UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INFORMÁTICA

TESIS

FACTORES DE DESERCIÓN PARA SEGMENTAR LOS ALUMNOS DEL INSTITUTO SUPERIOR TECNOLÓGICO PRIVADO ISTEPSA DURANTE EL PERIODO 2019

PRESENTADA POR:

WALTER BORDA NAVEDOS

PARA OPTAR EL GRADO ACADÉMICO DE:

MAGISTER SCIENTIAE EN INFORMÁTICA

MENCIÓN EN GERENCIA DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES

PUNO, PERÚ

2021

DEDICATORIA

Dedico este trabajo con mucho cariño a mis familiares, que siempre estuvieron ahí para apoyarme cuando los necesité, entre ellos mis padres, hermanos y querida tía; porque son parte esencial de mi día a día, fuente de inspiración para seguir adelante y lograr todas las metas que me propuse y siempre soñé alcanzar.

AGRADECIMIENTOS

Primeramente, agradecer a mi asesor por demostrar compromiso e identificación con este trabajo de investigación y que además me enseñó la importancia de trabajar en equipo.

De igual manera, agradezco a los docentes de la Escuela de Postgrado de la Universidad Nacional del Altiplano de Puno, por transmitirme sus conocimientos y experiencias con paciencia y dedicación. Gracias a los jurados quienes me han ayudado a reconocer los errores, orientado por nuevos caminos y de esa manera pulir las ideas.

ÍNDICE GENERAL

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

CAPÍTULO III

MATERIALES Y MÉTODOS

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

Puno, 19 de mayo del 2021

ÁREA : Informática.
TEMA : Factores de deserción.
LÍNEA: Desarrollo de aplicaciones.

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

La deserción de alumnos en los niveles superiores de estudio es preocupante por ello esta investigación se desarrolla en el Instituto Superior Tecnológico Privado ISTEPSA de la ciudad de Andahuaylas el cual tiene 427 alumnos matriculados en el semestre académico 2019-II, para lo cual se ha planteado el siguiente problema; ¿Cuáles son los factores y patrones que permiten segmentar los alumnos con riesgo de deserción del Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019?, para cuya solución se aplica técnicas de Aprendizaje Automático en WEKA: Se aplicó el método de evaluación CfsSubsetEval y el método de búsqueda BestFirst para identificar los factores, para establecer los patrones se usó el algoritmo de asociación A priori y para segmentar, se usó el algoritmo de Maximización del Valor Esperado "Expectation Maximissation" (EM) y mapas auto organizados de Kohonen en inglés Self Organizing Maps (SOM). Obteniendo los siguientes resultados: 06 factores significativos: Motivación de sesiones, Laboratorios y Aulas de la Institución, Aceptación de la carrera profesional, Cursos Repetidos en el colegio y Semestre Académico; para los patrones de deserción el 100% de los estudiantes que se retiran califican como deficiente la motivación, aulas y laboratorios; además el 96% consideran deficiente a la carrera profesional que estudian y 90% de los que se retiran son de cuarto semestre; En la segmentación se ha construido 3 grupos con el algoritmo EM y 4 grupos para el algoritmo SOM, donde se observa que los factores académicos son determinantes para la deserción de alumnos.

Palabras clave: Aprendizaje automático, deserción, motivación, instituto superior, segmentación de alumnos.

ABSTRACT

The desertion of students in the higher levels of study is worrying, for this reason this research is carried out at the ISTEPSA Private Technological Institute of the city of Andahuaylas, which has 427 students enrolled in the academic semester 2019-II, for which it has been proposed the next problem; What are the factors and patterns that allow the segmentation of students at risk of dropping out of the Higher Private Technological Institute ISTEPSA, during the 2019 period?, for whose solution Automatic Learning techniques are applied in WEKA: The CfsSubsetEval evaluation method and the BestFirst search method were applied to identify the factors, to establish the patterns the a priori association algorithm was used and to segment, the algorithm Maximization of Expected Value "Expectation Maximissation" (EM) and self-organizing maps of Kohonen in English Self Organizing Maps (SOM). Obtaining the following results: 06 significant factors: Motivation of sessions, laboratories and classrooms of the institution, acceptance of the professional career, repeated courses in the school and academic semester; For dropout patterns, 100% of students who dropout rate motivation, classrooms, and laboratories as deficient; in addition, 96% consider the professional career they are studying to be deficient and 90% of those who withdraw are from the fourth semester; in the segmentation, 3 groups have been constructed with the EM algorithm and 4 groups for the SOM algorithm, where it is observed that the academic factors are decisive for the dropout of students.

Keywords: Automatic learning, desertion, higher institute, motivation, student segmentation.

INTRODUCCIÓN

Las problemáticas nacionales referentes al abandono de alumnos en las entidades de formación profesional en sus diferentes niveles son muy preocupantes, para el caso de Institutos técnicos privados este problema es aún más preocupante porque las tazas de deserción son más altos, ello puede deberse a diversos factores y patrones como; Sociales, económicas, políticas, demográficas y académicas que son causales para la deserción de alumnos. En vista a lo descrito la presente investigación aborda la siguiente temática: Factores y patrones que permiten segmentar los alumnos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA de la ciudad de Andahuaylas, Región Apurímac; para ello se utilizará metodologías y algoritmos de minería de datos y otros conceptos de inteligencia artificial; los cuales serán aplicados a la información obtenida de todos los alumnos de las 04 carreras profesionales de este Instituto, Desarrollo de Sistemas de Información, Contabilidad Computarizada, Administración de Negocios Internacionales y Administración de Empresas Turísticas y Hoteleras.

El móvil de la presente investigación son los alarmantes índices de deserción que se han identificado en el Instituto, puesto que actualmente se tiene un aproximado de 34% de alumnos que desertan durante la formación profesional el cual dura un periodo de 03 años, esta problemática afecta directamente a los objetivos del Instituto Superior Tecnológico Privado ISTEPSA, como es la de formar profesionales con capacidades técnicas, ser una entidad auto sostenible entre otros objetivos.

Para la comprensión adecuada de esta problemática es necesario identificar plenamente las causales que determinan la deserción de alumnos, para ello nos enfocaremos en características sociales, económicas, demográficas y académicas de los alumnos; específicamente en aquellas características que pueden representarse de manera cuantitativa puesto que los algoritmos de minería de datos elegidos trabajan con distancias. Por otro lado, la metodología usada en la presente investigación es el KDD, por sus siglas en inglés que significa Descubrimiento de Conocimiento en Base de Datos; la metodología propone 07 fases: a) Determinación de las fuentes de información, b) Diseño del esquema de un almacén de datos, c) Implantación del almacén de datos, d) Selección, limpieza y transformación de los datos que se van a analizar, e) Selección y aplicación del método apropiado de mineración, f) Evaluación, interpretación,

transformación y representación de los patrones extraídos y g) Difusión y uso del nuevo conocimiento.

Las técnicas de minería de datos y aprendizaje automático serán aplicadas a la información obtenida de todos los alumnos matriculados en el semestre académico 2019-II siendo un total de 427, para el recojo de dicha información se utilizó una ficha elaborada y contrastada con antecedentes a este proyecto, asimismo la participación de la directora del Instituto.

Una vez aplicadas la metodología y técnicas tenemos como objetivos: *Identificar* los factores de deserción, *Establecer* los patrones de deserción y *Segmentar* los alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.

El presente trabajo de investigación está estructurado por capítulos de la siguiente manera: CAPITULO I, Este capítulo contiene toda la parte teórica necesaria para fundamentar y comprender el trabajo de investigación, asimismo los antecedentes recopilados de trabajos de investigación recientes que servirán como referencia y con ello encaminar de manera apropiada esta investigación; CAPITULO II, en esta sección se plantea el problema desde una perspectiva nacional, regional y local, enmarcado en principios metodológicos y científicos, en este capítulo también se plantean los objetivos que se persiguen para resolver la problemática planteada, finalmente tenemos la justificación, donde se expone la importancia de este trabajo de investigación y el beneficio de los resultados para la entidad donde se desarrollará la investigación; CAPITULO III, en este capítulo se describe los materiales y métodos utilizados durante el proceso de investigación los cuales permitirán la validación de las hipótesis planteadas; CAPITULO IV, en esta sección se escriben los resultados obtenidos con los algoritmos de minería de datos, siendo específicamente los factores, patrones y segmento de alumnos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA, en esta sección también tenemos la discusión sobre los argumentos que justifican los resultados obtenidos.

CAPÍTULO I

REVISIÓN DE LITERATURA

1.1. Marco teórico

1.1.1. Inteligencia Artificial (IA)

Mathivet (2018), indica que la Inteligencia Artificial es: "Un concepto difícil de definir con precisión, porque puede adoptar muchas formas. Resulta difícil, también, medirla, y las pruebas de C.I. están sesgadas. Podría resumirse como la capacidad de adaptación al entorno para resolver los problemas que se le presentan" (p 28). En efecto al respecto hay mucha información y diversas definiciones que se utilizan de acuerdo a la perspectiva de cada autor y es que éste término fue acuñado en 1956 por John McCarthy y desde entonces ha tenido avances muy importantes y también su definición ha evolucionado sin embargo todas tienen algo en común, así como (Terrones, 2018), quién define la Inteligencia Artificial como, "La idea de crear y dar forma a programas de ordenador o también máquinas que sean capaces de desarrollar conductas que serían consideradas inteligentes si las realizara un ser humano" (p 145).

1.1.2. Redes Neuronales Artificiales

Las redes neuronales artificiales son modelos matemáticos que buscan representar el funcionamiento del cerebro humano mediante la implementación de técnicas en los procesadores o computadoras que cuentan con altos niveles de rendimiento con la finalidad de explorar y reproducir conocimiento, en la actualidad se aplica a muchos campos con la finalidad de

resolver problemas complejos que hasta la fecha solo posible resolverlas mediante el razonamiento humano; es importante diferenciar entre un proceso tradicional y el proceso de una red neuronal puesto que el primero refiere a una situación donde el computador realizará un conjunto de secuencias establecido en la instrucción, sin embargo para el segundo caso consiste en aplicar un modelo matemático que en función a la data histórica y datos de entrenamiento aproximarán un resultado a partir de los casos conocidos simulando el comportamiento del cerebro humano (Redondo, 2016).

1.1.3. Aprendizaje Automático

También conocido como machine Learning en inglés; Baviera (2016), menciona que: "El aprendizaje automático nació en el campo de la informática cuyo procesamiento de datos es una suerte de aprendizaje. Dicho con otras palabras: la máquina no se programa para que responda de una determinada forma según las entradas recibidas, sino más bien para que extraiga patrones de comportamiento a partir de las entradas recibidas, y en base a dicha información aprendida o asimilada, realice la evaluación de nuevas entradas." (p 36).

Como podemos apreciar se mencionan los términos aprendida y asimilada por lo que podemos concluir que el aprendizaje automático demanda un grado de independencia en el proceso de aprendizaje, esto tiene un significado muy importante en el avance de la Inteligencia Artificial, se tiene los siguientes tipos de aprendizaje automático.

A. Aprendizaje no Supervisado

Mativet (2017), menciona que este tipo de aprendizaje es el menos común, puesto que no se espera ningún resultado y es usada para hacer agrupamiento o también llamado Clustering como por ejemplo en una base de datos de clientes donde se busca obtener las distintas categorías en función a características económicas o sociodemográficas de acuerdo a los objetivos de la investigación, para ello se desconoce inicialmente cuántos grupos se obtendrá o cuales son las características de los grupos, ésta técnica se caracteriza por que busca minimizar la distancia entre los datos de un grupo

(Consistencia de datos) y maximizar la distancia entre grupos conformados, mientras más separados se encuentren los grupos significa que la técnica ha sido apropiada para el caso de estudio.

B. Aprendizaje Supervisado

El aprendizaje supervisado es la técnica que permite a las computadoras aprender sin antes haber sido programadas explícitamente mediante el entrenamiento de algoritmo a través de ejemplos o datos de entrenamiento que serán denominados objetos, luego se compara el resultado obtenido por la red con los resultados que son conocidos o también llamados etiquetas, en caso de que el resultado obtenido por la red presente márgenes de error se deben aplicar técnicas para disminuir el error y finalmente dar o aproximar con los resultados esperados, los parámetros usados en el entrenamiento serán las variables que el investigador manipule para determinar un modelo apropiado para el caso de estudio (Mativet 2017), como se puede observar en este caso se supervisa el entrenamiento del algoritmo de forma iterativa hasta encontrar el modelo que aproxime los resultados reales

1.1.4. Clustering

Cestero y Caballero (2018), refiere que todo objeto científico es afines con la clasificación y la reducción de las generalidades a modelos más sencillos de manejar, la clasificación nos ayuda a constatar hipótesis sobre un grupo de elementos observados, en la actualidad se observa que existen clasificaciones de todo tipo de elementos mediante los cuales se busca constituir perfiles y patrones de comportamiento, los modelos de Clustering para fines de precisión en los resultados obtenidos recurren a modelos matemáticos el cual es más objetivo, sin embargo la subjetividad también es requerida como por ejemplo al momento de decidir la cantidad de grupos o segmentos en las que se va agrupar los datos en observación, en la figura 1 se observa diferentes clasificaciones para los datos.

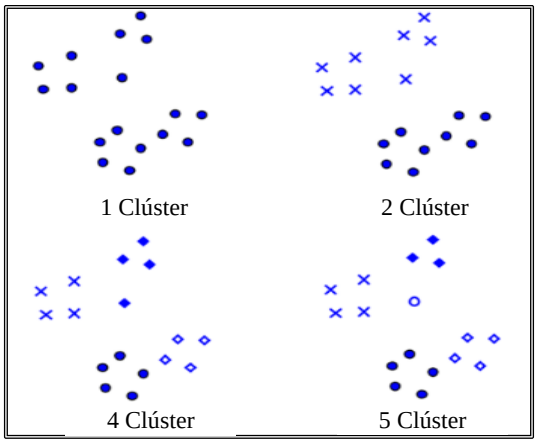


Figura 1. Diversos Clustering para datos en observación.

Fuente: Cestero y Caballero (2018).

El Clustering es usado en diversos escenarios y es por ello que se han desarrollado diversas técnicas de agrupamiento, para fines de lograr el objetivo de estudio en el presente caso de investigación se han considerado las siguientes técnicas por ser las más usadas y representativas.

1.1.5. Minería de Datos

Este término fue acuñado aproximadamente en 1960, sin embargo, logra consolidarse en 1980, de acuerdo a Himansu *et al.* (2017), este concepto engloba la idea de analizar grandes volúmenes de datos con la finalidad de extraer patrones y en base a ello los encargados de la toma de decisiones en las empresas puedan resolver problemas complejos, éste procedimiento demanda de la utilización de un algoritmo el cual debe ser elegido de acuerdo a la situación o problema que se desea resolver; se tiene diversas opciones como algoritmos de regresión, agrupamiento, clasificación, etc. Cada uno de éstos puede ser usado en cualquier escenario sin embargo el éxito y eficiencia estará sujeto a la experiencia del investigador para determinar el mejor algoritmo para ello se recomienda realizar las siguientes operaciones: Establecer claramente los resultados esperados, luego debe se debe preparar la data de acuerdo al algoritmo elegido y luego de la aplicación de la metodología se deberá realizar una adecuada interpretación de los resultados obtenidos.

1.1.6. Herramientas de Minería de Datos

A. Análisis Factorial

Hamilton (1992), indica que para validar el instrumento cuestionario se aplica la prueba de la medida de adecuación de la muestra Kaiser-Meyer-Olkin (KMO) la cual indica que las variables miden factores comunes cuando el índice es mayor a 0.7, asimismo, se realiza la prueba de esfericidad de Bartlett que permite definir estadísticamente si la matriz de interrelación es una matriz de identidad, una aplicación del análisis factorial es el método de factores principales, y el propósito fundamental es determinar la estructura de los dominios de los factores buscando la presencia de variables latentes no observables

B. WEKA 3.9.4

En español Entorno para el Análisis del Conocimiento de acuerdo al sitio oficial; está escrito en java y es una colección de algoritmos para trabajar con minería de datos a través del aprendizaje automático, WEKA contiene funcionalidades para la preparación de datos, clasificación, regresión, agrupación, extracción de reglas de asociación y visualización. WEKA es software libre bajo y se distribuye bajo la licencia GNU, una de sus grandes ventajas es su alto nivel de portabilidad y su facilidad de uso gracias a su interfaz sencilla; su nombre es en honor a un ave que no tiene la capacidad de volar sin embargo tiene como característica principal el comportamiento de investigar detalladamente. (WEKA3, 2019).

WEKA tiene implementado técnicas de evaluación y de búsqueda, estos reconocen a los factores como atributos que para la presente investigación será equivalente estos dos términos al momento de realizar las evaluaciones e interpretación de resultados.

1.1.7. Metodología KDD

Las tecnologías de la mineración de datos buscan encontrar patrones sobre una cantidad extensa de información y conocimiento extrayendo tendencias y modelos, donde la interpretación de dicha información represente un valor agregado, en el contexto del descubrimiento de Conocimiento en Bases de Datos (KDD), consta de una jerarquía que existe en una base de datos, además, entre datos, información y conocimiento.

