Figura 4. Adaptación del instrumento para la recolección de datos.

Fuente: Urbina-Nájera, et al. (2019)

Como se observa en la Figura 4, el instrumento adaptado tiene 56 preguntas divididas en 9 categorías, mismas que se describen a continuación:

- Datos demográficos: Recolecta datos referentes a la edad, género, estado civil, número de hijos y empleo
- Antecedentes familiares: Busca identificar el nivel de estudios del padre/madre/tutor y del hermano(a) mayor, así como también la dependencia económica
- Escolaridad previa: Información sobre la preparatoria o bachillerato
- Rendimiento académico actual: Tipo de institución, área, promedio, entre otras
- Apoyos financieros: Tipos de apoyos financieros y dependencia que los otorga
- Ambiente y convivencia: Recoge información relacionada al ambiente en la institución

- **Infraestructura:** Estima el grado de satisfacción de los espacios universitarios
- **Seguimiento, tutorías y asesorías:** Identifica el grado de satisfacción del seguimiento académico proporcionado por la parte académica y administrativa de la institución
- **Servicios:** Mide el grado de satisfacción de los servicios generales prestados por la institución en atención al estudiante

4.3.2 Validez del Instrumento

A pesar de que el instrumento fue adaptado, se consideró validar la fiabilidad de este mediante el coeficiente de Cronbach que es una función estadística que evalúa el grado en que un conjunto de variables encaja entre sí. Es aplicado típicamente para determinar si un grupo de variables puede combinarse válidamente en un solo índice o escala. Oscila entre 0 y 1, con valores cercanos a 1 que indican que los elementos están estrechamente relacionados y forman un índice efectivo (Longest, 2019). De este modo, el instrumento tuvo un coeficiente de 0.8767 y cada ítem obtuvo un coeficiente entre 0.8 y 0.9, reflejando con esto una fiabilidad relevante.

4.3.3 Aplicación del Instrumento

Se ha diseñado un formulario electrónico en *Google Forms* (Google, 2019), para la administración del instrumento de forma tal que sea más sencillo recolectar la información y facilitar el proceso de descubrimiento del conocimiento (véase Tabla 6).

Sección	Tipo de ítems	No. ítems	Sección	Tipo de ítems	No. ítems
Datos demográficos	Opción múltiple	4	Apoyos financieros	Escala Likert	3
	Dicotómica	1			
Antecedentes familiares	Opción múltiple	8	Infraestructura	Escala Likert	4
Escolaridad previa	Opción múltiple	5	Seguimiento, tutorías y asesorías	Escala Likert	10
Desempeño académico actual	Dicotómicas	2	Servicios	Escala Likert	7
	Abierta	1			
	Opción múltiple	7	Ambiente y convivencia	Opción múltiple	4

Tabla 6. Características del instrumento diseñado en *Google Forms*. Disponible en: <https://forms.gle/mR9tWmKZsTixqN6m8>

Fuente: Urbina-Nájera, et al. (2019)

4.4 Preparación de Datos

Una vez aplicado el instrumento de 56 preguntas a los 500 estudiantes de Instituciones de Educación Superior, es necesario pre-procesar las respuestas con el fin de obtener un conjunto de datos legible para la herramienta de software seleccionada para su procesamiento obteniendo con esto, 56 atributos y 500 instancias.

El proceso de codificación se realizó agrupando las posibles respuestas de los encuestados. Por ejemplo, en la columna A de la Figura 4, la edad se ha codificado por rangos: 1)18-19, 2)20-21, 3)22-23, 4) mayor

de 23, el mismo criterio fue aplicado para el resto de variables del conjunto de datos (véase columnas e, f,g,h,i,j...etc. de la Figura 4. Mientras que la fase de limpieza ha consistido en verificar que no existan datos vacíos o inconsistencias, si las hubiere, tomar la decisión de llenar estos datos con valores como: NA o cero. Es poco probable que existan inconsistencias en los datos puesto que el formulario está codificado y no acepta valores que no se hayan contemplado como respuesta. Dado que la herramienta de software utilizada para la aplicación de algoritmos de aprendizaje computacional lee archivos con extensión .csv, no se ha realizado el proceso de transformación pues en la hoja de cálculo se ha podido guardar el conjunto de datos con dicha extensión.

Finalmente, la carga de los datos se realiza en *Weka* que Frank, et al, (2016) definen como “una colección de algoritmos de aprendizaje automático para realizar tareas de minería de datos. Contiene herramientas para los datos de pre-procesamiento, clasificación, regresión, conglomerados, reglas de asociación y visualización. También es adecuado para el desarrollo de nuevos sistemas de aprendizaje de máquina” y ha sido seleccionada por que es la herramienta más popular en el área de aprendizaje computacional a pesar de que su interfaz de usuario es poco gráfica.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	2_Genero	3_Estado_civil	4_Tiene_Hijos	11_Lugar_de_nacimiento	13-De_quien_depende_economicamente	15_Institucion_procedente	18_tipo_Institucional	21_recibio_orientacion	22-Lugar_que_ocupa_el_carrera	23-Razon_eleccion_carrera	25_horario_en_el_que_estudio	27_El_horario_en_que_se_imparte_clases	28_promedio_no_desea_n_impactar	29_situacion_financiera	30_tiene_apoyos_financieros	32_No_tiene_apoyos_financieros	36-Falta_infraestructura	45-Falta_seguimiento	46-Falta_asesorias
1	1	2	1	No_aplica	4	2	2	2	3	2	2	5	2	3	1	1	2	2	1
2	1	2	1	No_aplica	1	1	1	1	1	6	4	3	5	3	5	1	5	3	3
3	1	2	1	No_aplica	1	1	1	1	1	1	4	3	4	2	5	3	4	3	4
4	1	2	1	No_aplica	1	1	1	1	1	2	1	5	2	1	5	1	1	2	2
5	1	2	1	No_aplica	2	1	1	2	1	1	4	5	5	5	1	3	3	5	5
6	1	2	1	No_aplica	1	2	1	2	2	3	1	4	4	3	5	3	4	4	4
7	1	2	1	No_aplica	1	2	1	1	1	2	1	5	3	5	1	5	5	5	5
8	1	2	1	No_aplica	1	2	1	1	1	2	1	5	5	2	5	2	1	4	4
9	1	2	1	No_aplica	1	1	1	1	1	1	3	5	5	5	5	4	5	5	5
10	1	2	1	No_aplica	1	2	1	1	2	7	1	4	1	3	5	2	5	2	2
11	1	2	1	No_aplica	1	1	2	1	3	2	1	5	5	2	1	1	3	2	2
12	1	2	1	No_aplica	1	2	2	2	1	1	4	5	5	3	1	4	4	4	4
13	1	2	1	No_aplica	4	1	1	2	1	6	4	1	1	4	5	2	5	5	5
14	1	2	1	No_aplica	1	1	2	1	2	4	2	2	2	2	5	5	2	2	2
15	1	2	1	No_aplica	4	1	1	2	1	2	2	1	1	1	5	1	2	2	4
16	1	2	1	No_aplica	1	1	1	2	1	2	1	3	4	3	5	2	3	3	3
17	1	2	1	No_aplica	1	1	1	1	2	1	4	2	4	4	5	2	4	3	3
18	1	2	1	No_aplica	1	1	1	2	1	6	1	1	3	3	5	3	3	3	3
19	1	2	1	No_aplica	1	1	2	2	2	1	1	5	5	5	1	2	4	5	5
20	1	2	1	No_aplica	1	1	1	1	1	1	1	3	3	1	5	1	2	3	3
21	1	2	1	No_aplica	4	1	1	2	2	7	2	2	1	1	1	1	5	2	2
22	1	2	1	No_aplica	1	1	1	2	1	2	1	2	3	4	5	1	5	5	5
23	1	2	1	No_aplica	1	1	1	1	1	2	4	2	4	1	1	1	3	3	3
24	1	2	1	No_aplica	1	1	1	1	1	2	4	2	4	1	1	1	3	3	3