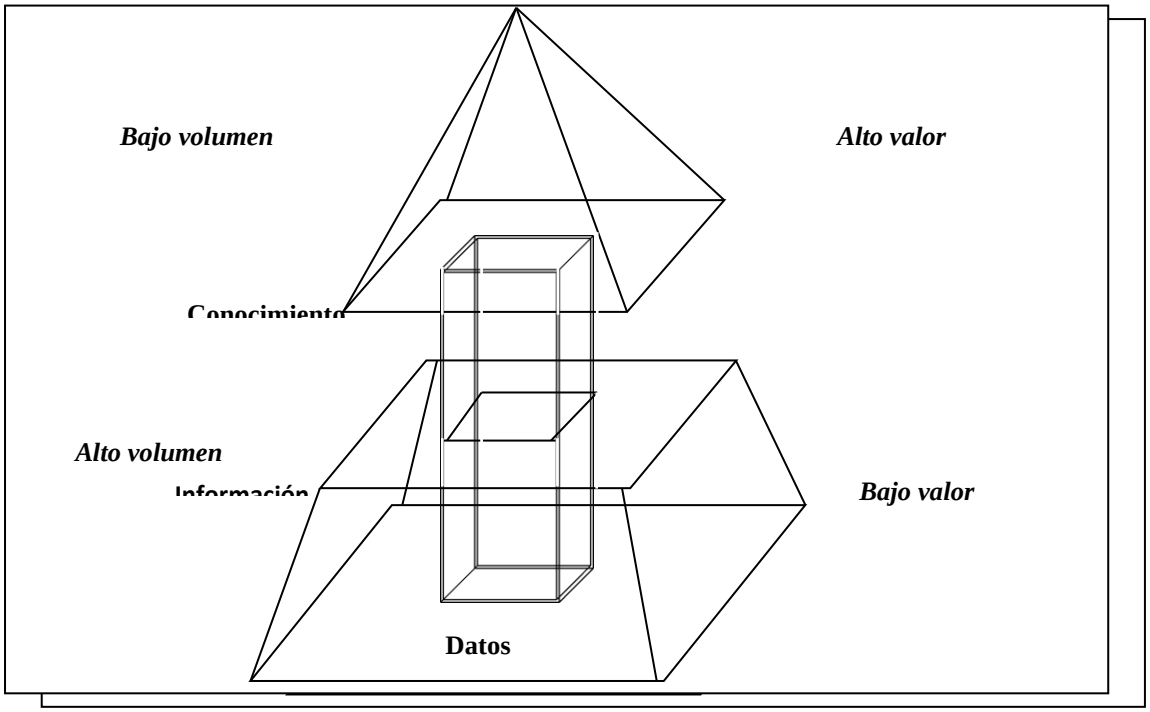


Figura 2. Jerarquía de la base de datos; entre datos, información y conocimiento.

1.1.8. Descubrimiento de Conocimiento en Bases de Datos (KDD)

El KDD es el "Proceso de extracción no trivial de identificar patrones válido, novedoso, útil y, comprensible a partir de los datos" para:

- ✓ Procesar automáticamente grandes cantidades de datos crudos.

- ✓ Identificar los patrones más significativos y relevantes.
- ✓ Presentar como conocimiento apropiado para satisfacer las metas del usuario.

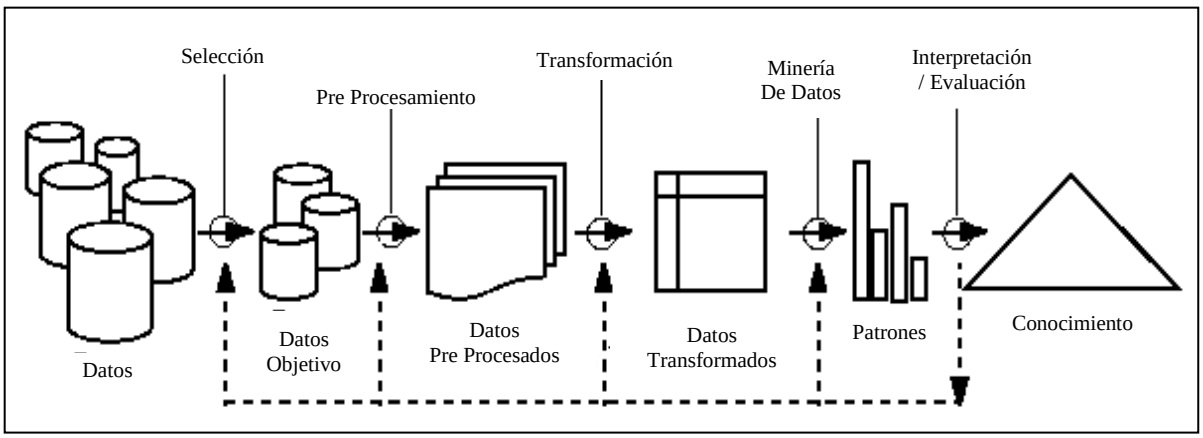


Figura 3. El proceso KDD de Extracción de Conocimiento.

La metodología seguida por la herramienta software Weka es el proceso KDD conocida como minería de datos (algoritmos) para extraer (identificar) conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post-procesamientos y cuya interpretación de los patrones extraídos es el nuevo conocimiento, asimismo, representa patrones de comportamiento observados en los valores de las variables o indicadores (atributos) del problema o relaciones de asociación entre dichas variables, en combinación con diversas técnicas generan distintos modelos, considerando que cada técnica requiere un pre procesado diferente de los datos, para ello, es necesario seguir las siguientes fases:

A. Determinación de las fuentes de información.

Este es la primera fase del proceso de investigación y para lo cual existen diversos instrumentos que el investigador puede usar, para nuestro caso utilizaremos el instrumento cuestionario estructurado el cual permitirá recoger datos cuantitativos de los alumnos del Instituto de Educación Superior

Tecnológico Privado ISTEPSA a través de preguntas formuladas en concordancia a otras investigaciones similares los cuales nos brindarán información estadística, así mismo se recurrirá a otras fuentes de información como son los reportes académicos de la Institución educativa como el número de alumnos matriculados en los anteriores periodos académicos en las carreras profesionales que ofrece esta entidad.

B. Diseño del esquema de un almacén de datos

Para unificar de manera operativa toda la información requerida y lograr los objetivos de esta investigación se ha elaborado el esquema de COPO DE NIEVE, puesto que las dimensiones identificadas requieren la implementación de más de una tabla de datos, ésta esquema esta en concordancia con el cuestionario aplicado a los alumnos de todas las carreras profesionales del Instituto de Educación Superior Tecnológico Privado ISTEPSA, este modelo presenta un grado mayor de normalización en comparación con el modelo ESTRELLA, lo cual permitirá la eliminación de datos redundantes.

ESQUEMA COPO DE NIEVE

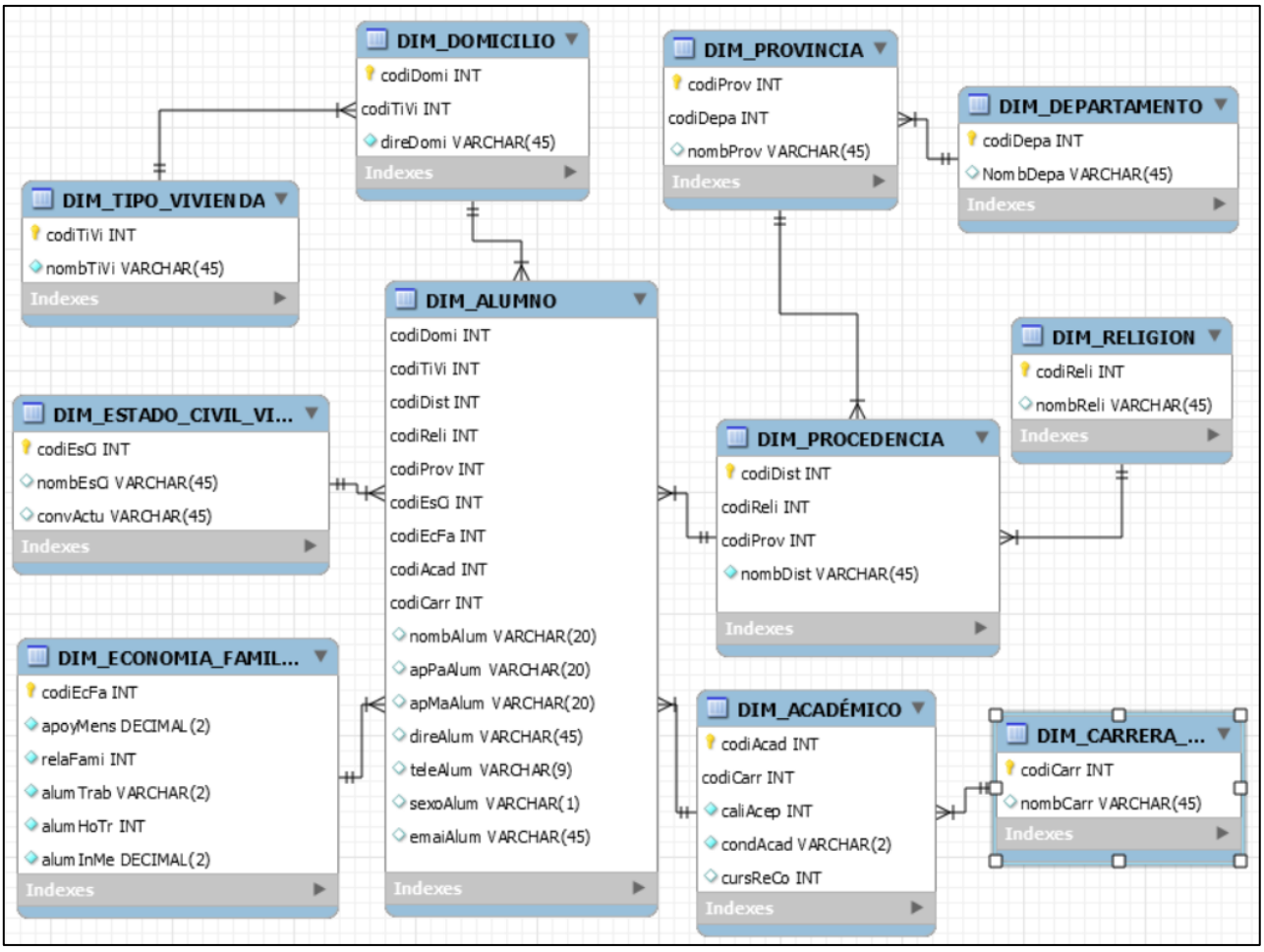


Figura 4. Esquema de copo de nieve.

C. Implantación del almacén de datos

Mediante este proceso se permitirá la "navegación" y visualización previa de sus datos, para discernir qué aspectos interesa ser estudiados. Para el análisis de los datos obtenidos de los alumnos a través del cuestionario estructurado se ha descargado a un libro Excel, puesto que los datos están se han organizado por dimensiones (tablas) como es los aspectos: Económicos, Académicos, Geográficos, Género y Social.

Dichas dimensiones están relacionadas entre sí a través de cada unidad de observación (Alumnos), los cuales se irán agrupando más delante de acuerdo a las similitudes y proximidades que exista ente cada objeto de estudio.

D. Selección, limpieza y transformación de los datos que se van a analizar

Consiste en estructurar adecuadamente los datos obtenidos detectando aquellos erróneos o irrelevantes para luego ser descartados, a continuación, los datos son descargados en el libro Excel, así mismo será necesario codificar los atributos puesto que los algoritmos no supervisados que se utilizarán están basados en distancias. La limpieza y pre-procesamiento de datos se logra diseñando una estrategia adecuada para manejar ruido, valores incompletos, secuencias de tiempo, casos extremos (si es necesario),

E. Selección y aplicación del método apropiado de mineración

Para esta fase se ha elegido la metodología KDD (Descubrimiento de Conocimiento de Base de Datos), esta metodología nos permitirá identificar modelos válidos, útiles y entendibles que describa patrones de deserción de los alumnos, es importante especificar que la metodología no es una fórmula maestra que nos permitirá obtener los patrones directamente, sino que es necesario tener en claro los objetivos del análisis y de acuerdo a ello trabajar con el algoritmo que satisfaga mejor las necesidades del estudio.

F. Evaluación, interpretación, transformación y representación de los patrones extraídos

Para el análisis de los datos recopilados se utilizarán algoritmos de análisis de clúster y de redes neuronales no supervisados como son: CFS: Selección de Características basada en Correlación, El algoritmo de Maximización del Valor Esperado “Expectation Maximissation” (EM) y Mapas auto organizados (SOM).

G. Difusión y uso del nuevo conocimiento

comprende incorporar el conocimiento descubierto al sistema (normalmente para mejorarlo) donde el conocimiento se obtiene para realizar acciones, ya sea incorporándolo dentro de un sistema de desempeño o simplemente para almacenarlo y reportarlo a las personas interesadas. Es

decir, el Instituto de Educación Superior Tecnológico Privado ISTEPSA con los resultados obtenidos en esta investigación podrá tomar medidas para fortalecer sus habilidades y evitar la deserción de alumnos puesto que tendrá un segmento focalizado para dirigir sus acciones.

1.1.9. Representación de patrones

Se distinguen dos técnicas de representación no simbólicas y simbólicas.

A. Técnicas no simbólicas

son las más numerosas y tradicionales apropiadas para variables continuas y con un conocimiento más claro de lo que se busca. El inconveniente de éstas técnicas es poca (o nula) inteligibilidad, destacan algoritmos basadas en: Redes Neuronales Artificiales, Lógica Difusa, Algoritmos Genéticos y combinaciones entre ellos.

B. Técnicas simbólicas

generan un modelo "legible" y además aceptan mayor variedad de variables y mayor riqueza en la estructura de los datos. Árboles de Decisión, Programación Inductiva y Otras Técnicas de Machine Learning.

1.1.10. Selección de subconjunto de Atributos

Según Hall & Smith (1998), postula que el problema de la selección de subconjuntos de atributos es muy conocido en estadística y reconocimiento de patrones. Sin embargo, muchas de las técnicas tratan exclusivamente con variables continuas, donde para muchos algoritmos prácticos de aprendizaje automático presenta la suposición común (monotonicidad), es decir, al aumentar el número de atributo no disminuye el rendimiento, el enfoque para la selección de subconjuntos de características en el aprendizaje automático utiliza técnicas de búsqueda y evaluación de atributos o sub conjunto de atributos.

De acuerdo a Gil (2018), indica que el modelo equivalente en estadística es el análisis de componentes principales, ésta es una técnica de aprendizaje

no supervisado puesto que a diferencia de las técnicas de aprendizaje supervisado donde hay un conjunto de valores que permiten predecir el resultado para el caso de aprendizaje no supervisado existe el total de atributos donde se buscará comportamientos similares y de esa manera se forma subconjuntos o subgrupos de factores. La técnica de análisis de componentes principales (PCA) aplica la reducción de dimensionalidad (Variables) mantenimiento la mayor cantidad de información posible de acuerdo a la Varianza de dichos atributos, PCA reduce el número de variables transformadas (Componentes Principales) que representan la variabilidad de los datos. Para cada componente principal que se genera con PCA será una combinación lineal de las variables originales. Otro método según Koller & Sahami (1996), elimina las características cuyo contenido de información (sobre otras características y la clase) está subsumido por algunas de las características restantes. Otros métodos intentan clasificar las características de acuerdo con una puntuación de relevancia (Kira & Rendell, 1992; Holmes & Nevill-Manning, 1995).

1.1.11. CFS: Selección de Características basada en Correlación

Utiliza un algoritmo de búsqueda que permite la selección de características para la obtención de un subconjunto pequeño que sea representativo del problema original. La heurística mediante el cual CFS mide la "bondad" de los subconjuntos de atributos tiene en cuenta la utilidad de los atributos individuales para predecir la etiqueta de clase junto con el nivel de intercorrelación entre ellas.

Los buenos subconjuntos de atributos contienen atributos altamente correlacionados predictivas de la clase:

$$G_S = \frac{k\overline{r}_{ci}}{\sqrt{k + k(k-1)\overline{r}_{ii}}} \qquad (1)$$

Dónde: G_S es el mérito heurístico del subconjunto S conteniendo k características, r_{ci} es el valor la correlación media entre la clase y la

característica y $\bar{r}_{ii}$ es la mejor correlación entre dos características del conjunto S. El método CFS asume que los atributos son independientes condicionalmente dada la clase, esto es una simplificación aceptable en algunos casos, pero si existe una fuerte interacción entre distintos atributos, entonces CFS no garantiza que los atributos seleccionados sean relevantes, k es el número de características en el subconjunto.

El coeficiente de incertidumbre simétrico se encuentra entre 0 y 1. Un valor de 0 indica que X e Y no tienen asociación; el valor 1 para la relación de ganancia indica que el conocimiento de Y predice completamente X; el valor 1 para el coeficiente de incertidumbre simétrico indica que el conocimiento de una variable predice completamente la otra. Ambos muestran un sesgo a favor de atributos con menos valores.

$$H(Y) = \sum_{y=1} p(y) \log_2 (p(y)) \qquad (2)$$

$$H(Y/X) = \sum_{X=1} p(x) \sum_{y=1} p(y \mid x) \log_2 (p(y \mid x)) \qquad (3)$$

$$ganancia = H(Y) - H(Y \mid X)$$

$$= H(X) - H(X \mid Y)$$

$$= H(Y) + H(X) - H(Y, X)$$

$$\text{Ratio ganancia} = \frac{ganancia}{H(X)}$$

$$\text{incertidumbre simétrica} = 2.0 * \frac{ganancia}{H(Y) + H(X)}$$

$$P(C_i \mid v_1, v_2, ..., v_n) = \frac{P(C_i)\prod P(v_j \mid C_i)}{P(v_1, v_2, ..., v_n)} \qquad (4)$$

1.1.12. Técnicas de Asociación

Según Tan et al. (2006), definen las reglas de asociación del algoritmo **a priori**, como:

Sea $\boldsymbol{I} = (\boldsymbol{i_1}, \boldsymbol{i_2}, \ldots \boldsymbol{i_n})$, un conjunto de atributos llamados ítems

Sea $\boldsymbol{D} = (\boldsymbol{t_1}, \boldsymbol{t_2}, \ldots \boldsymbol{t_n})$, un conjunto de transacciones almacenados en la base de datos

Cada transacción D tiene ID (identificador) único con subconjunto de ítems de I.

La fuerza de la asociación es medida de acuerdo con su soporte (Support) y su confidencia (confidence). El Soporte determina cómo a menudo una regla es aplicable a un conjunto de datos, por ende, constituye un índice de generación de las combinaciones entre los elementos Una regla se define como una implicación de la forma:

$$X \Rightarrow Y$$

Dónde: $X, Y \subseteq I$ y $X \cap Y \neq 0$ los conjuntos de items X y Y se denominan respectivamente "ANTECEDENTE" y "CONSECUENTE" de la regla.

Suport (cobertura): expresa el porcentaje o fracción de registros de D que satisfacen la unión de los elementos del antecedente y consecuente de la regla $s(X \Rightarrow Y) = s(X \cup Y)$.

Confidence (confianza): es la medida de la efectividad de la regla, representa el porcentaje de casos en los que dado el antecedente se verifica la implicación $c(X \Rightarrow Y) = s(X \Rightarrow Y) / s(X)$, puede utilizarse para estimar la probabilidad condicionada del consecuente dado el antecedente: $P(X/Y) = P(X \cup Y) / P(X) = c(X \Rightarrow Y)$

Lift (levantamiento): cuantifica la relación existente entre X e Y: se define como: lift($s(X \Rightarrow Y) = s(X \Rightarrow Y) / s(Y)$ según su valor obtenido se concluye:

$lift > 1$: X e Y positivamente correlacionados

$lift < 1$: X e Y negativamente correlacionados

$lift = 1$: X e Y son independientes

Leverage (apalancamiento):

$(X \Rightarrow Y) = s(X \Rightarrow Y) - s(X)s(Y) = P(X \cap Y) - P(X)P(Y)$

Conviction (convicción):

$(X \Rightarrow Y) = 1 - s(Y)/(1 - conf\,(Y \Rightarrow Y)) = P(X)P(Y')/P(X \cap Y')$

Tanto el soporte como la confianza definen el grado de interés de una regla de asociación, una regla con un valor de soporte ocurre simplemente por casualidad, un valor de confianza alto indica que el porcentaje de transacciones que contienen a X también a Y de manera conjunta.

Cada transacción está asociada con un identificador único, llamado TID. Sea X un grupo de elementos. Se afirma que una transacción T contiene X si y solo si X⊆T. Una regla de asociación se define como una expresión X⇒Y, donde X e Y son conjuntos de elementos no vacíos (es decir, X⊆I, Y⊆I). Esta regla se denomina antecedente, tal que X∩Y = ∅. La regla X⇒Y se cumple dentro del conjunto de transacciones D con soporte s, donde s% de transacciones en D que contienen X∪Y. La regla X⇒Y tiene confianza c, dentro del conjunto de transacciones D, siempre que el c% de las transacciones en D contengan X que también contenga Y.

Soporte: La regla X⇒Y tiene soporte s dentro del conjunto de transacciones D, si este es el caso de transacciones en D contiene X∪Y. Las reglas que

tienen una s mayor o igual a un soporte especificado por el usuario se denominan umbral de soporte mínimo (min_sup) Soporte (X⇒Y) = Soporte (X∪Y) = P(X∪Y)

Confianza: la regla X⇒Y tiene confianza c dentro del conjunto de transacciones D, si las transacciones recordadas en D contienen X que también contienen Y. Las reglas que tienen una c mayor o igual a una confianza especificada por el usuario se denominan umbral de confianza mínimo (min_conf).

Confianza (X⇒Y) = (apoyo(X∪Y)) / (apoyo (X)) = P (Y / X)

Por lo general, se utilizan valores de confianza grandes y un soporte menor. Las reglas que satisfacen cada soporte mínimo y confianza mínima se conocen como reglas sólidas. Dado la información grande y la preocupación de la alta dirección por la deserción de estudiantes, se predefine umbrales de apoyo y confianza para eliminar las reglas que no parecen ser tan notables o útiles. (Belamate et al. 2016).

A. Buscar todos los elementos (conjuntos de elementos) con un soporte de transacciones superior al soporte mínimo. Estos son los conjuntos de elementos frecuentes. Conjunto de elementos alternativo denominado conjuntos de elementos poco frecuentes.

B. Utilice los conjuntos de elementos frecuentes para obtener las reglas especificadas.

Existe una gran unión entre la literatura de que el subproblema principal es que el principal de los dos es necesario. Esto se debe a que lleva más tiempo debido al enorme espacio de búsqueda y, por lo tanto, la sección de generación de reglas se puede hacer en la memoria principal de una manera muy simple una vez que se encuentran los conjuntos de elementos frecuentes.

1.1.13. Extracción de segmentos

Recientemente en el análisis de clúster o la segmentación de casos en muchas disciplinas científicas se utilizan Sistemas Modulares, Mezcla de Expertos y Sistemas Híbridos previo a la búsqueda de soluciones al problema

que se plantea en cada momento, basada en el enfoque de visión de las distintas partes que forman el todo, transformando la tarea inicial compleja, en un conjunto de sub tareas más elementales, susceptibles de ser abordadas de manera más sencilla y eficiente, luego, requiere integrar los resultados parciales obtenidos de cada una de esas sub tareas y generar la solución al problema completo, una práctica conocida como método de "divide y vencerás" que aborda la mezcla de expertos, algunos de ellos basados en técnicas estadísticas extrapolables a las redes neuronales utilizados ampliamente en tareas genéricas (especialmente el Perceptrón Multicapa usando como algoritmo de aprendizaje el de Retropropagación del Error), o bien en tareas más específicas, típicamente de clasificación o clustering (cuyo exponente más habitual entre las redes neuronales artificiales lo forman los mapas autoorganizados de Kohonen y algoritmo de Maximización del Valor Esperado).

1.1.14. El algoritmo de Maximización del Valor Esperado "Expectation Maximissation" (EM)

Según Jordan & Jacobs (1994), una alternativa para el ajuste de los parámetros que definen la mezcla jerárquica de expertos es el uso del algoritmo de maximización del valor esperado (EM), el fundamento de éste algoritmo es la tarea de maximizar el parámetro *L* que sería más sencilla si pudiera conocerse los valores que toman un conjunto de parámetros que permanecen desconocidos, por ejemplo:

$$z_{ij} = \begin{cases} 1 \text{ si es el experto j del conglomera do i el que genera la salida} \\ 0 \text{ en cualquier otro caso} \end{cases}$$, para

$$L = \ln\left(\prod_{t=1}^{N} P(y^{(t)} / x^{(t)}, \theta)\right) = \sum_{t=1}^{N} \ln\left(\sum_{i=1}^{K} g_i^{(t)} \sum_{J=1}^{L} g_{j/i}^{(t)} P_{ji}(y^{(t)})\right) \quad (5)$$

Basado en sistemas modulares de Jacobs - Jordan. Un modelo que se ajusta fácilmente al caso particular de los sistemas compuestos por **mezcla o superposición de procesos estocásticos**, donde, cada módulo *i* constituye una regla o experto que produce una salida y_i, fruto de un proceso aleatorio cuya función de distribución para muchos casos prácticos suele considerarse

gaussiana de media μ_i. Este valor μ_i es el valor medio de la respuesta deseada ***y*** condicionado a conocer el vector de entradas **x**, con lo que su valores medios coincidirán: $y_i=\mu_i$.

Entonces la función de distribución de la salida deseada **y** condicionada al conocimiento de la entrada **x** es:

$$P(y/x)=\sum_{i=1}^{K} g_i P_i(y/x) \quad (6)$$

Siendo g_i las correspondientes salidas de las redes de puertas.

$$P(y/x)=\frac{1}{(2\pi)^{k/2}}\sum_{i=1}^{K} g_i \exp\left(-\frac{1}{2}\|y_i-\mu_i\|^2\right) \quad (7)$$

Para el caso particular de mezcla de gaussianas con matriz de covarianzas identidad, la expresión anterior se resume:

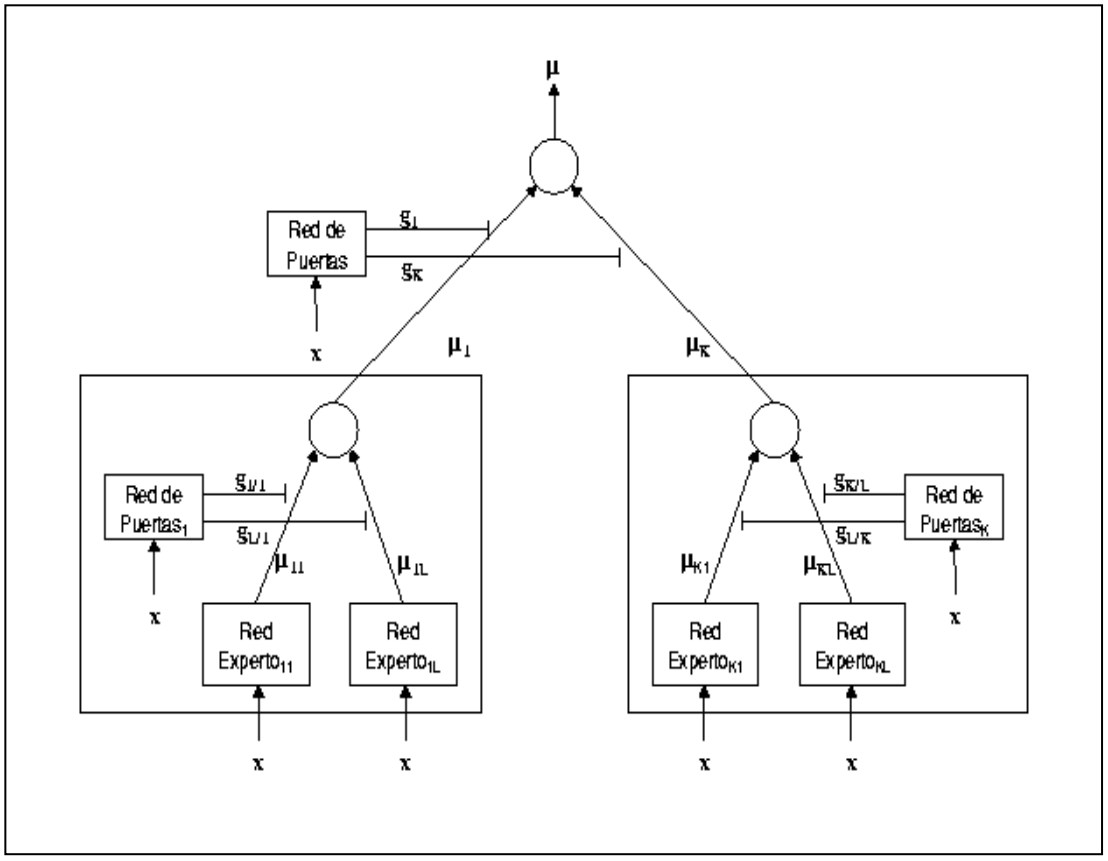


Figura 5. El modelo de Jacobs-Jordan de dos niveles de módulos expertos, y las redes de puertas (gating networks).
Fuente: Sancho (2000).

Un sistema genérico dispone una jerarquía de expertos, tal y como se aprecia en la figura 4 un modelo jerárquico formado por un árbol de dos niveles de expertos. El primer nivel, que es el más profundo, está constituido por K bloques de L expertos cada uno, cuyos resultados se combinan por varios módulos de redes de puertas, dando origen a K conglomerados de expertos, y éstos a su vez se combinan por otra red de puertas para generar la salida.

Para este sistema en particular, se considera que cada una de las redes de expertos lleva asociada una distribución de probabilidad P_{ij}, que será función implícita de los parámetros de los que dependa el experto $\mathbf{w}_{ij}$ y de las entradas y salidas que se hayan utilizado para su ajuste$\{(x^{(t)}, y^{(t)}), t=1...N\}$. Las denominadas redes de puertas del primer nivel generarán un conjunto de salidas $\{g_{j/i}\}$,i=1...K, j=1...L, y la red de puertas del segundo nivel generará un conjunto $\{g_i\}$, i=1...K que en ambos casos dependen de los parámetros $\{u_{ji}\}$ y $\{v_{ji}\}$ respectivamente, y además de los pares de entrada y salida deseada utilizados durante su ajuste. Dado que las salidas de todas las redes de puertas se van a comportar como distribuciones de probabilidades que ponderan la participación de cada módulo experto en la salida final, los $\{g_i\}$y $\{ g_{j/i} \}$ habrán de ser todos positivos y sumar uno; una manera de conseguir esto es mediante la utilización de la función softmax. Así, si se denomina por ξ_i a la activación correspondiente a la salida i-ésima de la red de puertas del segundo nivel, los valores g_i se generarían por medio de la fórmula:

$$g_i = \frac{\exp \xi_i}{\sum_{j=1}^{k} \exp \xi_j} \quad (8)$$

Una fórmula similar para los coeficientes $g_{j/i}$ es:

$$P(y/x,\Theta)=\sum_{i=1}^{K} g_i(x,v_i)\sum_{j=1}^{L} g_{j/i}(x,v_{j/i})P_{ji}(y/x,w_{ji}) \quad (9)$$

El sistema jerárquico a dos niveles, incluye de forma explícita la dependencia con los parámetros de todos los subsistemas:

Donde θ es el conjunto de parámetros que definen el sistema, que incluye los de los expertos $\mathbf{w}_{ji}$, y los de las redes de puertas $\mathbf{v}_i$ y $\mathbf{v}_{ji}$.

Note que el esquema expone todos los niveles y módulos que reciben como entrada el mismo vector **x**.

Previo a la descripción de algún método de ajuste de los parámetros del sistema, se definen las siguientes probabilidades condicionales a posteriori:

$$h_i=\frac{g_i\sum_{j=1}^{L} g_{j/i}P_{ji}(y)}{\sum_{i=1}^{k} g_i\sum_{j=1}^{L} g_{j/i}P_{ji}(y)} \quad (10)$$

El valor h_i representa la probabilidad de que el agrupamiento i-ésimo de expertos genere la respuesta deseada **y**.

También se define otro conjunto de probabilidades a posteriori, que dan cuenta de la probabilidad de que el experto j-ésimo del agrupamiento i-ésimo genere una determinada salida deseada **y**:

$$h_{j/i}=\frac{g_{j/i}P_{ji}(y)}{\sum_{j=1}^{L} g_{j/i}P_{ji}(y)} \quad (11)$$

Una manera de medir la bondad de los resultados obtenidos con el sistema es a través de la probabilidad de que dado un vector de entrada se obtenga su correspondiente vector de salida asociado. Si este mismo objetivo se debe cumplir simultáneamente para todos los pares de entrada y salida usados en el ajuste del sistema, un buen parámetro de evaluación sería el producto de las distribuciones de probabilidad que ofrece el sistema para todos los pares de datos utilizados en el entrenamiento:

$$Q = \prod_{t=1}^{N} P(y^{(t)} / x^{(t)}, \theta) \quad (12)$$

Este parámetro Q recibe el nombre de verosimilitud (**"likelihood"** en inglés). Cuanto mayor sea este parámetro, mayor será la probabilidad de que el sistema asocie todos los vectores de entrada con sus correspondientes salidas, y como es natural, durante el proceso de ajuste del sistema, el objetivo será hacer máximo su valor, o lo que es equivalente.