Figura 5. Conjunto de datos pre-procesada.

Fuente: Urbina-Nájera, et al. (2019)

4.5 Aplicación de técnicas de minería de datos educativa para análisis de datos

En esta etapa de la metodología se aplica el algoritmo de selección de atributos para identificar los atributos más importantes que determinan una deserción universitaria inminente, posteriormente se aplica el algoritmo de árbol de decisión al conjunto de datos con los 56 atributos y posteriormente se aplica únicamente a los atributos relevantes identificados, con estas pruebas se comparan los resultados determinando cual clasifica con mejor desempeño.

Para efectuar la fase de selección de atributos como parte del proceso KDD, en este trabajo se utilizan dos evaluadores de subconjuntos de atributos: *GainRatioAttributeEval* e *InfoGainAttributeEval*.

- *GainRatioAttributeEval*: Evalúa el valor de un atributo midiendo la relación de ganancia con respecto a la clase. Según la siguiente ecuación (Hall, 2011).

$$GananciaR(Clase, Atributo) = \frac{(H(clase) - H(Clase|Atributo))}{H(Atributo)}$$

- *InfoGainAttributeEval*: Evalúa el valor de un atributo midiendo la ganancia de información con respecto a la clase. Según la siguiente ecuación (Hall, 2011).

$$InfoGain(Clase, Atributo) = H(Clase) - H(Clase|Atributo)$$

Capítulo V: Resultados

En esta sección se presenta la aplicación del proceso para el descubrimiento de conocimiento en tres fases: 1) Aplicación de los evaluadores de selección de atributos, 2) Reconocimiento de reglas con árboles de decisión y 3) Clasificación con máquinas de soporte vectorial.

5.1 Selección de Atributos

Como se ha descrito con anterioridad, una de las fases del proceso para el descubrimiento del conocimiento es la selección de atributos. En la Figura 5 se presentan los resultados de haber aplicado los dos evaluadores (*GainRatioAttributeEval* e *InfoGainAttributeEval*), en la cual se denota una lista de los 27 principales atributos del conjunto de datos que contiene 56 atributos. Es notorio que el ranking de ambos evaluadores considera como importantes a los mismos atributos a pesar de otorgarles diferentes pesos. Dado que los resultados de ambos evaluadores coinciden, serán considerados para tipificar las reglas que pueden favorecer la identificación de una temprana deserción a partir de los árboles de decisión.

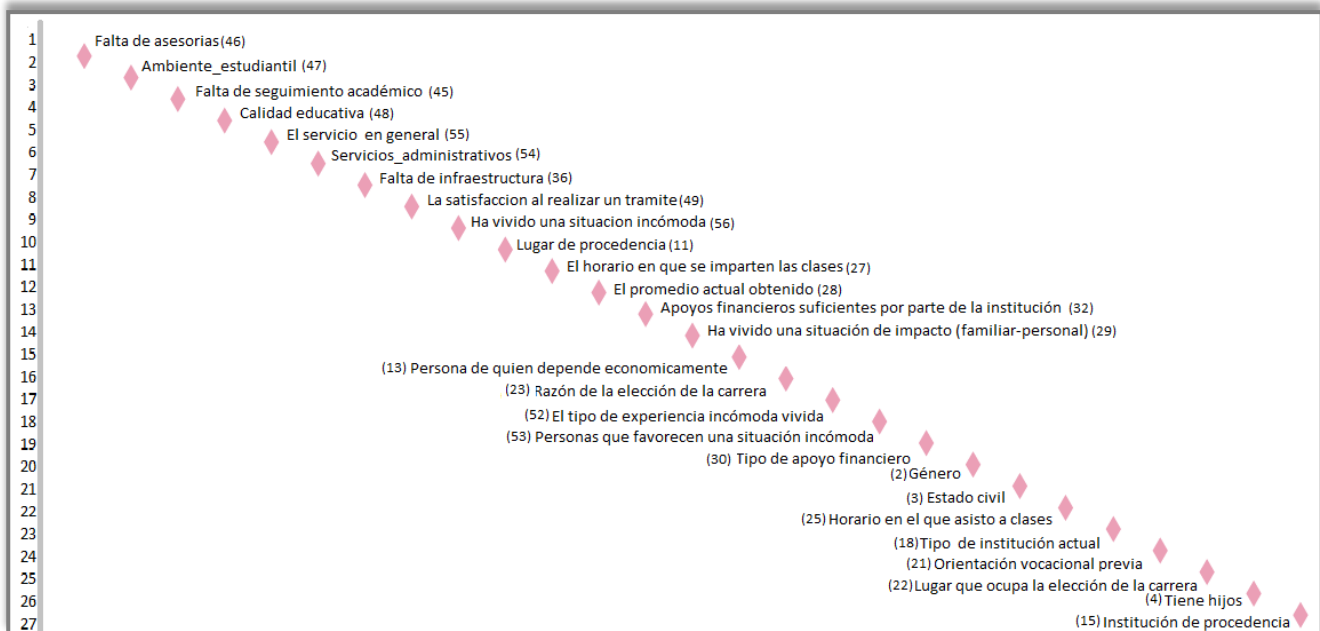


Figura 6. Atributos más importantes en la deserción escolar universitaria usando selección de atributos.

Fuente: Urbina-Nájera, et al. (2019)

5.2 Árboles de Decisión

En esta fase se presentan los hallazgos encontrados una vez que se ha aplicado el algoritmo árboles de decisión en sus dos fases; en la primera fase se ha utilizado el conjunto de datos completo, es decir, con los 56 atributos obteniendo un desempeño del 74.6% de exactitud (Tabla 8). En la segunda fase, se ha modificado el conjunto de datos basados en la aplicación de la selección de atributos descrita en la sección anterior. En otras palabras, se ha aplicado nuevamente el algoritmo árboles de decisión al

conjunto de datos solamente con los 27 atributos más relevantes de acuerdo con el algoritmo selección de atributos, obteniendo un desempeño de 92.6% de exactitud.