En EM la tarea de calcular el valor de la función de coste *L* sería trivial si se dispusiera de un conjunto **Z** constituido por $\{z_i\}$ y $\{z_{j/i}\}$. Estas variables hacen las veces de etiquetas que identifican para cada vector de entradas **x** cuál es el conglomerado de expertos que debe tenerse en cuenta en el proceso de generación de la salida **y**, y cuál de todos los módulos que forman el conglomerado i es el que en concreto genera la salida. Así se podría definir otro conjunto de variables $\{z_{ij}=z_i z_{j/i}\}$, tal que z_i y $z_{j/i}$ no son conocidas, ya que, si lo fueran, el problema del aprendizaje estaría resuelto, porque sólo se ajustaría el módulo y la conexión oportuna. Gracias a la introducción de estas variables, la expresión de *L* será:

$$\text{CC } L = \sum_{t=1}^{N} \ln\left(\sum_{i=1}^{K} g_i^{(t)} \sum_{J=1}^{L} g_{j/i}^{(t)} P_{ji}(y^{(t)}) \right) = \sum_{t=1}^{N} \ln\left(\prod_{i=1}^{K} \prod_{j=1}^{L} (g_i^{(t)} g_{j/i}^{(t)} P_{ji}(y^{(t)}))^{z_{ij}^{(t)}} \right)$$

(13)

Gracias a que la variable z_{ij} hace que cada término producto sea 1 en el caso de que no sea el experto responsable de esa salida deseada **y** (y por lo tanto no afecte al resto de términos del producto), y que valga justamente $P_{ji}(\mathbf{y})$ cuando se trate del módulo experto correcto. De esta forma la función logaritmo se reescribe como suma de logaritmos:

$$L = \sum_{t=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{L} z_{ij}^{(t)} \left(\ln g_i^{(t)} + \ln g_{j/i}^{(t)} + \ln P_{ji}(y^{(t)}) \right) \quad (14)$$

El algoritmo EM se lleva a cabo de forma iterativa en los siguientes pasos:

A. **Paso E**: cálculo de la esperanza matemática de sobre el conjunto formado por todos los pares de entrenamiento:

$$E(\theta, \theta^{(p)}) = E(L(\theta, Z) / X) = \sum_{i=1}^{K} \sum_{j=1}^{L} z_{ij}^{(t)} \left(\ln g_i^{(t)} + \ln g_{j/i}^{(t)} + \ln P_{ji}(y^{(t)}) \right) \quad (15)$$

donde $\theta^{(p)}$ es la estimación de los parámetros en la iteración p, **Z** es el conjunto de variables ocultas, y se ha tenido en cuenta además que $E(z_{ij}^{(t)} / \mathbf{X}) = h_{ij}^{(t)}$.

B. **Paso M**: obtener la siguiente estimación de los parámetros $\theta^{(p+1)}$ que

$$\theta^{(p+1)} = \arg\max_{\theta} E(\theta, \theta^{(p)}) \qquad (16)$$

C. maximice el valor esperado estimado calculado en la fase E:

Este problema de maximización se reduce a la maximización de cada uno de los tres sumandos más interiores que aparecen en las siguientes operaciones:

$$w_{ji}^{(p+1)} = \arg\max_{w_{ji}} \sum_{t=1}^{N} h_{ji}^{(t)} \ln P_{ij}(y^{(t)})$$
$$v_{i}^{(p+1)} = \arg\max_{v_{i}} \sum_{t=1}^{N} \sum_{k=1}^{K} h_{k}^{(t)} \ln g_{k}^{(t)} \qquad (17)$$
$$v_{ji}^{(p+1)} = \arg\max_{v_{ji}} \sum_{t=1}^{N} \sum_{k=1}^{K} h_{k}^{(t)} \sum_{l=1}^{L} h_{l/k}^{(t)} \ln g_{l/k}^{(t)}$$

Analizando los términos $\sum\sum h_{ji} \ln g_i$ y $\sum\sum h_{ji} \ln g_{j/i}$ se asimilan a entropías conjuntas, que miden la entropía de la distribución de los patrones **x** entre los conglomerados de expertos y los expertos respectivamente. De acuerdo con esta interpretación, el valor esperado *E* se maximiza cuando los conglomerados son mutuamente excluyentes, y disminuye cuando existen datos de entrada que hacen que se activen simultáneamente más de un conglomerado. De forma análoga se puede razonar con el segundo término para cada uno de los expertos que forman los conglomerados. En cuanto al tercer término, $\sum\sum h_{ji} P_{ji}(y)$ indica que los expertos que más pesan en el valor de *E* son aquellos cuya probabilidad h_{ij} es mayor (Moerland, 1997).

1.1.15. Mapas auto organizados

Los mapas auto organizados de Kohonen en inglés es Self Organizing Maps (SOM) son redes neuronales artificiales (RNA) llamada red de Kohonen usada para clasificar información y reducir el número de variables de análisis específico, no importa cuántas variables sean, esta RNA visualiza la información en mapas bidimensionales que preservan y reflejan la estructura de similitud entre la información entrante. El aspecto visual es una ventaja del método de clasificación frente a otros métodos cuando hay más de tres variables, la visualización del proceso de clasificación se vuelve enormemente compleja. RNA se caracteriza por su aprendizaje competitivo, es decir que los modelos de neuronas compiten entre sí para saber cuál es más parecido al patrón de entrenamiento presentado, con lo cual se actualiza el

peso de la neurona ganadora, en mayor proporción que el peso de las neuronas vecinas. La proporción de actualización en neuronas que pertenecen a la vecindad de la neurona ganadora disminuye en función de la distancia a esta. Cuanto mayor sea la similitud entre dos patrones de entrenamiento, menor será la distancia entre sus neuronas ganadoras. Esto proporciona la sensación de modelo auto organizado, porque a medida que se entrena la red, las neuronas ganadoras de patrones similares forman vecindarios independientes que finalmente reflejan los grupos de patrones similares.

Los SOM funcionan de manera similar a Escalamiento Multidimensional (MDS), pero en lugar de intentar reproducir distancias, su objetivo es reproducir la topología, intenta mantener los mismos vecinos. En tanto, si dos objetos de alta dimensión (p>2) son similares, entonces su posición en un lugar bidimensional también debería ser muy similar. En lugar de mapear objetos en un espacio continuo (2-D), SOM utiliza una cuadrícula regular de "unidades" en las que se mapean los objetos en un gráfico 2-D con MDS: una distancia entre dos objetos se interpreta directamente como una "estimación" de la distancia real entre los objetos en el espacio dimensional superior concentrando mayores diferencias, mientras que en un gráfico de SOM este no es el caso: es decir, los objetos mapeados en la misma unidad o en unidades vecinas son muy similares concentrándose en las mayores similitudes, son análogos a la agrupación de k-medias. En esa analogía, cada unidad del mapa SOM corresponde a un "grupo", el número de grupos se define por el tamaño de la cuadrícula, que normalmente se dispone de forma rectangular o hexagonal. Esto es lo que ocurre en el cerebro y es similar al método de modelado de aprendizaje supervisado de las redes neuronales, Sea $X = \{x_1, x_2, x_3, ..., x_m\}$ el conjunto de datos de entrada, el algoritmo básico para la generación de mapas auto organizados, requiere:

Crear la red de N neuronas e iniciar los vectores de peso w de manera aleatoria.

A. Presentar el dato $x(t)$ y encontrar la neurona ganadora, n_c, como sigue: $$\| x(t) - w_c \| = \min_i \min^N \{ \| x(t) - w_i \| \} \quad (18)$$

B. Actualizar los vectores de referencia con la siguiente regla de aprendizaje:

$$w_i(t+1) = w_i + \alpha(t) h_{ci}(t)[x(t) - w_i(t)]$$

dónde: $$h_{ci}(t) = \exp\left(\frac{\left\| r_c - r_i \right\|^2}{2\sigma^2(t)} \right) \quad (19)$$

$h_{ci}(t)$ Es llamada función vecindad; α es el factor de aprendizaje; r_c y r_i son, respectivamente, los vectores de localización (en la retícula plana) de la neurona ganadora y la neurona que está siendo actualizada.

C. Si se alcanza el número de iteraciones deseadas, el algoritmo termina su ciclo, de lo contrario recurre al paso 2.

Función de vecindad. En la ecuación (3) se utiliza una función vecindad, $h_{ci}(t)$, de forma gaussiana. Esta función controla el grado de conexión entre las neuronas de la retícula plana durante el entrenamiento (mayor distancia corresponde a una interacción más débil). Así, los vectores de peso correspondientes a las neuronas más cercanas a la neurona ganadora, se actualizan usando un factor de mayor magnitud.

De acuerdo a la fórmula (3), la función de vecindad $h_{ci}(t)$ depende del tiempo y el rango de alcance (respecto de neuronas vecinas). Este rango depende de los valores que asume la función $\sigma(t)$, la cual determina la

amplitud de la gaussiana. Generalmente se elige una función $\sigma(t)$ decreciente, para que el radio de influencia de la neurona ganadora se vaya estrechando a medida que procede el entrenamiento.

Una manera socorrida de definir sigma es la siguiente:

$$\sigma(t) = \left\{ \begin{array}{ll} R & para\ \ t < t_g \\ R(1 - \frac{t}{t_{\max}}) & para\ \ t \geq t_g \end{array} \right\}, \qquad (20)$$

Donde el parámetro R (llamado "radio máximo de la gaussiana") se escoge en proporción al tamaño de la retícula y tg es el tiempo de ordenamiento global de la red. El tiempo restante es para cubrir la etapa de refinamiento.

Factor de Aprendizaje. La función alfa es la responsable de garantizar la convergencia del proceso de entrenamiento:

$$\alpha(t) = \left\{ \begin{array}{l} \alpha_{\max}, t < t_g \\ \alpha_{\min}, t \geq t_g \end{array} \right\} \qquad (21)$$

Inicialmente, en esta ecuación, para t pequeño, $\alpha(t)$ asume un valor relativamente grande (cercano a 1). Esto permite el ordenamiento global de la red, durante una primera fase. Posteriormente, el valor de $\alpha(t)$ se disminuye para hacer un ajuste de menor escala en los vectores de peso. A

esta etapa del entrenamiento se le llama refinamiento.

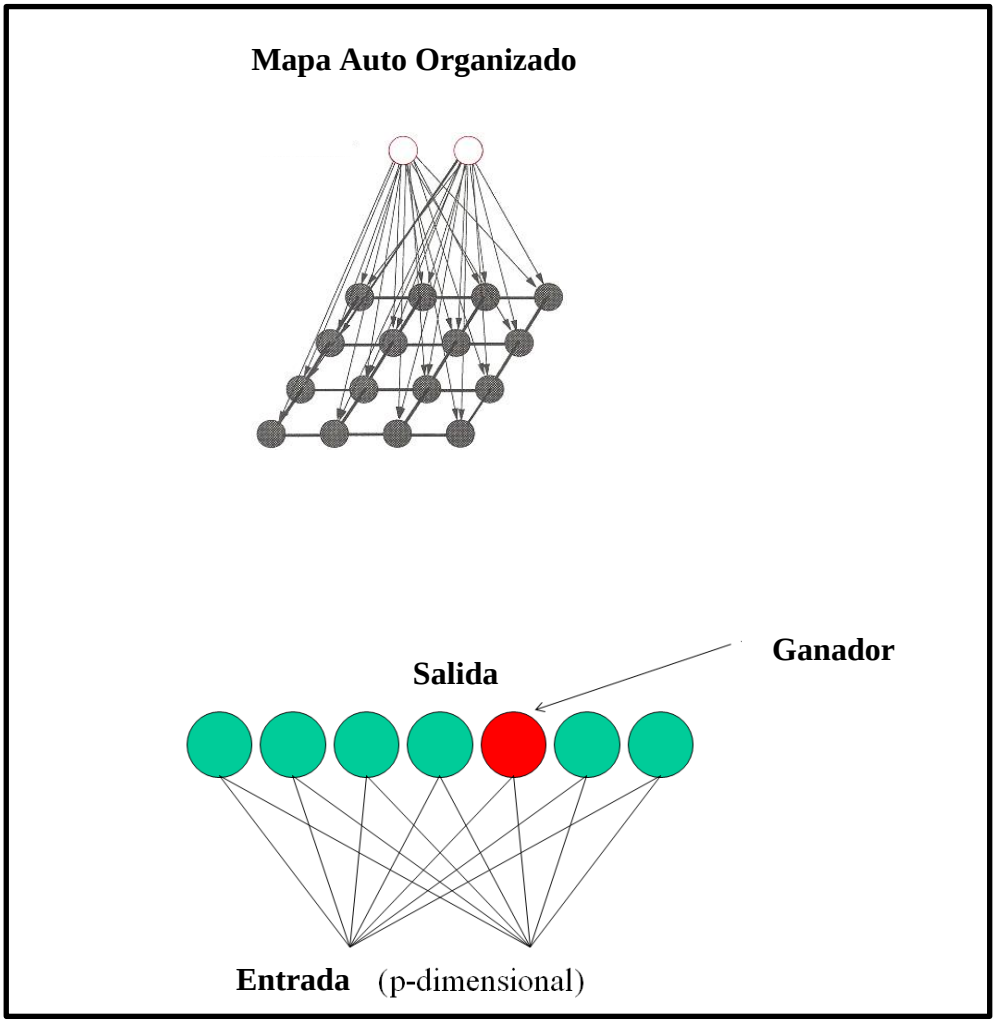


Figura 6. Entrada de estímulos nerviosos.

Considerando que los patrones de entrenamiento se forman únicamente con las variables de análisis del proceso de clasificación y en consecuencia no es necesario incluir variables de salida como por ejemplo el grupo al que pertenece cada patrón, se dice que el entrenamiento de este tipo de redes clasifica como no supervisado. Este aspecto resulta ser otra gran ventaja frente a otros métodos de clasificación de información, que en general necesitan el número predefinido de grupos.

El esquema de arquitectura de este tipo de red neuronal artificial y en la Figura 06 se muestra el diagrama de flujo de su proceso de entrenamiento.

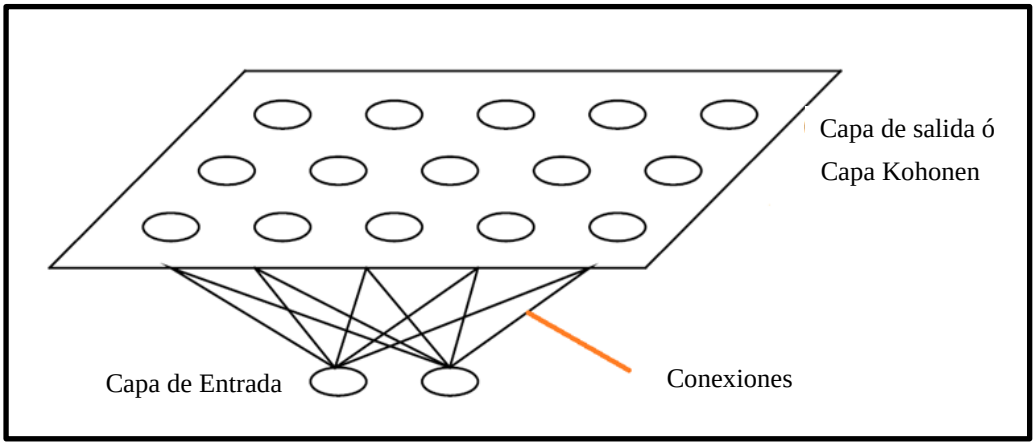


Figura 7. Arquitectura de las redes de Kohonen.

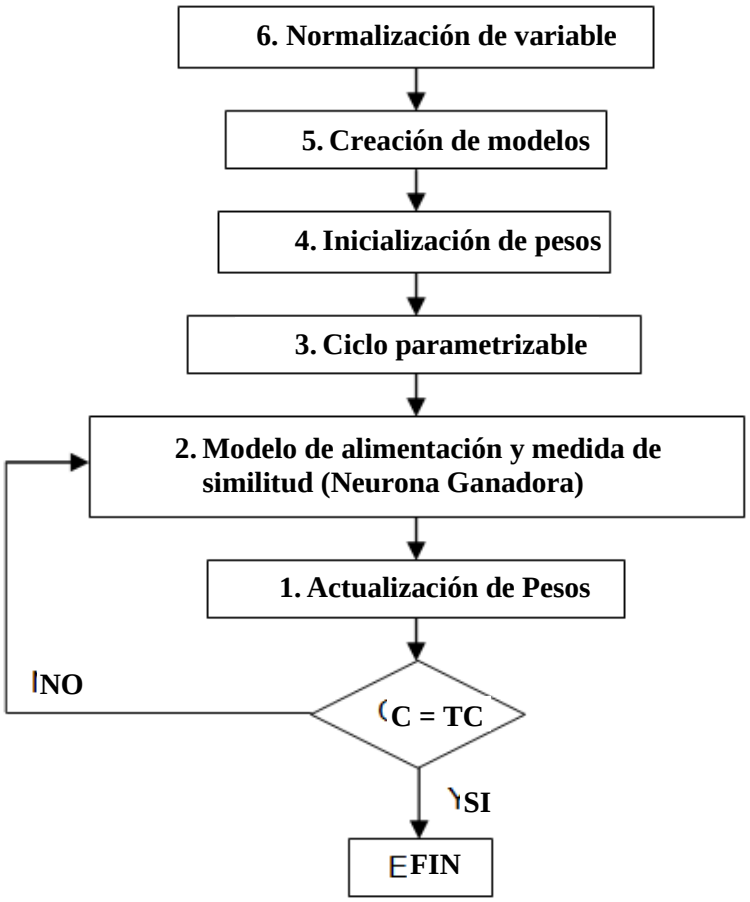


Figura 8. Funcionamiento de la red de Kohonen (C: ciclo y Tc).

En general, la distancia euclidiana se utiliza como métrica de similitud y la actualización del peso de la neurona se da de acuerdo con la ecuación.

$$W_j(t+1) = W_j(t) + \eta(t)h_j(t)(X - W_j(t)) \quad (22)$$

Donde W_j representa el grupo de pesos de la neurona j, t el ciclo correspondiente, la tasa de aprendizaje para el ciclo actual, h_j el factor de ponderación de la neurona j dependiendo de la vecindad establecida para el ciclo actual y respecto a la neurona ganadora y X el conjunto de variables del patrón presentado a la red. El factor de ponderación de la neurona j en función del ciclo y vecindad, con respecto a la neurona ganadora, normalmente se decide con una función gaussiana como la que aparece en la siguiente ecuación:

$$h_j = \exp\left(-\frac{\|u_j - u_j^*\|^2}{2\sigma^2}\right) \quad (23)$$

Los resultados obtenidos en Weka se ilustran a continuación:

Esquema: weka. clusterers. SelfOrganizingMap -L 1.0 -O 2000 -C 1000 -H 2 -W 2

Relación: tesis_deserción

Instancias: 427

Atributos: 23, de manera similar que EM

Modo de prueba: evaluación de clases a grupos en datos de entrenamiento

Modelo de agrupamiento (conjunto de entrenamiento completo) SOM

Número de conglomerados seleccionados mediante validación cruzada: 4

Número de iteraciones realizadas: 1

Parámetros de los segmentos de instancias por Clúster.