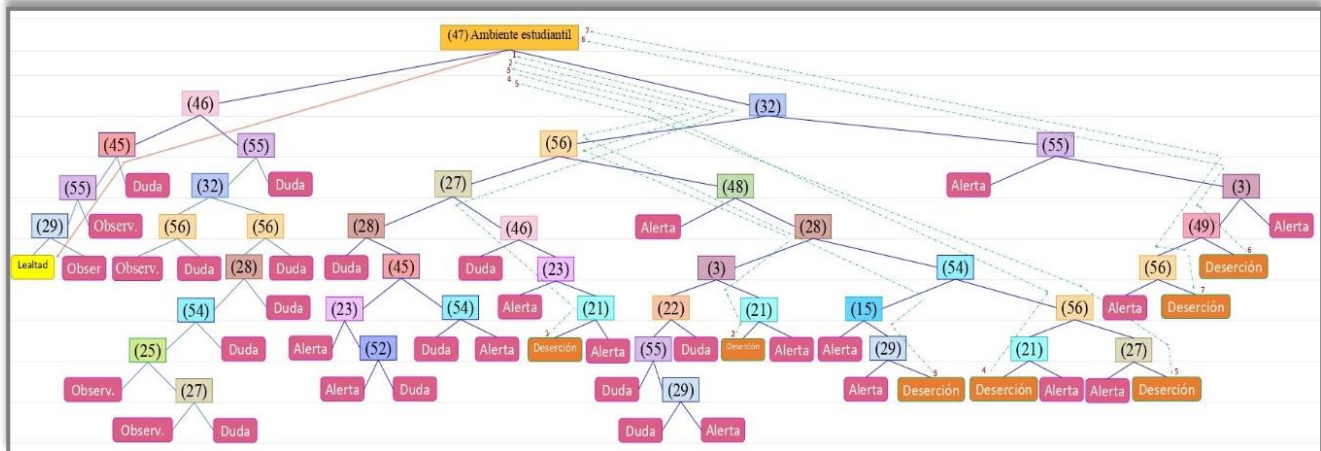


Figura 7. Árbol de decisión con la mejor clasificación y causas que predicen una inminente deserción universitaria.

Fuente: Urbina-Nájera, et al. (2019)

En la Figura 7 se muestra el árbol obtenido con los mejores atributos y las reglas que definen un patrón para aquellos jóvenes que tienen lealtad y convicción de permanecer estudiando. Dicho patrón está relacionado con la regla siguiente:

```

47_AMBIENTE ESTUDIANTIL <= 3
| 46_FALTA DE ASESORÍAS <= 2
| | 45_FALTA DE SEGUIMIENTO ACADÉMICO <= 2
| | | 55_SERVICIO GENERAL <= 1
| | | | 29_HA VIVIDO UNA SITUACIÓN DE IMPACTO <= 1: LEALTAD
  
```

La cual se lee de la siguiente manera: SI el ambiente estudiantil es neutral a satisfactorio y la falta de asesoría y la falta de seguimiento académico casi no existe y el servicio en general es bueno y no ha vivido una situación de impacto (muerte de un familiar, divorcio de padres, ruptura sentimental) ENTONCES permanezco estudiando. En otras palabras, un estudiante universitario decide continuar con sus estudios si existe un ambiente estudiantil bueno a excelente, si las asesorías, seguimiento académico y servicios en general son buenos a excelentes y no ha vivido una situación de impacto. Por el contrario, para identificar una inminente deserción se han encontrado siete reglas (Figura 7). La regla 1 y la regla 2, por citar ejemplos, son descritas a continuación:

Regla 1

```

47_AMBIENTE ESTUDIANTIL > 3
| 32_APOYOS FINANCIEROS INSUFICIENTES <= 4
| | 56_HA VIVIDO UNA SITUACIÓN INCÓMODA <= 3
| | | 27_HORARIO EN EL QUE SE IMPARTEN LAS CLASES > 4
| | | | 46_FALTA DE ASESORÍAS > 2
| | | | | 23_RAZÓN DE LA ELECCIÓN DE LA CARRERA > 2
| | | | | | 21_RECIBIO ORIENTACIÓN PREVIA <= 1: DESERCIÓN
  
```

Regla 2

```

47_AMBIENTE ESTUDIANTIL > 3
| 32_APOYOS FINANCIEROS INSTITUCIONALES <= 4
| | 56_HA VIVIDO UNA SITUACIÓN INCÓMODA > 3
| | | 48_SERVICIO GENERAL > 2
| | | | 28_PROMEDIO ACTUAL > 4
| | | | | 54_SERVICIOS ADMINISTRATIVOS <= 4
| | | | | | 15_INSTITUCIÓN DE PROCEDENCIA = PÚBLICA
| | | | | | | 29_HA VIVIDO UNA SITUACIÓN DE IMPACTO > 2: DESERCIÓN
  
```

La regla 2 se puede explicar de la siguiente forma: SI el ambiente estudiantil va de bueno a pésimo y los apoyos financieros proporcionados por la institución son insuficientes, ha experimentado una situación incómoda (acoso, discriminación, mal trato), el servicio general, el promedio general no es el esperado y los servicios administrativos son buenos a excelentes y procede de una preparatoria pública y ha vivido una situación de impacto ENTONCES decide desertar.

Para conocer el desempeño del algoritmo árboles de decisión, clasificando a la clase “pronóstico” identificada por 1: Lealtad, 2: Observancia, 3: Duda, 4: Alerta y 5: Deserción, se encuentran las dos matrices de clasificación obtenidas una vez que se ha probado con el conjunto de datos a) 56 atributos y b) 27 atributos más relevantes (Tabla 7).

a) Clasificación con 56 atributos					
a	b	c	d	e	Clasificación
65	0	24	0	4	a=Observancia
3	102	29	13	0	b=Alerta
21	30	115	3	0	c=Duda
1	19	0	42	0	d=Deserción
3	0	0	0	4	e=Lealtad

b) Clasificación con 27 atributos					
a	b	c	d	e	Clasificación
88	0	4	0	1	a=Observancia
1	135	10	1	0	b=Alerta
4	6	179	0	0	c=Duda
0	8	1	53	0	d=Deserción
1	0	0	0	8	e=Lealtad

Tabla 7. Matriz de confusión obtenida con a) 56 atributos y b) 27 atributos más relevantes.

Fuente: Urbina-Nájera, et al. (2019)

En la Tabla 7 se han seleccionado los valores en la diagonal de cada una de las matrices, la cual indica las instancias correctamente clasificadas en función de la clase pronóstico. Por ejemplo, si se desea hacer un comparativo entre las matrices relacionada al atributo <Deserción>, se observa que en la matriz a 42 instancias fueron correctamente clasificadas como deserción, mientras que 1 fue incorrectamente clasificada como observancia y 19 instancias como alerta; en comparación con la matriz b donde 53 instancias fueron clasificadas correctamente como deserción, 8 fueron clasificadas como alerta y 1 como duda. De esta manera, en la matriz b, el número de instancias correctamente clasificadas fue mayor que en la matriz a, situación que se nota para el resto de los atributos. Del mismo modo, en la Tabla 8 se presentan las métricas de desempeño más importantes del algoritmo árbol de decisión mencionadas con anterioridad.

Clasificación	Exactitud	Precisión	Recuperación	Medida-F	Instancias clasificadas correctamente
56 atributos	74.6	74.9	74.6	74.6	373
27 atributos	92.6	92.7	92.6	92.6	463

Tabla 8. Métricas de desempeño del árbol de decisión aplicado en dos fases.

Fuente: Urbina-Nájera, et al. (2019)

Se observa que utilizando el algoritmo selección de atributos la exactitud y precisión del algoritmo árbol de decisión aumenta considerablemente; aunque lo más importante es conocer cuáles son los factores más relevantes que se deben considerar como alerta en la decisión de desertar, así como aquellas reglas o patrones que un estudiante universitario sigue para desertar de sus estudios universitarios.

5.3 Máquina de Soporte Vectorial

Al igual que se realizó con el árbol de decisión, en esta fase se presentan los hallazgos encontrados una vez que se ha aplicado el algoritmo máquinas de soporte vectorial en dos etapas; en la primera se ha utilizado el conjunto de datos completo, es decir, con los 56 atributos, aplicando variantes en los *kernel* a utilizar (lineal, binomial, trinomial y de base radial) obteniendo el mejor desempeño de 83.0% de exactitud con el *kernel* binomial (Tabla 9), cabe destacar que el desempeño del *kernel* de base radial obtuvo un desempeño bajo comparado con los experimentos realizados en este estudio (63.6%) y el *kernel* trinomial no pudo encontrar el hiperplano que separara satisfactoriamente la clase <Lealtad>, mientras que el *kernel* de base radial no fue capaz de separar las clases <Lealtad> ni <Deserción>.

En la segunda etapa, se ha modificado el conjunto de datos basados en la aplicación de la selección de atributos descrita anteriormente. En otras palabras, se ha aplicado nuevamente el algoritmo de máquinas de soporte vectorial con sus variantes en el *kernel* al conjunto de datos solamente con los 27 atributos más relevantes de acuerdo con el algoritmo selección de atributos, obteniendo el mejor desempeño con el *kernel* lineal de 86.6% de exactitud, se presenta nuevamente un bajo desempeño con el *kernel* de base radial alcanzando tan solo 64.8% de exactitud, además de no tener la capacidad de separar las clases <Lealtad> y <Deserción>.

Tal cual se presentó en la sección anterior, en el caso del algoritmo SVM se observa una clasificación mejor (mayor exactitud de clasificación) al conjunto de datos después de aplicar el algoritmo de selección de atributos comparado con el conjunto de datos completo como se puede ver en la tabla 9.

	<i>Kernel</i>	Exactitud	Precisión	Recuperación	Medida-F	Instancias clasificadas correctamente
56 Atributos	Lineal	82.4	82.2	82.4	82.1	412
	Binomial	83.0	83.3	83.0	82.6	415
	Trinomial	81.8	-	81.8	-	409
	Base Radial	63.6	-	63.6	-	318
27 Atributos	Lineal	86.6	86.6	86.6	86.5	433
	Binomial	86.0	86.0	86.0	85.9	430
	Trinomial	84.4	84.8	84.4	84.1	422
	Base Radial	64.8	-	64.8	-	324

Tabla 9. Métricas de desempeño de la máquina de soporte vectorial con variantes de *kernel* en dos etapas.

Fuente: Elaboración propia

En la tabla 10 se muestra la matriz de confusión de la variante de *kernel* con el mejor desempeño de clasificación en cada una de las etapas del experimento, resaltando la diagonal de instancias correctamente clasificadas. Se puede observar que, en ambas etapas, las instancias de la clase “Lealtad” incorrectamente clasificadas son confundidas con <Observancia>, y aquellas de <Deserción> son confundidas con <Alerta>.

a) Clasificación con 56 atributos y <i>kernel</i> binomial					
a	b	c	d	e	Clasificación
75	0	18	0	0	a=Observancia
0	117	21	9	0	b=Alerta
10	12	167	0	0	c=Duda
0	8	0	54	0	d=Deserción
7	0	0	0	2	e=Lealtad

b) Clasificación con 27 atributos y <i>kernel</i> lineal					
a	b	c	d	e	Clasificación
78	0	14	0	1	a=Observancia
0	125	13	9	0	b=Alerta
12	6	171	0	0	c=Duda
0	8	0	54	0	d=Deserción
4	0	0	0	5	e=Lealtad

Tabla 10. Matriz de confusión obtenida con a)56 atributos y kernel binomial y b)27 atributos más relevantes y kernel lineal

Fuente: Elaboración propia

Capítulo VI: Análisis y Discusión

Durante este estudio se ha presentado la manera de cómo aplicar la minería de datos educativa a un conjunto de datos e identificar hallazgos y patrones que coadyuven a desarrollar planes de acción y toma de decisiones en este sentido.

En este caso en particular se adaptó un cuestionario utilizando como base el proyecto Alfa Guía 2014 de acuerdo con las particularidades de México, considerando 56 atributos relacionados con datos demográficos, antecedentes familiares, escolaridad previa, rendimiento académico actual, apoyos financieros, ambiente y convivencia, infraestructura, seguimiento, tutorías y asesorías, servicios de las instituciones. Sivakumar, Venkataraman & Selvaraj (2016), Márquez-Vera et al. (2016) y Barbosa Manhães, Serra de Cruz & Zimbrão (2014) también incluyen algunos de estos rubros en sus estudios. En cambio, Abu-Oda & El-Halees (2015), Dekker, Pechenizkiy & Vleeshouwers (2009) y Kotsiantis, Pierrakeas, & Pintelas (2003) dan más peso al desempeño académico actual en los cursos.

Para recabar la información se utilizó un formulario electrónico. En este sentido Yukselturk, Ozeke & Türel (2014) también obtuvieron sus datos por un medio electrónico a través de la adaptación de diversos cuestionarios al contexto de Turquía. En ambos casos se validó el instrumento valiéndose del coeficiente de Cronbach obteniendo en su caso valores entre 0.7 y 0.9, en el presente estudio entre 0.8 y 0.9, reflejando con esto una fiabilidad relevante en ambos casos.