1.1.16. Institutos de Educación Superior en el Perú

De acuerdo a la Ley Peruana N° 30512, se define a los Institutos de Educación Superior (IES) en el Perú como:

Instituciones educativas de la segunda etapa del sistema educativo nacional, con énfasis en una formación aplicada. Los IES brindan formación de carácter técnico, debidamente fundamentada en la naturaleza de un saber que garantiza la integración del conocimiento teórico e instrumental a fin de lograr competencias requeridas por los sectores productivos para la inserción laboral, además, estudios de especialización, de perfeccionamiento profesional en áreas específicas y otros programas de formación continua, y otorgan los respectivos certificados (Ley Peruana N° 30512, 2016, p.1).

Los IES han tomado un rol importante en el desarrollo de la sociedad Peruana puesto que se presentan como alternativa de formación de capacidades a corto plazo para los jóvenes con bajos recursos económicos, quienes buscan insertarse en el campo laboral en tiempos más cortos en comparación a lo ofrecido por las universidades tanto nacionales como privadas.

A. Instituto Superior Tecnológico Privado ISTEPSA

De acuerdo a entrevistas sostenidas con la gerencia y revisión de la documentación del Instituto Superior Tecnológico Privado, ISTEPSA (2019), y se sabe que mediante la R.M. N° 0267-2006 E.D. el Ministerio de Educación autorizó a este Instituto el funcionamiento del I semestre del año académico 2006 para ofrecer las carreras técnicas de Computación e Informática y de Contabilidad, actualmente se tiene implementado más dos carreras profesionales, las cuales son: Administración de Negocios Internacionales y Administración de Empresas Turísticas y Hoteleras; con un total de 427 alumnos matriculados en las 04 carreras profesionales en los 6 semestres académicos.

B. Deserción de Alumnos en la ISTEPSA

La deserción de alumnos en el Instituto es una problemática que desde los inicios de funcionamiento se ha presentado y que a la fecha no se ha abordar adecuadamente, se conoce que esta problemática es una de las principales causantes para la quiebra y cierre de las empresas ubicadas en este rubro; por ello es sumamente importante tomar acciones que reduzcan éstos índices.

Se ha evaluado la cantidad de alumnos matriculados en los 6 semestres académicos de las 04 carreras profesionales durante los último 3 años y se conoce que actualmente se tiene un 34 % de deserción de alumnos en el Instituto, las causas muchas veces se desconocen por ello no ha sido posible tomar acciones estratégicas efectivas. El presente trabajo de investigación busca encontrar estos patrones ocultos que determinan el perfil de los alumnos con riesgo de deserción.

1.2. Antecedentes

1.2.1. Antecedentes nacionales

En la ciudad de Tacna, Mollo (2018), desarrolló el trabajo de investigación Análisis predictivo de la deserción estudiantil utilizando data warehouse y minería de datos en la Universidad Nacional Jorge Basadre Grohmann – Tacna, 2012-2018, quién construyó Data Warehouse utilizando la metodología Ralph Kimball y para minería de datos CRISP-DM, juntamente con técnicas de árboles de decisión, regresión logística y redes bayesianas, obteniendo que los indicaron de deserción estudiantil en la Universidad Nacional Jorge Basadre Grohmann fueron el índice de masa corporal (Factores individuales), tupo de ingreso (factores académicos), para esta investigación no se tuvo hipótesis puesto que es de carácter descriptivo, siendo las principales acciones responder los objetivos planteados.

Por otro lado, Holgado (2018), en el trabajo de investigación titulado Detección de patrones de bajo rendimiento académico mediante técnicas de minería de datos de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios, aplicó minería de datos con la metodología CRISP-DM y los

algoritmos Random Forest, obteniendo que las variables más influyentes son: Para el bajo rendimiento académico de estudiantes se considera la cantidad de asignaturas y el servicio de comedor universitario, además la carrera profesional también es influyente en el rendimiento académico donde se deduce que la elección acertada del estudiante en la carrera profesional será muy determinante.

Rivera (2016), realizó la investigación, Los factores determinantes y su relación con la deserción escolar en los alumnos del primero al sexto grados del nivel primario de la institución educativa N° 33160, de Monzón, 2010 al 2015; Donde determinó que el factor principal para el abandono de estudios de los alumnos está centrado básicamente en el factor económico, puesto que se ha determinado que los ingresos mensuales no sobrepasan los S/. 500.00 siendo clasificados en condición económica muy bajas; sumado a ello otros factores personales, puesto que la mayoría de estudiantes reconocen que la integración familiar que tienen no es adecuada lo que permite entender que probablemente existan conflictos en la familia, así mismo también se suma la carga familiar que en promedio tienen de 4 a 6 integrantes y que los estudiantes deben ayudar trabajando para el sustento familiar, disminuyendo el tiempo para dedicarlo al estudio.

En la ciudad de Lima Yamao (2018), desarrolló el trabajo de investigación titulado Predicción del rendimiento académico mediante de minería de datos es estudiantes del primer ciclo de la escuela profesional de ingeniería de Computación y sistemas de la Universidad de San Martín de Porres, donde se logró predecir el rendimiento de los estudiantes ingresante mediante las técnicas planteadas de minería de datos, además se logró identificar de manera temprana a los alumnos que podrían tener dificultades académicas en el futuro y tomar acciones para mitigar el riesgo de esta eventualidad, además se identificó la técnica Support Vector Machines como inapropiada para este trabajo puesto que no arrojó los resultados esperados a pesar de ser una técnica más avanzada a razón de que los datos utilizados para este estudio no guardan la estructura necesaria para dicha técnica.

Torres (2018), en la investigación titulada, Segmentación demográfica y relaciones con los clientes en la empresa Hotel Cielo, Distrito de Tarapoto, utilizó la prueba de Chi – cuadrado de Pearson donde concluye que el nivel de segmentación demográfica en el Hotel Cielo, el 30% es malo, el 64% regular y tan solo el 6% es bueno, por lo que se recomienda optimizar los grupos segmentados hasta obtener grupos homogéneos y así enfocar apropiadamente las estrategias comerciales, en la investigación de tipo descriptivo se ha priorizado la atención a los objetivos planteados en vista que no se ha establecido hipótesis, llegando a la conclusión de que se debe utilizar otras técnicas para obtener segmentos de clientes con características similares y tomar acciones de fidelización apropiadas de acuerdo a los rasgos y necesidades de cada segmento.

Por otro lado De la Cruz (2017), desarrolló el proyecto de tesis, Segmentación de Clientes con Inteligencia Analítica para Personalizar las Ventas de los Servicios de Agencias Turísticas; dicha investigación la realizo en la ciudad de Lima con una población de 1100 clientes que visitaron lugares a través de los servicios de las agencias turísticas, considerando un tamaño de muestra de 570 clientes; llegando a las siguientes conclusiones; La implementación del modelo de inteligencia analítica basada en redes neuronales artificiales K-medias identifica los factores externos sociodemográficos, económicos y factores intrínsecos de lealtad logrando segmentar y definir el perfil de los clientes que han utilizado los servicios de las agencias turísticas.

Además, las redes neuronales son usadas para el pronóstico. Como se evidencia en trabajo de investigación titulado, Pronóstico de la Exportación Pesquera por Redes Neuronales y Modelo Arima, el mismo que tuvo lugar en la ciudad de Trujillo con la finalidad de explicar dos tipos de modelos usados para modelar una serie y determinar el modelo más eficiente para realizar tareas de pronóstico, llegando a las siguientes conclusiones; El mejor modelo para pronóstico de exportación pesquera en el Perú es el modelo Arima asimismo que el modelo más apropiado con redes neuronales para pronóstico de exportación pesquera en redes neuronales es aquella que tiene una capa oculta en la función de activación (Zavala, 2017).

Otra experiencia interesante que podemos observar es en el trabajo de investigación de Linarez (2019), titulado como Predicción de renuncia de socios de una cooperativa utilizando técnicas supervisadas de aprendizaje automático, desarrollado en la ciudad de Arequipa, puesto que la entidad donde se realizó tenía pocos datos y con la finalidad de obtener un resultado más confiable se procedió a generar datos sintéticos, los cuales guardan relación a los datos originales, se aplicó técnicas con las librerías del lenguaje Python obteniendo los siguientes resultados; Que la técnica aprendizaje supervisado automático tiene mayor precisión para la predicción alcanzando un 90.6% de precisión, así mismo se observó que la técnica de bosque aleatorio y potenciación de gradiente son las más adecuadas para la identificación de variables determinantes en la renuncia de socios.

También están los algoritmos de aprendizaje supervisado, como es el algoritmo K-NN el mismo que fue usado por Quezada (2017), en el proyecto de investigación titulado. K-vecino más Próximo en una Aplicación de Clasificación y Predicción en el Poder Judicial del Perú, quién llegó a las siguientes conclusiones; Se encontró el modelo óptimo de clasificación y predicción cuando el valor de k es 3 vecinos más próximos, debido a que el error cuadrático (registro de errores de selección) para tres vecinos es 0.12% mientras que para 4 y 5 vecinos es mayor al 20%, evidenciando que el modelo construido para 3 vecinos es más eficiente por otro lado el modelo de 3 vecinos más próximos encontrado se ejecuta con precisión para tamaño muestra de datos de entrenamiento. Esto debido a que los grupos son distintos. Se demuestra mediante las pruebas estadísticas no paramétricas de Kruskal-Wallis y la Mediana, en ambas pruebas rechazamos la hipótesis de igualdad de promedios y medianas poblacionales respectivamente, y concluimos que los grupos (pequeña, mediana y grande) comparados difieren en cada una de las seis variables. Por tanto, los grupos son distintos.

La minería de datos y las redes neuronales son ampliamente usados para segmentación de datos, como es el caso de Ochoa (2016), Quién desarrollo el trabajo de investigación, Estudio Comparativo de Técnicas no Supervisadas de Minería de Datos para Segmentación de Alumnos, aplicado en el II semestre de la Escuela Profesional de Ingeniería de Sistemas de la

Universidad Católica de Santa María ciudad de Arequipa, utilizó diversas técnicas de agrupamiento y la metodología CRISP-DM ideal para desarrollo de proyectos de minería de datos, después de evaluar la calidad de los agrupamientos a través del Coeficiente de Silueta llegó a las siguientes conclusiones; Que, el algoritmo K-means agrupó los datos con mayor similitud en los clúster y mayor separación entre separación entre los grupos formados, concluyendo finalmente que el algoritmo K-means es la técnica que permite obtener grupos de mejor calidad.

Dentro de la minería de datos esta la rama de la inteligencia de negocios y de manera muy similar a los casos anteriores se basa en los algoritmos de redes neuronales como es el caso de la investigación titulada, Análisis Predictivo Basado en Redes Neuronales no Supervisados Aplicando Algoritmo K-Medias y Crisp-DM para Pronóstico de Riesgo de Morosidad de los Alumnos en la Universidad Peruana Unión, para tal objetivo se utilizó la información sociodemográfico y económica de los alumnos, una aplicado la metodología y algoritmo se llegó a las siguientes conclusiones; Que, el haber utilizado la herramienta BA (Business Analytics) facilitó el trabajo en las fases de definición, diseño y exploración de modelos para la toma de decisiones, así se logró el objetivo propuesto; Además con la creación de un modelo clúster y BA se ha mejorado la toma de decisiones a través del manejo dinámico de reportes para lo cual se consideró una muestra de 130 alumnos (Pacco, 2015).

Aranciaga (2021), realizó el trabajo de investigación titulado Factores asociados a la deserción de estudiantes en el instituto de educación superior privado de Lima, se trabajó con una muestra de 51 alumnos que previamente había abandonado sus estudios, la técnica utilizada fue medir los factores determinantes en la deserción de alumnos en dos programas de estudio, se utilizaron métodos estadísticos para realizar la comparación de los factores propuestos y así obtener los valores de prevalencia entre los factores de deserción, donde se obtuvo que no existe diferencia entre retiro voluntario e involuntario es decir que lo casos presentados de abandono son indistintos a la voluntad del estudiante de de abandonar o no los estudios, otro dato importante es que no existe relación entre factores institucionales, personales,

académicos y económicos en los estudiantes de los programas de Computación y Diseño gráfico.

1.2.2. Antecedentes internacionales

En la ciudad de Guanajuato de México, Castillo (2017), aplicó las técnicas del aprendizaje automático para la predicción de clientes potenciales en procesos de mercadotecnia en vista a que en los últimos años la inversión en publicidad y mercadotecnia se ha incrementado notablemente por ello en esta investigación se ha centrado en identificar una técnica de aprendizaje automático que permita predecir que clientes tienen mayor probabilidad de realizar la compra de un producto como resultado de la mercadotecnia directa, para ello se ha utilizado información demográfica y socioeconómica de los clientes; la predicción se realizó con técnicas de clasificación y regresión a través de los algoritmos Random Forest, Gradient Boosting y eXtreme Gradient Boosting; concluyendo finalmente que el modelo con mejor rendimiento es el eXtreme Gradient Boosting por lo que su aplicación en el proceso de mercadotecnia permitirá desarrollar campañas más eficiente en la empresa.

Cifuentes (2016), realizó el trabajo de investigación titulado, Clasificación Automática de Tweets utilizando K-NN y K-Means como algoritmos de clasificación automática, aplicando TF-IDF y TF-RFL para las ponderaciones, el cual se basó en analizar y evaluar el desempeño de los algoritmo K-NN y K-Means en la minería de opinión los mismos que fueron aplicados a un conjunto de Tweets en relación a una empresa de marketing Falabella, obteniendo las siguientes conclusiones; Que la incorporación de la ponderación TF-RFL aumenta el índice de aciertos generados por los algoritmos, además se pudo observar que para el caso de ambos algoritmos al aumentar el porcentaje de entrenamiento generó bajo impacto por lo que no es necesario de gran cantidad de datos para obtener resultados positivos.

La línea del aprendizaje automático consiste en la aplicación y entrenamiento de técnicas con la finalidad de simular conocimiento para posteriormente resolver problemas complejos, este conocimiento adquirido a través de la data histórica y el entrenamiento también puede ser denominado

como nuevo conocimiento; Pavón (2016), desarrolló un trabajo de investigación basado en el aprendizaje automático para resolver problemas complejos que las compañías de cualquier sector viene enfrentando, la idea fundamental es que en función a determinadas variables se defina un procedimiento denominado AIPAKA el cual servirá como herramienta para la solución de problemas complejos, concluyendo que; el modela AIPAKA apoya de gran manera en el análisis de los datos para la toma de decisiones sin embargo no significa que reemplace el análisis de especialistas o profesionales sino que se debe tomar como apoyo, por otro lado el modelo permite la trazabilidad, repetitividad, monitorización y desarrollo, haciendo que sea un modelo predictivo parametrizado y estandarizado para su réplica, finalmente se concluyó que el modelo AIPAKA permite focalizar a la empresas en información importante lo cual conlleva a la eficiencia de recursos para lograr los objetivos deseados.

Berón (2020), en su trabajo de investigación titulado Principales causas de ausentismo laboral: una aplicación desde la minería de datos, de la Universidad Nacional de Colombia, se estudiaron diez variables independientes: sexo, edad, contrato, hijos, casado, antigüedad, turno, trabajo, sindicalizado y escolaridad usando el algoritmo J48 en el programa WEKA, se seleccionaron las variables más influyentes en el ausentismo con una efectividad superior a 94.72%, obteniendo como determinantes las siguientes variables: Sindicalización, hijos, sexo, contrato, estudios, casado y efectividad, siendo las demás variables clasificadas como poco influyentes en el ausentismo laboral.

Tenemos otra experiencia interesante en la investigación realizada por Miranda (2017), en la investigación titulada Análisis de la deserción de estudiantes Universitarios usando técnicas de minería de datos, donde se trabajó con una muestra de 9195 alumnos, y se utilizaron los algoritmos de minería de datos: Redes bayesianas, redes neuronales y árbol de decisión; para extraer modelos, patrones e interpretar se utilizó el proceso KDD, los resultados obtenidos con primera técnica es que se logró clasificar a un 33.9% la cantidad de alumnos que tienden a abandonar sus estudios; y mediante la técnica de árbol de decisión se obtuvo que aquellos alumnos con beneficios

como becas tienen un 89.3% de probabilidad de permanencia; por otro lado se construyó un clasificador mediante redes neuronales mediante el algoritmo Perceptrón multicapa obteniendo las siguientes variables en orden de importancia: Promedio de prueba PSU, beneficios en los últimos 3 periodos, promedio ponderado de postulación, puntaje de nota de enseñanza media del estudiante y beneficios económicos en los últimos 2 periodos.