Una vez obtenidos los datos se procedió al pre-procesamiento de los mismos, realizando transformaciones y codificación para su procesamiento y análisis en Weka, software comúnmente utilizado en este tipo de estudios. Abu-Oda & El-Halees (2015) describen su proceso de pre-procesamiento donde utilizaron técnicas para discretizar la información y para evitar el efecto de sobre muestreo (Over sampling en inglés); en el caso de Manhães, Serra de Cruz & Zimbrão (2014) y Kotsiantis, Pierrakeas, & Pintelas (2003) también pre-procesan su set de datos a través de categorizar los atributos.

El presente estudio propone utilizar Selección de Atributos para identificar a aquellos más relevantes para la deserción universitaria previo a la aplicación del algoritmo de aprendizaje computacional, en este caso Árbol de Decisión. En este sentido se realizaron dos pruebas, el primero aplicando el árbol de decisión a los 56 atributos y el segundo aplicándolo únicamente a los 27 identificados con selección de atributos. Enfoques similares se proponen en la literatura, por ejemplo Sivakumar, Venkataraman & Selvaraj (2016) utilizan selección de atributos además de aplicar entropía de *Renyi*, ganancia de información y función de asociación al árbol de decisión; Márquez-Vera, et al. (2016) aplican 3 experimentos, el primero utilizando cinco algoritmos diferentes de clasificación entre ellos árbol de decisión, el segundo aplicando previamente selección de atributos y el tercero aplicando algoritmos diseñados para datos desbalanceados; Yukselturk, Ozeke & Türel (2014) utilizaron un algoritmo genético para selección de atributos y posteriormente aplicaron algoritmos de clasificación entre ellos árbol de decisión; por último, Dekker, Pechenizki & Vleeshouwers (2009) en sus tres experimentos aplicaron algoritmos de clasificación con todos los atributos y posteriormente realizaron la prueba utilizando selección de atributos.

En la Tabla 11 se muestra la exactitud de clasificación de árbol de decisión y máquinas de soporte vectorial obtenido en la literatura revisada:

Literatura	Árbol de Decisión Todos los Atributos	Árbol de Decisión Selección de Atributos	SVM Todos los Atributos	SVM Selección de Atributos
Sivakumar, Venkataraman & Selvaraj (2016)	92.5%	97.5%	-	-
Márquez-Vera et al. (2016) <i>El rango depende del número de atributos por etapa</i>	90.7% - 95.9%	90.7% - 96.4%	85.7% - 98.1%	86.0% - 96.4%
Abu-Oda & El-Halees (2015)	98.1%	-	-	-
Yukselturk, Ozekes & Türel (2014)	-	79.7%	-	-
Barbosa Manhães, Serra da Cruz & Zimbrão (2014)	82.8%	-	87.4%	-
Dekker, Pechenizki & Vleeshouwers (2009) <i>Experimento 1</i>	69%	71%	-	-
Dekker, Pechenizki & Vleeshouwers (2009) <i>Experimento 2</i>	79%	-	-	-
Dekker, Pechenizki & Vleeshouwers (2009) <i>Experimento 3</i>	80%	-	-	-
Kotsiantis, Pierrakeas, & Pintelas (2003) <i>El rango depende del número de atributos por etapa</i>	63.0% - 83.8%	-	58.1% - 83.0%	-
Presente Estudio	74.6%	92.6%	83.0%	86.6%

Tabla 11. Exactitud en la clasificación de Árboles de Decisión y SVM en la literatura comparado con el presente estudio.
 Fuente: Elaboración propia con información de Urbina-Nájera, et al. (2019)

Se puede observar que salvo en la aplicación de SVM de Márquez-Vera et al. (2016), en el resto de los casos donde se aplicó previamente selección de atributos al algoritmo de minería de datos educativa se obtiene mayor exactitud que cuando no fue aplicado. También es posible ver que este estudio alcanza una exactitud de clasificación muy competitiva comparada con los mejores resultados de la literatura revisada.

También es de destacar que el tiempo de ejecución de SVM es mayor al de árbol de decisión, esto también se resalta en el estudio de Barbosa Manhães, Serra da Cruz, & Zimbrão (2014), donde se descartó a SVM por este factor y por la dificultad de interpretación de los resultados para el usuario, los autores finalmente seleccionaron el algoritmo de *Näive Bayes*, el cual tuvo el peor desempeño de clasificación y un buen tiempo de ejecución, sin embargo, considero que las conclusiones a las que llegan los autores se fundamentan en un análisis descriptivo de las variables a través de gráficas de barra y línea, comparando el comportamiento de las clases variable por variable sin aprovechar el poder de la minería de datos educativa, el cual da la posibilidad de hacer un comparativo de las clases considerando todas las variables en su conjunto, es decir, si los autores buscan un algoritmo fácil de interpretar y con buena precisión de clasificación, la opción a utilizar es árbol de decisión, pues este genera reglas claras y su precisión es cercana a los mejores algoritmos de este experimento.

En otro orden de ideas, este estudio identifica que las cinco principales causas de la deserción escolar universitaria se deben a la falta de asesorías, a un inadecuado ambiente estudiantil, a la falta de seguimiento académico, a la deficiente calidad educativa y al servicio en general. Estos hallazgos contrastan sensiblemente con los encontrados en la literatura, en donde Márquez-Vera et al. (2016) indica que los principales factores son baja calificación, factores familiares, factores económicos (trabajar más

de cuatro horas), malos hábitos (alcohol y cigarro) y grupos grandes (más de 40 estudiantes); Sivakumar, Venkataraman & Selvaraj (2016) identificaron como factores principales razones familiares, ambiente universitario, disgusto por la carrera, extrañar a la familia, baja tasa de colocación; coincidiendo en el ambiente estudiantil únicamente. También, mediante la utilización de los evaluadores para la selección de atributos se pudo identificar que los aspectos relacionados con el disgusto por la carrera se encuentran en el lugar 13 y 19 de los 27 principales factores encontrados, así también, que los factores relacionados al aspecto económico ocupan los lugares 17 y 25 del ranking obtenido a través del uso del algoritmo. Esto indica que los jóvenes universitarios han ampliado la gama de factores que satisfacen el deseo de permanecer estudiando, al menos para la muestra obtenida, y que discrepan de los estudios analizados en la Tabla 2. Por otro lado, Yukselturk, Ozekes & Türel (2014) identifican factores específicos para los cursos en línea como son las percepciones iniciales de los participantes sobre las creencias de autoeficacia específicas del entorno en línea, la preparación para el aprendizaje en línea y el conocimiento previo sobre el contenido del curso, lo que parece ser un atributo común para los cursos en línea, pues Kotsiantis, Pierrakeas, & Pintelas (2003) también lo incluye en sus atributos, ellos además de factores demográficos, familiares y académicos, también incluyen el conocimiento previo de informática y el uso de este en el trabajo actual. En contraste los autores Dekker, Pechenizki & Vleeshouwers (2009), Barbosa Manhães, Serra da Cruz, & Zimbrão (2014) y Abu-Oda & El-Halees (2015) se enfocan por completo en el desempeño académico de los estudiantes con diferentes variables en este sentido.