Por otro lado, en la Universidad Santo Tomás de Colombia se desarrolló el trabajo de investigación Caracterización de los Estudiantes de una Institución de Educación Superior Mediante Big Data por Hoyos & Aponte (2019), para el proceso de Big Data se utilizó la metodología SMART con sus respectivas etapas: definición de la estrategia; captura y medición de los datos; análisis de los datos; generación de un informe de resultados y transformación del negocio. La data estuvo compuesta por las siguientes características de los estudiantes: componente social (ciudad e institución de origen, domicilio, conformación grupo familiar, intereses, aficiones, pertenencia a grupos); componente académico (resultado proceso de admisión, promedios, permanencia, repitencia). Siendo un total de 3908 registros del año 2017 y 12957 registro del 2014 al 2017 obtenidos del sistema académico. Echa la evaluación de datos se ha obtenidos los siguientes resultados, en cuanto a la distribución se observa que en los semestres impares duplica a la cantidad de alumnos, por otro lado se obtuvo que existe una relación fuerte entre las siguientes variables: Niveles cursados, asignaturas aprobadas y número de matrículas; siendo el primer segmento de estudiantes; Por otro lado, se observa que en el caso de la edad, el estrato socioeconómico, y el número de asignaturas perdidas no presentan una relación fuerte con las demás variables analizadas.

Pérez (2020), en su trabajo de investigación titulado Comparación de Técnicas de Minería de Datos Para Identificar Indicios de Deserción Estudiantil a Partir del Desempeño Académico; para lo cual utilizó la técnica de clasificación binaria siendo la metodología CRISP-DM la más adecuada para el proceso de identificación de indicios de deserción; las fase de ésta metodología son: Comprensión del negocio, Conocimiento de la base de datos, preparación de los datos, modelado, evaluación e implementación; la

data para este proceso de análisis fue dotado por la Universidad Privada en Bogotá – Colombia, los datos obtenidos de 762 alumnos en total se han estructura en 4 tablas y 43 columnas; luego de realizar el procedimiento se obtuvo los siguientes hallazgos, que el rendimiento de los cursos de ingeniería de sistemas están relacionados con los cursos de física y matemáticas, por lo que se concluye que los cursos relaciones con números tienen mayor impacto en la deserción escolar.

De manera similar tenemos el trabajo de investigación de Urbina et al. (2020), en su trabajo de investigación titulado Deserción Escolar Universitaria: Patrones Para Prevenirla Aplicando Minería de Datos Educativa, para el análisis de la data se ha utilizado algunas técnicas de selección de atributos y reducción de dimensionalidad para tener una data mas consistente, para ello se ha utilizado algunas técnicas como: Filter, Wrapper y Ranker; por otro lado se ha utilizado el método de búsqueda: CfsSubsetEval y método para medir el grado de redundancia: ConsistencySubsetEval; Se ha visto conveniente utilizar la metodología KDD (Descubrimiento de Conocimiento en Bases de Datos con sus respectivas Fases: Colección de datos, preprocesamiento de los datos, selección de atributos y aplicación de algoritmos de aprendizaje computacional. La población para este estudio consta de 230788 estudiantes de educación superior del Estado de Puebla, México en base el Sistema Nacional de Información y Estadística Educativa (SNIE), para determinar la muestra se ha utilizado la técnica de selección aleatoria simple y se aplicó 26 preguntas organizadas en 9 categorías: Datos demográficos, Antecedentes Familiares, Escolaridad previa, Rendimiento académico, Apoyos Financieros, Ambiente y convivencia, Infraestructura, Seguimiento y tutorías, y Servicios; obteniéndose los siguientes resultados. Que la principal causa de deserción estudiantil es la falta de asesoría, inadecuado ambiente estudiantil, falta de seguimiento académico, deficiente calidad educativa y servicio en general; los hallazgos encontrados en este estudio permitirán a las IES implementar estrategias dirigidas para contrarrestar los índices de deserción estudiantil.

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1. Identificación del problema

Los Institutos técnicos a nivel nacional son la principal alternativa para los jóvenes que desean formarse académicamente, de acuerdo a un estudio realizado por Arellano Consultoría durante el 2018 se ha podido determinar que la demanda de este tipo de centros de estudio ha crecido en un 19%. Puesto que los Institutos o también llamadas escuelas de educación superior tecnológica brindan formación especializada con fundamentación científica y están orientadas a capacitar a los estudiantes en el dominio de ciencias aplicadas, generando capacidades en un periodo de 03 años, equivalente a 06 semestres académicos.

En la ciudad de Andahuaylas de manera similar existen diversos Institutos Tecnológicos Privados siendo la ISTEPSA una de las más antiguas y reconocidas de la provincia, esta entidad ha logrado constituirse y mantenerse en el mercado pese a las diversas dificultades que en los primeros años se presentaron, actualmente cuenta con 04 carreras profesionales; Computación e Informática, Contabilidad Computarizada, Administración de Negocios Internacionales y Administración de Empresas Turísticas y Hoteleras, teniendo a la fecha un total de 427 alumnos matriculados en el periodo 2019-II. Sin bien es cierto en los últimos años la demanda de estudiantes ingresantes ha crecido sin embargo también hay factores que demandan la necesidad de implementar acciones estratégicas, como la aparición de nuevos Institutos Tecnológicos Privados que son competencia directa sumado a ello los niveles considerables de deserción de alumnos son una constante preocupación de la gerencia que debe atenderse con prioridad. La deserción de alumnos es una intriga constante en cada inicio de semestre, de acuerdo a los datos obtenidos de la

gerencia se sabe que existe aproximadamente el 34% de deserción de alumnos en todas las carreras que ofrece el instituto.

Por lo expuesto la presente investigación tiene como objetivo resolver la siguiente problemática: ¿Cuáles son los factores y patrones que permiten segmentar los alumnos con riesgo de deserción del Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019?

Para resolver el problema se propone determinar los factores y patrones para segmentar los alumnos con riesgo deserción, para tal objetivo existe diversos tipos de técnicas en la minería de datos como, por ejemplo: CFS: Selección de Características basada en Correlación, El algoritmo de Maximización del Valor Esperado "Expectation Maximissation" (EM) y Mapas auto organizados (SOM).

2.2. Enunciados de problema

2.2.1. Problema general

¿Cuáles son los factores y patrones que permiten segmentar los alumnos con riesgo de deserción del Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019?

2.2.2. Problemas específicos

A. ¿Cuáles son los factores que afectan en la deserción de alumnos del Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019?

B. ¿Cuáles son los patrones significativos en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019?

C. ¿Cuál es el segmento de alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019?

2.3. Justificación

Es sumamente importante y urgente enfrentar la problemática actual e implementar estrategias orientadas a reducir los niveles de deserción de alumnos en

el Instituto de Educación Superior Tecnológico Privado ISTEPSA puesto que se ha identificado que en la actualidad el 34% de alumnos ingresantes desertan durante el programa formativo.

Por ello el presente trabajo de investigación en primera instancia identificará los factores que provocan la deserción de alumnos, con esta información el Director y administrador de la Institución podrán plantear acciones específicas y a medida que permitan mitigar el impacto de estos factores en los alumnos.

Además de identificar los factores de deserción es importante descubrir los patrones de deserción que presentan los alumnos es decir que entre los factores de deserción existen algunos que son más determinantes y que causan mayor impacto en los alumnos, con este conocimiento la Gerencia de la entidad podrá enfocar mayores recursos humanos y económicos en mitigar dichos factores y consecuentemente los patrones de deserción.

Identificado los factores y patrones de deserción será posible segmentar a los alumnos con riesgo de deserción es decir, del 100% de alumnos matriculados será posible diferenciar el segmento conformado por un 34% aproximadamente que tienen riesgo de deserción en un futuro cercano, entonces la Gerencia, Administración y Director de la Entidad podrán aplicar una lista de acciones establecidas en función a los factores y patrones de deserción identificados oportunamente, al segmento de alumnos con este riesgo de deserción.

Focalizar los esfuerzos permitirá optimizar el uso de recursos humanos, económicos y de tiempo, además esto generará mayor impacto en la problemática de deserción de alumnos en el Instituto, este modelo será de gran fortaleza para la institución ya que permitirá ofrecer servicios adecuados al alumnado fortaleciendo el nivel de competitividad frente a otras entidades locales y regionales que se encuentran en el mismo rubro.

2.4. Objetivos

2.4.1. Objetivo general

Determinar los factores y patrones para segmentar los alumnos con riesgo deserción del Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.

2.4.2. Objetivos específicos

A. Identificar los factores de deserción en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.

B. Establecer los patrones de deserción en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.

C. Segmentar los alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.

2.5. Hipótesis

2.5.1. Hipótesis general

Los factores y patrones permiten segmentar significativamente los alumnos con riesgo de deserción del Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.

2.5.2. Hipótesis específicas

A. El Segmento de alumnos con riesgo de deserción en el estudio corresponde al 30% del total en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1 Lugar de estudio

El lugar donde se desarrollará la presente investigación será en el Instituto de Educación Superior Tecnológico Privado ISTEPSA de la ciudad de Andahuaylas, región Apurímac.

3.2 Población

La población estará representada por todos los alumnos matriculados de los seis semestres académicos del periodo 2019 – II, en las 04 carreras profesionales que ofrece la ISTEPSA; Computación e Informática, Contabilidad Computarizada, Administración de Negocios Internacionales y Administración de Empresas Turísticas y Hoteleras; el mismo que asciende a un total de 427 alumnos.

3.3 Muestra

Por las técnicas que se utilizarán en el presente caso de investigación, no se extraerá una muestra representativa de la población, es decir se utilizará el 100% de datos de la población, el mismo que asciende a un total de 427 alumnos.

3.4 Método de investigación

Como procedimiento de la Minería de Datos Riquelme et al. (2006), definen al proceso KDD es interactivo e iterativo conteniendo los siguientes pasos:

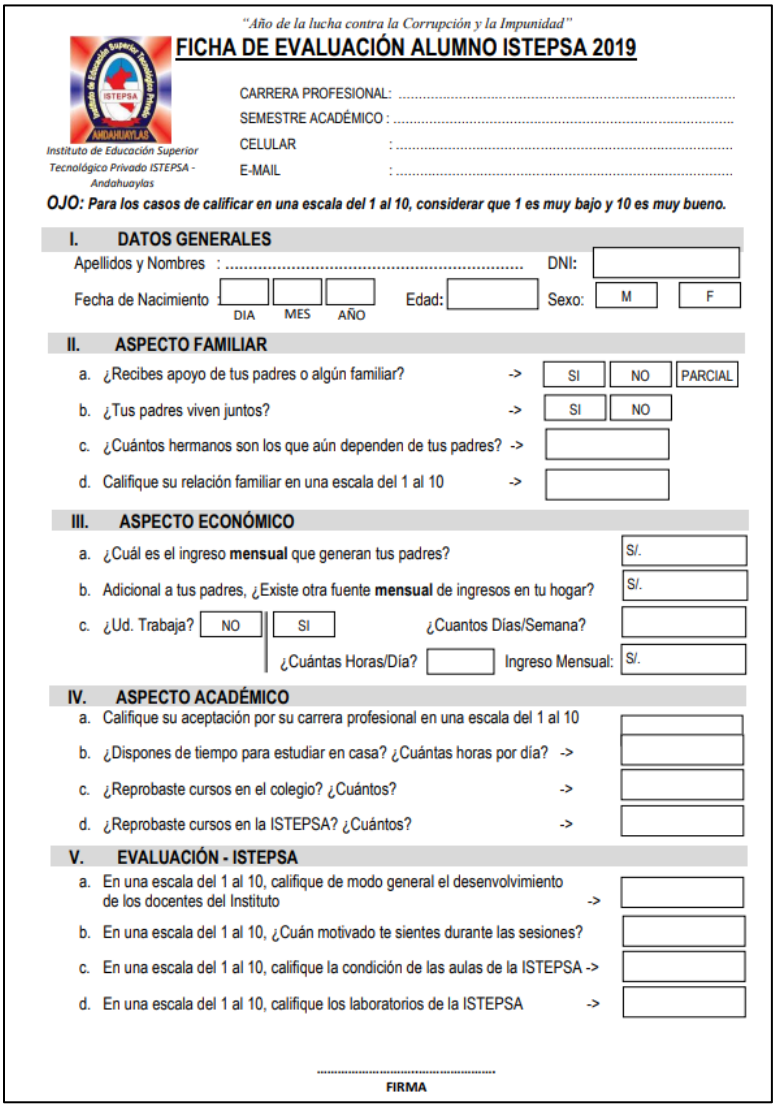
3.4.1 Comprender el dominio de aplicación

Incluye el conocimiento relevante previo y las metas de la aplicación. Se identifican en el Instituto Tecnológico Privado ISTEPSA la cantidad variable de estudiantes ingresantes en cada proceso de admisión, donde en los últimos semestres se observa que durante el proceso de formación profesional aproximadamente el 34% de alumnado deserta y esto es una problemática que debe abordarse de manera adecuada y se formula a Determinar los factores y patrones para segmentar los alumnos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.

3.4.2 Extraer la base de datos objetivo

En cuanto a la recogida, evaluación de la calidad y el análisis exploratorio de los datos se tiene: inicialmente el Instituto de Educación Superior Tecnológico Privado ISTEPSA no cuenta con algún tipo de información socioeconómica de sus alumnos, por lo que, para la investigación se requirió elaborar y aplicar una ficha de información, para validar el instrumento cuestionario de la encuesta se aplicó la prueba de la medida de adecuación de la muestra Kaiser-Meyer-Olkin (KMO) la cual indica que las variables miden factores comunes cuando el índice es mayor a 0.7, finalmente, se practicó la prueba de esfericidad de Bartlett que permite definir estadísticamente si la matriz de interrelación es una matriz de identidad, para el análisis factorial se seleccionó el método de factores principales, teniendo en cuenta que el propósito fundamental era determinar la estructura de los dominios de deserción buscando la presencia de variables latentes no observables (Hamilton, 1992). Para definir el número de factores que se debían incluir, se tuvo en cuenta el método de Kaiser (Valores propios mayores a 1) en la estructura factorial se evaluó también el método de cargas factoriales por rotaciones varimax, para determinar si ofrecían las mismas condiciones de interpretación que el método de componentes principales, el análisis de consistencia interna se llevó a cabo mediante los coeficientes alfa de Cronbach para establecer que ítem tenían una medida de homogeneidad ente 0,7 y 0,9.

Se obtuvo autorización de la Dirección para aplicar la ficha a la totalidad de los alumnos matriculados durante el semestre académico 2019-II, los datos considerados en la ficha de diagnóstico refieren a aspectos sociales, económicos y demográficos de los alumnos, con ello se busca garantizar que la información obtenida sea suficiente y adecuada para lograr los objetivos del proyecto, la aplicación se realizó a las 04 carreras profesionales existentes (Computación e informática, Contabilidad computarizada, Administración de negocios internacionales y Administración de empresas turísticas y hoteleras); a continuación, se muestra la ficha aplicada a los estudiantes:



"Año de la lucha contra la Corrupción y la Impunidad"

FICHA DE EVALUACIÓN ALUMNO ISTEPSA 2019

Instituto de Educación Superior Tecnológico Privado ISTEPSA - Andahuaylas

CARRERA PROFESIONAL: .....
SEMESTRE ACADÉMICO : .....
CELULAR : .....
E-MAIL : .....

OJO: Para los casos de calificar en una escala del 1 al 10, considerar que 1 es muy bajo y 10 es muy bueno.

I. DATOS GENERALES

Apellidos y Nombres : ..... DNI:

Fecha de Nacimiento: DIA MES AÑO Edad: Sexo: M F

II. ASPECTO FAMILIAR

a. ¿Recibes apoyo de tus padres o algún familiar? -> SI NO PARCIAL
b. ¿Tus padres viven juntos? -> SI NO
c. ¿Cuántos hermanos son los que aún dependen de tus padres? ->
d. Califique su relación familiar en una escala del 1 al 10 ->

III. ASPECTO ECONÓMICO

a. ¿Cuál es el ingreso **mensual** que generan tus padres? S/.
b. Adicional a tus padres, ¿Existe otra fuente **mensual** de ingresos en tu hogar? S/.
c. ¿Ud. Trabaja? NO SI ¿Cuantos Días/Semana?
¿Cuántas Horas/Día? Ingreso Mensual: S/.

IV. ASPECTO ACADÉMICO

a. Califique su aceptación por su carrera profesional en una escala del 1 al 10
b. ¿Dispones de tiempo para estudiar en casa? ¿Cuántas horas por día? ->
c. ¿Reprobaste cursos en el colegio? ¿Cuántos? ->
d. ¿Reprobaste cursos en la ISTEPSA? ¿Cuántos? ->

V. EVALUACIÓN - ISTEPSA

a. En una escala del 1 al 10, califique de modo general el desenvolvimiento de los docentes del Instituto ->
b. En una escala del 1 al 10, ¿Cuán motivado te sientes durante las sesiones?
c. En una escala del 1 al 10, califique la condición de las aulas de la ISTEPSA ->
d. En una escala del 1 al 10, califique los laboratorios de la ISTEPSA ->

.....
FIRMA

Figura 9. Ficha de aplicación al alumno – ISTEPSA.