En todos los casos se llega a la misma conclusión, la deserción escolar no depende de un solo factor o causa, sino que se deriva de circunstancias multifactoriales interrelacionadas entre sí, además de que se atribuye en la mayoría de los casos, al entorno y antecedentes del estudiante junto con su desempeño escolar y asesoramiento que recibe.

Conclusiones y Recomendaciones

México cuenta con una proporción de adultos (25-64 años) con título universitario de 17%, muy por debajo del promedio de los países miembro de la OCDE que es de 37%. Es por ello por lo que la deserción escolar toma gran relevancia como un problema de dimensiones considerables en México y específicamente en el estado de Puebla, el cual ocupa el sexto lugar nacional en este tema. De ahí que el propósito de este estudio es identificar los patrones que inciden en una deserción escolar universitaria inminente, para lo cual se aplicó el proceso de descubrimiento de conocimiento a un conjunto de 500 datos de estudiantes de IES en Puebla obtenidos mediante la aplicación de un instrumento adaptado del proyecto Alfa Guía 2014 sobre abandono en la educación superior, el cual consta de 56 preguntas clasificadas en 9 categorías para recabar información demográfica, antecedentes familiares, escolaridad previa, rendimiento académico actual, apoyos financieros, ambiente y convivencia, infraestructura, asesorías y servicios escolares.

A través de la aplicación de minería de datos educativa utilizando los algoritmos selección de atributos, árbol de decisión y máquinas de soporte vectorial, es posible concluir que las cinco principales causas de deserción escolar universitaria se deben a la falta de asesorías, a un inadecuado ambiente estudiantil, a la falta de seguimiento académico, a la deficiente calidad educativa y al servicio en general. En este sentido, se han logrado identificar siete patrones que conducen a una deserción escolar; de la misma manera, se han logrado clasificar patrones que son dignos de atención mostrados en la Figura 7 como aquellos en donde se encuentre duda, observancia o alerta. De la misma forma, se ha determinado el patrón que el estudiante universitario defiende para permanecer estudiando, el cual considera que exista un ambiente estudiantil adecuado, que le proporcionen asesorías y seguimiento académico, que los servicios en general sean adecuados y que no haya vivido una situación de impacto que le genere duda en la decisión de continuar con sus estudios.

Si bien se sabe que la mayoría de IES cuentan con mecanismos de tutoría y asesoría, pareciera que estos son genéricos para toda la población estudiantil, Los hallazgos obtenidos en este estudio dan la pauta para que las IES desarrollen programas y acciones personalizados y enfocados a alumnos que presenten las características identificadas como lo puede ser una situación de impacto en su vida, que podría denotar un factor de riesgo emocional del estudiante y que implique la toma de acciones como por ejemplo, atención psicopedagógica.

El gobierno mexicano se ha interesado principalmente por el factor de cobertura y calidad de la educación (infraestructura, servicios generales y acceso a la educación) y al factor financiero a través de otorgar becas y apoyos, en todos los casos tomando en consideración los factores sociales enfocados en los sectores de la población más marginados y rezagados; si bien son factores importantes, de acuerdo con este estudio no todos son los más relevantes o simplemente no están del todo en su control directo. Sin embargo, en los dos sexenios anteriores han puesto gran énfasis en apoyar a los estudiantes de nuevo ingreso en su proceso de incorporación al ambiente estudiantil, es aquí donde las IES a través de toda su comunidad universitaria y las familias deben involucrarse para contribuir en crear ese entorno donde se habiliten canales adecuados de comunicación, convivencia sana e integradora de todos los estudiantes. Un reto que ha tenido México y América Latina en general es habilitar mecanismos que permitan alinear

los programas educativos al mercado laboral, pues a pesar de contar con programas de calidad, la oferta de egresados no ha sido suficiente para satisfacer la demanda de habilidades y competencias específicas que requieren las empresas.

Uno de los hallazgos de esta revisión es que el gobierno mexicano desarrolla estrategias principalmente para la educación obligatoria o de manera general para todo el sistema educativo, direccionando estrategias de manera específica a la educación superior en pocas ocasiones. Y si bien se han implementado acciones desde el gobierno para combatir la deserción escolar, esta no ha sufrido una disminución significativa en los últimos años.

Finalmente, alineado a la revisión de la literatura sobre deserción escolar donde se analiza el fenómeno a través de métodos de investigación cualitativos y cuantitativos básicos, como en aquellos que utilizan algoritmos avanzados de minería de datos educativa, se ha confirmado que la deserción escolar no depende de un solo factor, sino que es causada por un conjunto de factores e interacción de estos, y que dichos factores variaran en función de la región, procedencia, nivel socioeconómico o incluso creencias. Lo que este estudio ha proporcionado con relación a otros estudios tradicionales es la introducción de minería de datos educativa para encontrar patrones que prevengan una inminente deserción y actuar antes que suceda. Por ello, para enriquecer este estudio es imperativo ampliar la muestra a otros estados o ciudades de manera que se puedan aplicar diversos algoritmos que proporcionen más información que conduzca al establecimiento de mecanismos certeros para disminuir los índices de deserción universitaria.

En este sentido, a través de minería de datos educativa se ha podido confirmar, alineado a la literatura revisada, que la aplicación del algoritmo de Selección de Atributos previo a la aplicación de algún algoritmo de clasificación como Máquina de Soporte Vectorial y Árbol de Decisión, se obtiene un mejor desempeño de clasificación que al aplicar el algoritmo de clasificación sin selección de atributos. Del mismo modo en la literatura revisada se ha podido observar que típicamente las máquinas de soporte vectorial clasifican con mejor desempeño que el árbol de decisión, tal como sucede en este estudio al aplicar dichos algoritmos al conjunto de datos completo (56 atributos), sin embargo, después de aplicar selección de atributos la situación se revierte (27 atributos más relevantes), clasificando significativamente mejor el árbol de decisión que la máquina de soporte vectorial (92.6% y 86.6% respectivamente). Además, un beneficio adicional que se obtiene del árbol de decisión es su facilidad de interpretación por su origen simbólico, pues como parte de los resultados se generan una serie de reglas que permiten a los tomadores de decisiones detonar acciones al respecto.