3.4.3 Preparar los datos

Incluye limpieza, transformación, integración y reducción de datos. Los campos seleccionados para aplicar las técnicas de minería de datos son: de tipo cualitativos y cuantitativos, contienen información socioeconómica, académica y demográfica, siendo los campos de tipo texto (12) los siguientes: C_PROFESIONAL, S_ACADÉMICO, CELULAR, E-MAIL, NOMBRE, A_PATERNO, A_MATERNO, DNI, SEXO, APOYO_FAMILIAR, PADRES_VIVEN_JUNTOS y APOYO_FAMILIAR; mientras que los datos de tipo numérico (14) son: EDAD, HERMANOS_DEPENDEN_PADRES, RELACION FAMILIAR, ¿CUANTOS DIAS A LA SEMANA?, CUANTAS HORAS/DIA?, ACEPTACION_CARRERA_PROFESIONAL, CUANTAS_HORAS_DISPONES_PARA_ESTUDIO, CURSOS_ REPROBADOS_COLEGIO, CURSOS_REPROBADOS_ISTEPSA, DOCENTE_ISTEPSA, MOTIVADO_SESIONES, CALIFICACIÓN_AULAS_ISTEPSA y CALIFICACIÓN_LABORATORIOS_ ISTEPSA; campos de tipo moneda (03) son: INGRESO_MENSUAL_PADRES, INGRESO_ADICIONAL y INGRESO_MENSUAL; y campo de tipo fecha (01): F_NACIMIENTO, de dichos atributos no entran al análisis DNI . Además, es necesario uniformizar los atributos a partir de las técnicas de discretización a fin de contar información simbólica con características de aplicación de técnicas no supervisadas de Machine Learning.

3.4.4 Minería de datos

Como se ha señalado anteriormente, esta es la fase fundamental del proceso. Está constituido por selección de atributos, Asociación y Clustering siguiente:

B. Selección del subconjunto de atributos

La selección de subconjunto de atributos o factores (Indicadores) o simplemente características de mayor importancia que expliquen significativamente la deserción estudiantil, se realiza a partir del ranking de subconjuntos de atributos en función a la evaluación heurística obtenido de la

correlación de la clase con cada atributo considerado en el estudio, para eliminar atributos que tienen muy alta correlación los cuales son atributos redundantes, método de evaluación que determina la calidad del subconjunto de atributos para discriminar la clase se retira el estudiante. Se distinguen dos categorías, en la primera parte se utiliza métodos de evaluación *CfsSubsetEval* clasificador específico para seleccionar medir la calidad del subconjunto de atributos a través de la tasa de error resultado de un proceso completo de entrenamiento y evaluación para cada caso de búsqueda, que se encuentra implementado en Weka.

Segundo método es de búsqueda que determina la forma de realizar la búsqueda eficiente y exhaustiva de subconjuntos de atributos en base a la evaluación planteada en Weka como son: Bestfirst y otros.

C. Extracción de patrones de Asociación

Para extraer los patrones de comportamiento se utiliza el algoritmo no supervisado Apriori implementado en WEKA, por no existir relaciones conocidas a priori a contrastar la validez de los resultados, se evalúa si las reglas son estadísticamente significativas. Este algoritmo únicamente busca reglas entre atributos simbólicos, razón por la que se requiere previamente la discretización de todos los atributos numéricos.

C. Segmentación de estudiantes con riesgo de deserción

Para la segmentación de estudiantes con riesgo de abandono de sus estudios se utiliza los algoritmos de redes neuronales Maximización del Valor Esperado en inglés Espectation Maximizzation (EM) y los mapas autoorganizados de Kohonen en inglés es Self Organizing Maps (SOM) implementados en Weka bajo enfoque de redes neuronales artificiales (RNA).

3.5 Descripción detallada de métodos por objetivos específicos

3.5.1 Método para identificar los factores de deserción

El método utilizado para identificar los factores de deserción de los alumnos es el de escalamiento multidimensional, para la aplicación de este método se utilizó la herramienta WEKA, en el módulo Select attributes para

la selección de indicadores (atributos) se utilizó el algoritmo *CfsSubsetEval* con el método de búsqueda BestFirst.

Para establecer el orden de importancias de los atributos que explican la deserción de estudiantes se utilizó el evaluador de atributos ChiSquaredAttrubeEval, con el método de búsqueda Ranker.

3.5.2 Método para descubrir los patrones de deserción

Luego de haber identificado y eliminado aquellos indicadores que no son relevantes para el proceso de mineración de datos, se procedió a descubrir los patrones de deserción, para ellos se utilizó el módulo Associate del WEKA, y dentro de ella se trabajó con el algoritmo Apriori a través del algoritmo a priori para los atributos seleccionados con *CfsSubsetEval*.

3.5.3 Método para segmentar alumnos con riesgo de abandono de estudios

Para la segmentación de alumnos se utilizó técnicas de redes neuronales de aprendizaje no supervisado, los cuales son: Maximización del Valor Esperado (EM) y mapas auto-organizados de Kohonen (SOM); Ambas técnicas viene implementadas en el WEKA.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. Resultados

4.1.1. Identificar los factores o atributos de deserción de ISTEPSA, durante el periodo 2019

La calidad de los datos es uno de los factores más importantes a tomar en cuenta para aplicar los algoritmos de la Minería de Datos, si la información es irrelevante o redundante, o si los datos son ruidosos y poco confiables, entonces el descubrimiento de conocimiento durante el entrenamiento es más difícil. La selección del subconjunto de atributos (Indicadores) es el proceso de identificar y eliminar la mayor cantidad de información irrelevante y redundante posible. Los algoritmos de aprendizaje difieren en la cantidad de énfasis que ponen en la selección de atributos.

Tabla 1

Indicadores para el análisis de deserción estudiantil

N°	Indicadores	Descripción
1	C_Pro	Carrera Profesional
2	S_acad	Semestre Académico
3	Edad	Edad
4	Sexo	Genero 0:mujer; 1: varón
5	A_familia	Apoyo_familiar
6	Padres_junt	Padres_Viven_Juntos
7	Her_dep_pad	Hermanos_dependen_padres
8	Re_familia	Relacion_familiar
9	Ingreso_m_p	Ingreso_mensual_padres
10	Ingreso_a	Ingreso_adicional
11	Trabaja	
12	Tra_dia_sem	¿Cuántos días a la semana?
13	Horas_dia	¿Cuántas horas/dia?
14	Ingreso_mes	Ingreso_mensual
15	Acepta_c_p	Aceptación_carrera_profesional
16	Horas_est	Cuántas_horas_dispones_para_estudio
17	Curso_rep_co	Cursos_reprobados_colegio
18	Curso_rep_inst	Cursos_reprobados_istepsa
19	Cal_docen_inst	Calificación_docentes_istepsa
20	Motiv_sesiones	Motivado_sesiones
21	Cal_aulas_inst	Calificación_aulas_istepsa
22	Cal_lab_inst	Calificación_laboratorios_istepsa

De los 22 indicadores codificados para el análisis en la Tabla 1, no todos tienen la misma importancia en la explicación de la deserción estudiantil (valor de retira = sí). En este estudio, el uso de las estrategias de selección automática de rasgos para identificar los indicadores con mayor poder discriminante, se utilizó Weka que tiene una variedad de técnicas de selección de indicadores que tratan de explorar qué subconjuntos de indicadores son los que mejor clasifican la clase de estudiantes que se retiran. Esta selección de indicadores tiene dos componentes:

A. Un método de evaluación que determina la calidad del conjunto de indicadores para discriminar la clase. Se distingue dos categorías de métodos de evaluación, en la primera se utiliza directamente un clasificador específico para medir la calidad del subconjunto de indicadores a través de la tasa de error del clasificador. Estos métodos necesitan un proceso completo de entrenamiento y evaluación en cada caso de búsqueda, por eso resultan de un elevado coste computacional. La alternativa es la utilización de métodos que no utilizan un clasificador específico, por ejemplo, el método *CfsSubsetEval* que se encuentra implementado en *Weka* y que se basa en calcular la correlación de la clase con cada atributo, y eliminar indicadores que tienen una correlación muy alta como indicadores redundantes. Según este método los subconjuntos preferidos son aquellos altamente correlacionados con el atributo que define las clases y con poca correlación entre ellos.

B. Un método de búsqueda determina la forma de realizar la búsqueda de subconjuntos, su evaluación exhaustiva se convierte en un problema combinatorio inabordable cuando el número de indicadores es elevado. Por tanto, se necesitan estrategias de búsqueda más eficientes. Una de las estrategias más efectiva, por su rapidez, es el *BestFirst*, que se basa en elegir primero el mejor atributo, y realizar un proceso iterativo de ir añadiendo indicadores que aporten más información hasta llegar a la situación en la que añadir un nuevo atributo empeora la situación.

En la Tabla 2 se observa los subconjuntos de indicadores obtenidos por *Weka* utilizando el método de evaluación *CfsSubsetEval* y diferentes métodos de búsqueda. En todos los casos el atributo que define las clases es el atributo RETIRA=SÍ, utilizado para comprobar la deserción del estudiante de sus estudios. Como se aprecia, los subconjuntos obtenidos son iguales, y tienen una gran

similitud, en concreto en la última fila de la tabla se incluyen los indicadores seleccionados por todos los métodos de búsqueda.

Tabla 2

Subconjuntos de indicadores seleccionados con el método evaluador de atributos CfsSubsetEval

Método de Búsqueda	N° Indic.	Atributos o Indicadores
Best first	6	S_acad, Acepta_c_p, Curso_rep_co, Motiv_sesiones, Cal_aulas_inst, Cal_lab_inst
EvolutionarySearch	6	S_acad, Acepta_c_p, Curso_rep_co, Motiv_sesiones, Cal_aulas_inst, Cal_lab_inst
GreedyStepwise	6	S_acad, Acepta_c_p, Motiv_sesiones, Curso_rep_co, Cal_aulas_inst, Cal_lab_inst
LinearForwardSelection	6	S_acad, Acepta_c_p, Curso_rep_co, Motiv_sesiones, Cal_aulas_inst, Cal_lab_inst
ScatterSearchV1	6	S_acad, Acepta_c_p, Curso_rep_co, Motiv_sesiones, Cal_aulas_inst, Cal_lab_inst
TabuSearch	6	S_acad, Acepta_c_p, Motiv_sesiones, Curso_rep_co, Cal_aulas_inst, Cal_lab_inst
SubsetSizeForwardSelection	6	S_acad, Acepta_c_p, Curso_rep_co, Motiv_sesiones, Cal_aulas_inst, Cal_lab_inst

Tabla 3

Ranking de atributos según su calidad para medir la tasa de éxito.

Chi_Cuadrado	Ranking	Indicadores
38.4598	20	Motiv_sesiones
28.9913	21	Cal_aulas_inst
24.8499	22	Cal_lab_inst

20.0739	2	S_acad
17.759	15	Acepta_c_p
14.4278	19	Cal_docen_inst
10.2819	17	Curso_rep_co
5.9907	5	A_familia
2.0945	11	Trabaja
1.3149	1	C_Pro
0.544	6	Padres_junt
0.0534	4	Sexo
0	3	Edad
0	18	Curso_rep_inst
0	16	Horas_est
0	13	Horas_dia
0	12	Tra_dia_sem
0	8	Re_familia
0	14	Ingreso_mes
0	10	Ingreso_a
0	9	Ingreso_m_p
0	7	Her_dep_pad

Además del método evaluador de subconjuntos de atributos, *Weka* dispone de prorrateadores de atributos (*AtributeEval*) que no seleccionan indicadores, sino que los ordenan por relevancia de acuerdo a un ranking establecido. Utilizando el prorrateador *ChiSquaredAttributeEval* que evalúa el valor de un atributo mediante el cálculo del estadístico chi-cuadrado con respecto a la clase, los 22 indicadores de los

datos son ordenados tal como se muestra en la Tabla 3. Como, los 6 atributos que son seleccionados por todos los métodos de selección de atributos se encuentran entre los 7 primeros del ranking anterior.

4.1.2. Establecer los patrones de deserción de ISTEPSA, durante el periodo 2019

Según Tan, Steinbach y Kumar (2006), definen las reglas de asociación del algoritmo a priori, tras la discretización de los datos, puesto que este algoritmo trabaja con datos categóricos, inicialmente, se eliminaron los elementos que no tienen buen desempeño en la mineración de datos. De todos los conjuntos de reglas generadas, en la Tabla 4 se describen las más importantes. Para la interpretación se utiliza cuatro métricas conocidas Confianza (Conf), Elevación (Lift), y los indicadores de Apalancamiento (Leverage=Lev) y Convicción (Conviction-Conv). El apalancamiento mide la proporción de casos de X e Y por encima de lo esperado, si X e Y son independientes entre sí. La convicción determina el efecto del incumplimiento del consecuente de la regla.

Tabla 4

Reglas de asociación obtenidas

Reglas	Conf	Lift	Lev	Conv
1. Motiv_sesiones=deficiente Cal_aulas_inst=deficiente retira=desertor 47 ==> Cal_lab_inst=deficiente 47	1	1.96	0.05	23
2. Acepta_c_p=deficiente Motiv_sesiones=deficiente Cal_aulas_inst=deficiente 50 ==> Cal_lab_inst=deficiente 48	0.96	1.88	0.05	8.16
3. Curso_rep_co=0 Motiv_sesiones=deficiente	0.95	1.89	0.06	6.86

Cal_lab_inst=deficiente 55 ==> Cal_aulas_inst=deficiente 52				
4. Curso_rep_co=0 Cal_aulas_inst=Bueno Cal_lab_inst=bueno 55 ==> retira=0 51	0.93	1.25	0.02	2.86
5. Curso_rep_co=0 Cal_aulas_inst=bueno 81 ==> retira=0 75	0.93	1.25	0.04	3.01
6. S_acad=4 Motiv_sesiones=deficiente Cal_lab_inst=deficiente 51 ==> Cal_aulas_inst=deficiente 47	0.92	1.84	0.05	5.09
7. Motiv_sesiones=deficiente Cal_lab_inst=deficiente retira=1 51 ==> Cal_aulas_inst=deficiente 47	0.92	1.84	0.05	5.09
8. Motiv_sesiones=deficiente retira=1 56 ==> Cal_lab_inst=deficiente 51	0.91	1.78	0.05	4.57
9. Curso_rep_co=0 Motiv_sesiones=Bueno Cal_lab_inst=Bueno 55 ==> retira=0 50	0.91	1.23	0.02	2.38

De acuerdo con las reglas se identifican un papel relevante del indicador Evalúa laboratorios del instituto como deficiente, presente en seis reglas que destacan la deserción. Según el minado, 100% de los estudiantes que se retiran califican como deficiente tanto la motivación de las sesiones de clase asimismo a las aulas de la institución, entonces los laboratorios de la institución son deficientes; además 96% de los estudiantes consideran

deficiente a la carrera profesional que estudian a diferencia de la primera regla, es más 92% se retiran considerando que las aulas de la institución son deficientes; Por otro lado tres reglas señalan que el 92% de los que se retiran ratifican que la motivación que tienen es deficiente; por último, 92% de los estudiantes que se retiran son de cuarto semestre académico. También se observa que los valores de lift son superiores a 1, por lo cual se asume que los indicadores seleccionados se asocian de forma positiva, lo cual indica que la regla hacia el futuro tiene más probabilidades de que se repita.

4.1.3. Segmentación de los alumnos con riesgo de abandono de estudios de ISTEPSA, durante el periodo 2019

El análisis de clúster o la segmentación se aplica en muchas disciplinas científicas así como en el comportamiento de los estudiantes de una Institución Educativa, con algoritmos de machine Learning no supervisadas susceptibles de ser abordadas de manera más sencilla y eficiente, bajo una buena práctica conocida es el método "divide y vencerás" basado en las técnicas estadísticas extrapolables a las redes neuronales artificiales como mapas autoorganizados de Kohonen y Maximización del Valor Esperado (EM)

Con la base teórica expuesto implementada en Weka se obtiene para EM:

EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Relación: tesis_deserción

Instancias: 427

Atributos: 23 ->C_Pro, S_acad, Edad, Sexo, A_familia, Padres_junt, Her_dep_pad, Re_familia, Ingreso_m_p, Ingreso_a, Trabaja, Tra_dia_sem, Horas_dia, Ingreso_mes, Acepta_c_p, Horas_est, Curso_rep_co, Curso_rep_inst, Cal_docen_inst, Motiv_sesiones, Cal_aulas_inst, Cal_lab_inst, Ignorado=retira.