Referencias

- Abarca Rodriguez, A., & Sanchez Vindas, M. (2005). La deserción estudiantil en la educación superior: El caso de la Universidad de Costa Rica. *Actividades investigativas en educación*, 5, 1-22. Obtenido de <https://www.redalyc.org/html/447/44759911/>
- Abu-Oda, G. S., & El-Halees, A. M. (2015). Data mining in higher education: university student dropout case study. *International Journal of Data Mining & Knowledge Management Process*, 5(1), 15-27. doi:10.5121/ijdkp.2015.5102
- Alkhasawneh, R., & Hargraves, R. (2014). Developing a Hybrid Model to Predict Student First Year Retention in STEM Disciplines Using Machine Learning Techniques. *Journal of STEM Education*, 5(3), 35-42. Obtenido de <https://eric.ed.gov/?id=EJ1048907>
- Barbosa Manhães, L. M., Serra de Cruz, S. M., & Zimbrão, G. (2014). The Impact of High Dropout Rates in a Large Public Brazilian University: A Quantitative Approach Using Educational Data Mining. *CSEDU*, 124-129.
- Benitez, R. A. (2012). *Proyecto R.A.M.O.N.* Obtenido de [proyectoramon.wordpress.com](http://proyectoramon.wordpress.com/la-desercion-universitaria/): <https://proyectoramon.wordpress.com/la-desercion-universitaria/>
- Betancourt, G. A. (2005). Las Máquinas de Soporte Vectorial (SVMs). *Scientia et Technica*, 1(0122-1701). doi:DOI: <http://dx.doi.org/10.22517/23447214.6895>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. USA: O'Really Media, Inc.
- Bishop, C. M. (2007). *Pattern recognition and Machine Learning*. Singapore: Springer.
- Cabrera, L., Bethencourt, J., Alvarez, P., & González, M. (2006). El problema del abandono de los estudios universitarios. *RELIEVE*, 171-203. Obtenido de http://www.uv.es/RELIEVE/v12n2/RELIEVEv12n2_1.htm
- Camarena, R., Chávez, A., & Gómez, J. (1984). Eficiencia terminal en la UNAM 1970-1981. *Revista Perfiles Educativos*.(7).
- Carvajal Olaya , P., & Trejos Carpintero, A. (Febrero - Mayo de 2016). Revisión de estudios sobre deserción estudiantil en educación superior en latinoamérica bajo la perspectiva de Pierre Bourdieu. *VI CLABES. SEXTA CONFERENCIA LATINOAMERICANA SOBRE EL ABANDONO EN LA EDUCACION SUPERIOR*. (págs. 1-10). Perspectiva Económica de America Latina.
- Chatti, M., Dyckhoff, A., Schroeder, U., & Thüs, H. (2013). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 318-331. Obtenido de <https://pdfs.semanticscholar.org/63f9/71de987f6659cb9b6906ae3b8d21b030d69f.pdf>
- Cortés, S., Álvarez, P., Llanos, M., & Castillo., L. (Febrero - Mayo de 2019). Deserción universitaria: La epidemia que aqueja a los sistemas de educación superior. *REV. PERSPECTIVA*, 1, págs. 13-25. doi:ISSN 1996-5389. DOI: <https://doi.org/10.33198/rp.v20i1.00017>



- De Los Santos, V. J. (2004). Los procesos de permanencia y abandono escolar en educación superior. *Revista Iberoamericana de Educación*, 7, 1-7.
- de Vires, W., León Arenas, P., Romero Muñoz, J. F., & Hernández Saldaña, I. (Octubre-Diciembre de 2011). ¿Desertores o decepcionados? Distintas causas para abandonar los estudios universitarios. *REVISTA DE EDUCACIÓN SUPERIOR*, XL (4)(160), 29-49.
- Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *Educational Data Mining* (págs. 41-50). Cordoba, España: International working group on Educational Data Mining.
- Dominguez, D., Álvarez, J. F., & Gil-Jaurena, I. (2016). Analítica del aprendizaje y Big Data: heurísticas y marcos interpretativos. *DILEMATA, International Journal of Applied Ethics*, 87-103. Obtenido de <http://www.dilemata.net/revista/index.php/dilemata/article/view/412000042/450>
- El Financiero. (24 de 08 de 2017). ¿Cuánto se gana en las carreras más populares en México? Obtenido de elfinanciero.com.mx: <https://www.elfinanciero.com.mx/universidades/cuanto-se-gana-en-las-carreras-mas-populares-en-mexico>
- El Informador. (12 de Enero de 2019). *Combaten deserción con estrategias de nivelación*. Obtenido de informador.mx: <https://www.informador.mx/Combaten-desercion-con-estrategias-de-nivelacion-l201901120001.html>
- Erazo, L., Valle, R., Rojo, L., Campillo, L., & Ruiz, J. A. (Febrero - Mayo de 2013). Factores de Abandono Escolar en las carreras de Área de las Ciencias Biológicas, Químicas y de la salud de la UNAM. *III CLABES. TERCERA CONFERENCIA LATINOAMERICANA SOBRE EL ABANDONO EN LA EDUCACION SUPERIOR UNAM*. (págs. 1-10). Perspectiva Económica de America Latina.
- Fozdar, B., Kumar, L., & Kannan, S. (Febrero - Mayo de 2006). A Survey of a Study on the Reasons Responsible for Student Dropout from the Bachelor of Science Programme at Indira Gandhi National Open University. *International Review of Research in Open Distance Learning*.(38), 1-15.
- Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"* (Fourth ed.). Morgan Kaufmann.
- Freitas, A. A. (2002). *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. The Netherlands: Springer-Verlag.
- Garza Carranza, S. (Febrero - Mayo de 2000). *Determinación de los factores que propician la deserción escolar en el Instituto Tecnológico de Nuevo Laredo*. San Nicolas de los Garza N. L.: Universidad Autonoma de Nuevo Leon. Facultad de Ingenieria Mecanica y Electrica. Subdireccion de Estudios de Postgrado.
- Google. (2019). *Google*. Obtenido de Formularios: https://www.google.com/intl/es_mx/forms/about/
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Amsterdam: Morgan Kaufmann.



- INEE. (2019). *AT02d – Tasa de abandono total*. Obtenido de Instituto Nacional para la Evaluación de la Educación México: <https://www.inee.edu.mx/evaluaciones/panorama-educativo-de-mexico-isen/at02d-tasa-de-abandono-total>
- INEGI. (2018). *Estadísticas a propósito del día mundial de la población (11 de julio)*. Ciudad de México: INEGI. Obtenido de https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2018/poblacion2018_Nal.pdf
- Jimenez Moscovitz, L., & Rengifo, P. (2010). Al Interior de una Máquina de Soporte Vectorial. (U. d. Valle, Ed.) *Revista de Ciencias*, 14, 73-85. Obtenido de https://s3.amazonaws.com/academia.edu.documents/50836645/Art05-73.pdf?response-content-disposition=inline%3B%20filename%3DAI_Interior_de_una_Maquina_de_Soporte_Ve.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20200215%2Fus-e
- Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2003). Preventing Student Dropout in Distance Learning Using Machine Learning Techniques. *International conference on knowledge-based and intelligent information and engineering systems*, (págs. 267-274). Springer, Berlin, Heidelberg.
- Lavado, P., & Gallegos, J. (2005). La dinámica de la deserción escolar en el Perú: un enfoque usando modelos de duración. *Centro de Investigación de la Universidad del Pacífico*, 1-46.
- Londoño Ardila, L. F. (Febrero - Mayo de 2013). Factores de riesgo presentes en la deserción estudiantil en la Corporación Universitaria Lasallista. *Revista Virtual. Universidad Católica del Norte*.(38), 183-194.
- Longest, K. C. (2019). *Using Stata for Quantitative Analysis*. California, USA: SAGE Publications.
- Lugo, B. (Octubre-Diciembre de 2013). La Deserción Estudiantil: ¿ Realmente es un problema social?. *Revista de Postgrado. FACE-UC*, 7(12), 289-309.
- M. P. van der Aalst, W. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. London, UK: Springer-Verlag.
- Malene, R., Bjornoy, Sommersel, H., & Sogaard Larsen, M. (Abril de 2013). Evidence on Dropout Phenomena at Universities. *Danish Clearinghouse for Educational Research*, 1-52. Obtenido de <http://edu.au.dk/en/research/research-areas/danish-clearinghouse-for-educational-research/>
- Márquez-Vera, C., Cano, A., Romero, C., Mohamad Noaman, A. Y., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 107-124.
- Márquez-Vera, C., Romero Morales, C., & Ventura Soto, S. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. *Revista Iberoamericana de Tecnologías del/da Aprendizaje/Aprendizagem*, 7(3), 109-117. Obtenido de <https://bit.ly/2zoZKmo>
- Mateos Ayala, A. (2018). *Deserción Escolar: Causas, Consecuencias y Características*. Obtenido de Lifeder.com: <https://www.lifeder.com/desercion-escolar/>



- MathWorks. (2019). *Algoritmos de Machine Learning para clasificación (SVM)*. Obtenido de la.mathworks.com: <https://la.mathworks.com/discovery/support-vector-machine.html>
- Mitchell, T. M. (2000). *Decision Tree Learning*. Obtenido de Washington State University: <https://bit.ly/2N1AI32>
- Molano, M. J. (24 de 01 de 2017). *Mejor, mínimo, un salario*. Obtenido de México cómo vamos: <https://mexicocomovamos.mx/?s=contenido&id=639>
- Morillas, A. (2014). *Muestreo en poblaciones finitas*. Obtenido de Academia.edu: http://www.academia.edu/download/40425074/muestreo_en_las_poblaciones_finitas.pdf
- Muñoz Izquierdo, C. (1979). El síndrome del atraso escolar y el abandono del sistema educativo. *Revista latinoamericana de estudios Educativos*., IX.
- Navarro Sandoval, N. L. (2001). Marginación escolar en los jóvenes. Aproximación a la causas de abandono. *Revista de información y Análisis*.(15), 43-50.
- OCDE. (2019). *Educación Superior en México. Resultados y Relevancia para el Mercado Laboral*. París: OECD Publishing. doi:<https://doi.org/10.1787/9789264309432-en>.
- OECD. (2017). *Skills Strategy Diagnostic Report: Mexico 2017, OECD Skills Studies*. París: OECD Publishing. doi: <http://dx.doi.org/10.1787/9789264287679-en>.
- OECD. (2019). *Higher Education in Mexico: Labour Market Relevance and Outcomes, Higher*. París: OECD Publishing. doi:<https://doi.org/10.1787/9789264309432-en>
- Páramo, G. J., & Correa, C. A. (1999). Deserción Estudiantil Universitaria. Conceptualización. *Revista Universidad Eafit*, 65-78. Obtenido de <http://publicaciones.eafit.edu.co/index.php/revista-universidad-eafit/article/view/1075/967>
- Ruiz Calderón, L. (2009). Deserción en la Educación Superior recinto Las Minas. Período 2001-2007. *Ciencia e Interculturalidad*., 4(2), 30-46.
- Secretaría de Gobernación. (30 de Abril de 2019). *Plan Nacional de Desarrollo 2019-2024*. Obtenido de gaceta.diputados.gob.mx: <http://gaceta.diputados.gob.mx/PDF/64/2019/abr/20190430-XVIII-1.pdf>
- SEP. (22 de Noviembre de 2007). *Programa Sectorial de Educación 2007-2012*. Obtenido de dof.gob.mx: http://dof.gob.mx/nota_detalle.php?codigo=5028684&fecha=17/01/2008
- SEP. (Diciembre de 2013). *Programa Sectorial de Educación 2013 - 2018*. Obtenido de Gobierno de México: https://www.gob.mx/cms/uploads/attachment/file/11908/PROGRAMA_SECTORIAL_DE_EDUCACION_2013_2018_WEB.compressed.pdf
- SEP. (2019). *Abandono escolar*. Ciudad de México: SEP.
- Silva Laya, M. (24 de Junio de 2011). El primer año universitario. Un tramo crítico para el éxito académico. *Perfiles educativos*, 33(ISSN 0185-2698), 102-114. Obtenido de

http://www.scielo.org.mx/scielo.php?pid=S0185-26982011000500010&script=sci_arttext&tlng=en

- Sistema Nacional de Información Estadística Educativa. (2019). *Estadísticas Educativas*. Obtenido de <http://www.sniesep.gob.mx>: http://www.sniesep.gob.mx/x_entidad_federativa.html
- Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive Modeling of Student Dropout Indicators in Educational Data Mining Using Improved Decision Tree. *Indian Journal of Science and Technology*, 9 (4)(0974-5645), 1-5. doi:10.17485/ijst/2016/v9i4/87032
- Urbina-Nájera, A. B., Cruz-Barbosa, R., Morales-Salgado, M., Olivera-Pérez, E., Juárez-Maldonado, O., & Camino-Hampshire, J. C. (2019). *Mecanismos para reducir el índice de deserción escolar universitaria aplicando aprendizaje automático: Caso IES Puebla*. Puebla, México: UPAEP-Universidad. Proyecto interdisciplinario financiado a través de la convocatoria 2018-Fondo a la Investigación UPAEP.
- Valle Gómez, R., Eslava Gómez, G., Manzano Patiño, A., & García Minjares, M. (2014). *Encuesta Internacional sobre el Abandono en la Educación Superior*. Unión Europea. Obtenido de <https://rida2.utp.ac.pa/bitstream/handle/123456789/3140/Informe-resultados-encuesta-abandono-Educacion-Superior.pdf?sequence=2&isAllowed=y>
- Vélez, A., & López Jiménez, D. (2004). Estrategias para vencer la deserción universitaria. *Revista Educación y Educadores*, 7, 177-203.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (Third ed.). Burlington, MA: Morgan Kaufmann Publishers.
- Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2a. ed.). New Zealand: Elsevier.
- Witten, I., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. ELSEIVER.
- Yamao, E., Saavedra, L. C., Campos Pérez, R., & Hurtado, V. (2018). Prediction of academic performance using data mining in first year students of peruvian university. *CAMPUS*, XXIII, 151-160. doi:10.24265/campus.2018.v23n26.05
- Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. *European Journal of Open, Distance and e-Learning*, 17(1). doi:10.2478/eurodl-2014-0008