Modo de prueba: evaluación de clases a grupos en datos de entrenamiento

Modelo de agrupamiento (conjunto de entrenamiento completo) EM

Número de conglomerados seleccionados mediante validación cruzada: 3

Número de umbral iteraciones realizadas: 1

Tabla 5

Parámetros de los segmentos de instancias por Clúster a partir de EM

Clúster	0 (0.35)	1 (0.34)	2 (0.31)
	C_Pro		
1	78.2108	93.538	89.2512
2	39.9999	34.5984	22.4017
3	14.0011	16.95	5.0489
4	20.9999	5.4343	18.5658
[total]	153.2117	150.5207	135.2676
	S_acad		
1	16.1469	14.652	11.2012
2	36.9998	39.6483	29.3519
3	6.9999	15.0156	9.9845
4	64.727	58.4952	50.7779
5	22.3383	23.6559	13.0058
6	7.9998	1.0538	22.9464
[total]	155.2117	152.5207	137.2676
	Edad		
mean	21.416	22.3703	22.0219
std. dev.	2.7291	3.536	2.6583

Sexo			
0	95.065	87.1718	74.7631
1	56.1467	61.3488	58.5045
[total]	151.2117	148.5207	133.2676
A_familia			
1	116.2108	46.3567	88.4325
2	19.001	64.3977	26.6013
3	16.9999	38.7662	19.2338
[total]	152.2117	149.5207	134.2676
Padres_junt			
1	110.727	83.8897	97.3833
2	40.4847	64.6309	35.8843
[total]	151.2117	148.5207	133.2676
Her_dep_pad			
mean	2.3948	2.5154	2.3701
std. Dev.	1.5139	1.653	1.4409
Re_familia			
mean	3.0336	2.1829	3.0816
std. dev.	1.065	1.2712	0.9527
Ingreso_m_p			
mean	891.6879	798.5939	1231.7679
std. dev.	1127.6181	1183.9844	924.1328
Ingreso_a			

mean	37.6714	28.0314	88.4209
std. dev.	144.5044	90.1022	209.5509
	Trabaja		
0	150.0635	3.3966	12.5399
1	1.1483	145.124	120.7277
[total]	151.2117	148.5207	133.2676
	Tra_dia_sem		
mean	0.0107	4.8292	3.8533
std. dev.	0.143	1.808	1.9804
	Horas_dia		
mean	0.0234	6.0899	6.2102
std. dev.	0.3051	1.6973	1.802
	Ingreso_mes		
mean	1.3712	519.1776	336.1846
std. dev.	16.4235	413.64	242.2249
	Acepta_c_p		
mean	3.0091	2.1886	3.3925
std. dev.	1.0367	1.1597	0.5512
	Horas_est		
mean	1.4209	0.8167	1.206
std. dev.	0.5495	0.5583	0.5189
	Curso_rep_co		
mean	0.5769	1.3695	0.4971
std. dev.	1.0937	1.5957	0.7066

Curso_rep_inst			
mean	1.0211	1.2181	1.0145
std. dev.	1.2966	1.2359	1.2197
Cal_docen_inst			
1	38.0013	75.7455	8.2533
2	15.3383	14.9837	6.6779
3	76.8723	43.7352	92.3925
4	22.9998	16.0563	27.9439
[total]	153.2117	150.5207	135.2676
Motiv_sesiones			
1	43.1475	85.3837	6.4689
2	15.0007	22.1743	5.825
3	67.7254	34.4295	93.8451
4	27.3382	8.5332	29.1286
[total]	153.2117	150.5207	135.2676
Cal_aulas_inst			
1	71.4848	117.7449	27.7703
2	15	12.9828	16.0172
3	52.726	15.0287	73.2454
4	14.001	4.7643	18.2347
[total]	153.2117	150.5207	135.2676
Cal_lab_inst			
1	68.4848	114.3936	38.1215
2	10	9.399	18.601

3	54.7254	21.8815	58.3932
4	20.0016	4.8466	20.1519
[total]	153.2117	150.5207	135.2676

Tiempo necesario para crear el modelo (datos de entrenamiento completos): 1.09 segundos

Modelo y evaluación del conjunto de entrenamiento

Tabla 6

Instancias agrupadas

Clúster	Instancias	Asignado
0	147 (34%)	Sin clase
1	148 (35%)	Se retiran
2	132 (31%)	No se retiran

Probabilidad de registro: -44.11639, considerando el atributo de clase: se retira, representado por el clúster: segmento de los estudiantes que se retiran y el clúster 2 es el segmento de los estudiantes que no se retiran.

Tabla 7

Clase se retira asignado al clúster

0	1	2	Asignado a Clúster
113	90	113	No se retira
34	58	19	Se retira

Instancias agrupadas incorrectamente: 256 (59.9532 %)

4.1.4. Segmentación con Mapas Autoorganizados (SOM) de Kohonen

Estas redes neuronales artificiales (RNA) llamada redes de Kohonen permitieron clasificar la información y reducir el número de segmentos visualizando la información en mapas bidimensionales que preservan y reflejan la estructura de similitud entre la información entrante respecto a una clase “se retira”. Los resultados obtenidos en Weka se ilustran a continuación:

Esquema: weka. clusterers. SelfOrganizingMap -L 1.0 -O 2000 -C 1000 -H 2 -W 2

Relación: tesis_deserción

Instancias: 427

Atributos: 23, de manera similar que EM

Modo de prueba: evaluación de clases a grupos en datos de entrenamiento

Modelo de agrupamiento (conjunto de entrenamiento completo) SOM

Número de conglomerados seleccionados mediante validación cruzada: 4

Número de iteraciones realizadas: 1

Parámetros de los segmentos de instancias por Clúster

Tabla 8

Parámetros de los segmentos de instancias por Clúster a partir de SOM

Clúster **Atributo**	**0** **(110)**	**1** **(48)**	**2** **(186)**	**3** **(83)**
		C_pro		
value	0.8	0.86	0.5754	0.3768
min	0	0	0	0
max	3	3	3	3
mean	0.8364	0.875	0.6398	0.3976

std. dev.	1.0538	1.0842	0.9778	0.7641
		S_acad		
value	2.6538	2.2807	3.0669	1.9736
min	0	0	0	0
max	5	4	5	5
mean	2.4818	2.2292	2.6935	1.9518
std. dev.	1.3994	1.3247	1.3979	1.3057
		Edad		
value	21.4466	21.8546	22.2085	21.791
min	17	18	17	18
max	34	32	34	39
mean	21.3364	21.7292	22.3118	21.9759
std. dev.	2.6311	3.0579	2.9591	3.5919
		Sexo		
value	0.3337	0.4027	0.4216	0.4373
min	0	0	0	0
max	1	1	1	1
mean	0.3727	0.4167	0.414	0.4217
std. dev.	0.4857	0.4982	0.4939	0.4968
		A_familia		
value	0.3089	0.3808	0.4819	0.8908
min	0	0	0	0
max	2	2	2	2
mean	0.3273	0.375	0.6398	0.9398

std. dev.	0.6788	0.6058	0.7811	0.7547
		Padres_junt		
value	0.2341	0.3222	0.2153	0.4748
min	0	0	0	0
max	1	1	1	1
mean	0.2545	0.3333	0.2796	0.506
std. dev.	0.4376	0.4764	0.45	0.503
		Her_dep_pad		
value	2.6443	1.666	2.4324	2.8391
min	0	0	0	0
max	5	5	5	5
mean	2.6091	1.6667	2.3602	2.7831
std. dev.	1.515	1.3422	1.5332	1.5777
		Re_familia		
value	3.5313	1.5082	3.3235	1.177
min	2	1	1	1
max	4	4	4	3
mean	3.5636	1.5	3.328	1.1325
std. dev.	0.5163	0.7989	0.7241	0.4354
		Ingreso_m_p		
value	1074.5017	803.5596	1317.7137	951.3236
min	0	0	0	0
max	3500	10000	10000	8200
mean	895.2727	803.7917	1058.4409	937.5904

std. dev.	884.7533	1512.153	1149.8311	991.2675
		Ingreso_a		
value	41.6345	21.7229	66.572	20.1409
min	0	0	0	0
max	1500	500	1000	500
mean	40.7364	23.75	74.6452	22.0482
std. dev.	160.7083	80.2556	188.448	73.9604
		Trabaja		
value	0.0079	0.1378	0.8988	0.9835
min	0	0	0	0
max	1	1	1	1
mean	0.0091	0.1458	0.9355	0.988
std. dev.	0.0953	0.3567	0.2463	0.1098
		Tra_dia_sem		
value	0.0079	0.4931	4.0251	4.8504
min	0	0	1	1
max	1	5	7	7
mean	0.0091	0.5	4.2366	4.8434
std. dev.	0.0953	1.2377	2.0048	1.7425
		Horas_dia		
value	0.0317	0.5928	6.1218	6.5767
min	0	0	1	3
max	4	6	8	8
mean	0.0364	0.625	6.129	6.4699

std. dev.	0.3814	1.5246	1.7872	1.4085
		Ingreso_mes		
value	0.1587	43.7443	375.7966	538.6583
min	0	0	0	0
max	20	600	3000	2500
mean	0.1818	43.5417	398.3871	532.4699
std. dev.	1.9069	115.9143	358.7088	341.4713
		Acepta_c_p		
value	3.3379	2.2234	3.1316	1.8239
min	1	1	1	1
max	4	4	4	4
mean	3.3545	2.2292	3.1613	1.8193
std. dev.	0.7491	1.1713	0.8919	0.9771
		Horas_est		
value	1.4872	1.3017	1.1794	0.5739
min	0	0	0	0
max	2	3	3	2
mean	1.4545	1.3125	1.1559	0.6265
std. dev.	0.5182	0.6242	0.4792	0.599
		Curso_rep_co		
value	0.508	0.7019	0.5361	2.2134
min	0	0	0	0
max	4	4	4	4
mean	0.5182	0.6875	0.4946	2.0482

std. dev.	1.0292	1.2404	0.859	1.5765
		Curso_rep_inst		
value	1.0263	0.9906	1.0483	1.2991
min	0	0	0	0
max	4	4	4	4
mean	1.0182	0.9792	1.0753	1.2651
std. dev.	1.3544	1.1938	1.2541	1.1696
		Cal_docen_inst		
value	1.7759	1.1107	1.7958	0.9839
min	0	0	0	0
max	3	3	3	3
mean	1.7273	1.1458	1.7151	0.988
std. dev.	0.9473	1.1107	0.9751	1.0987
		Motiv_sesiones		
value	1.746	1.0598	1.8127	0.651
min	0	0	0	0
max	3	3	3	3
mean	1.6818	1.0417	1.7151	0.6988
std. dev.	1.0574	1.051	0.9639	0.9466
		Cal_aulas_inst		
value	1.323	0.7611	1.3167	0.5806
min	0	0	0	0
max	3	3	3	3
mean	1.2	0.7292	1.1129	0.5422

std. dev.	1.0904	1.0051	1.0921	0.8738
		Cal_lab_inst		
value	1.3905	0.9144	1.2457	0.5393
min	0	0	0	0
max	3	3	3	2
mean	1.2636	0.875	1.086	0.506
std. dev.	1.1226	1.1783	1.1117	0.8319

Tiempo necesario para crear el modelo (datos de entrenamiento completos): 4,98 segundos.

Modelo y evaluación del conjunto de entrenamiento - Instancias agrupadas.

Tabla 9

Instancias agrupadas por Clúster a partir de SOM

Clúster	Instancias	Asignado
0	110 (26%)	Sin clase
1	48 (11%)	Sin clase
2	186 (44%)	No se retiran
3	83 (19%)	Se retiran

Tabla 10

Clase se retira asignado al clúster

0	1	2	3	Asignado a Clúster
87	35	145	49	No se retira
23	13	41	34	Se retira

Instancias agrupadas incorrectamente: 248 (58.0796 %)

4.2. Discusión

Identificación de los factores o atributos que explican significativamente la deserción estudiantil de ISTEPSA, durante el periodo 2019 son seis, de acuerdo al orden son: Motiv_sesiones, Cal_aulas_inst, Cal_lab_inst Acepta_c_p, Curso_rep_co y S_acad, todos están asociados a los atributos académicos.

En el contexto de "motivación de sesiones" como el atributo de mayor importancia, implica que los estudiantes deben estar psicológicamente dispuestos a estudiar hasta la culminación de sus estudios, junto a las condiciones de infraestructura física como son las "aulas de la institución" sean adecuadas para el desarrollo académico, sin perder de vista que la formación profesional integral sea más práctico en base "laboratorios y talleres" de lo contrario recibirían una formación meramente teórico, por su puesto suman los perfiles profesionales adecuados que definen las carreras preferidas por los estudiantes, No obstante, también juega un rol el semestre académico en que cursan, ya que las admisiones son semestrales y generalmente en el segundo semestre del año lectivo los ingresos son muy reducidos. Ésta realidad, frente a otros trabajos como sostienen Pérez et al., (2018), sobre el modelo de predicción desarrollado con técnicas y métodos de Minería de Datos (DM) para identificar las causas de deserción estudiantil, consideran para la selección del subconjunto de atributos el método evaluador CfsSubsetEval y de búsqueda BestFirst encontrando indicadores de mayor influencia en la deserción y reprobación escolar de la Institución de Educación Superior (IES) del estado de México con un 66% de representación y un margen de error del 47% se logró una aproximación satisfactoria para abordar el fenómeno de la deserción o la reprobación estudiantil, similar a los resultados del presente estudio. Asimismo, según Eckert & Suénaga (2015) en el estudio de "Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos" aplican el método de evaluación CfsSubsetEval y el de búsqueda BestFirst, los que ofrecen una selección de subconjuntos de atributos de mayor calidad, se probó otras alternativas de algoritmos para cada método, para efectos prácticos, no se han encontrado variaciones significativas en los resultados finales.

Timaran & Jiménez (2014) obtiene en su trabajo que, el 100% de los estudiantes que desertan son solteros, su promedio de notas es menor que 2.4, han perdido materias en los primeros semestres (1 a 4) y todas las materias las han perdido una sola vez. El 16.1% del total de estudiantes (2.136) que ingresaron a la Universidad de Nariño y la Institución Universitaria CESMAG entre los años 2004 y 2006 cumplen con este patrón.

El método Expectation Maximization (EM) como método de agrupación bajo enfoque de Machine Learning, refina en forma iterativa un modelo de clústeres inicial para ajustar los datos y determina la probabilidad de que un punto de datos exista en un clúster. El algoritmo finaliza el proceso cuando el modelo probabilístico ajusta los datos, siendo el ajuste el logaritmo de la probabilidad de los datos dado el modelo. Este algoritmo se utiliza como algoritmo predeterminado porque proporciona numerosas ventajas comparado con la agrupación en clústeres K-mediana, entonces EM es escalable y no escalable creando clústeres más precisos. Por lo que, se justifica la aplicación de EM en el presente estudio.

Villamarín (2017), aplica Mapas Auto Organizados de Kohonen, para el análisis de la deserción estudiantil, reportando casos de éxito que coadyuva en la detección temprana de los posibles casos de deserción estudiantil en la Fundación Centro Colombiano de Estudios Profesionales (FCECEP), concretamente en las tecnologías en ingeniería las cuales, en los últimos años, han sido fuertemente golpeadas por la disminución de la población estudiantil. Los resultados fueron realizar reuniones (al menos una por semestre) con los padres de los estudiantes para que ellos entiendan y conozcan las actividades de apoyo a la permanencia, de bienestar, de índole académico que tienen sus hijos y como la institución invierte en su acompañamiento, previa visualización, identificación de los atributos en el mapa y su análisis basado en las colecciones de datos previamente identificados. Se procesaron los datos en el mapa, se entrenó el mapa, se validó el mapa, se logró generar la suficiente cantidad de mapas para determinar las variables que afectan la deserción estudiantil en la FCECEP.

En su trabajo Hernández (2011), sobre “Descubrimiento de conocimiento en la base de datos académica de una institución de educación superior usando redes neuronales”, utiliza mapas autoorganizativos de Kohonen. Según el Ministerio de

Educación Nacional (MEN) de Colombia, el riesgo de deserción en los estudiantes que asisten a instituciones públicas es de un 54% menor que en los que asisten a instituciones privadas, se desprende:

A. El abandono voluntario ocurre durante los primeros meses posteriores al ingreso a la institución;

B. Cinco de cada diez estudiantes desertan al inicio del segundo año;

C. Cuatro de cada diez estudiantes que comienzan el cuarto año, no obtienen el título profesional correspondiente; y

D. El mayor abandono se da en carreras con baja demanda.

Donde los estudiantes manifiestan serias dificultades para integrarse al medio académico y social de la Institución. Además, los atributos explicativos son: la edad, la madurez intelectual del estudiante, así como la falta de conocimientos y habilidades previas necesarias para realizar estudios superiores.

Díaz (2008), propone el modelo conceptual de deserción/permanencia que permita proveer a administradores de la educación superior el marco para construir un plan de retención de estudiantes incorporando las necesidades individuales de sus estudiantes. Realizar seguimiento y evaluación permanente de las variables que afectan la integración social y académica, como estrategias de intervención focalizadas para disminuir la deserción estudiantil. En el marco de la motivación (positiva o negativa), la que es afectada por la integración académica y social. A su vez, éstas están compuestas por las principales características preuniversitarias, institucionales, familiares, individuales y las expectativas laborales.

CONCLUSIONES

Mediante el método de evaluación *CfsSubsetEval* y el método de búsqueda *Best first* entre otros de Machine Learning se ha identificado los siguientes factores de deserción para los estudiantes de ISTEPSA durante el periodo 2019: Motivación de sesiones de aprendizaje, Calificación de laboratorios de la Institución, Calificación de las aulas de la Institución, Acepta a su carrera profesional, Cursos reprobados en el colegio y Semestre Académico, son atributos que influyen directamente en la deserción estudiantil del Instituto de Educación Superior Tecnológico Privado ISTEPSA.

Se logró establecer los patrones de deserción en ISTEPSA durante el periodo 2019, donde el 100% de los estudiantes que se retiran califican como deficiente tanto la motivación de las sesiones de clase asimismo a las aulas de la institución, entonces los laboratorios de la institución son deficientes; además 96% de los estudiantes consideran deficiente a la carrera profesional que estudian a diferencia de la primera regla, es más 92% se retiran considerando que las aulas de la institución son deficientes; Por otro lado tres reglas señalan que el 92% de los que se retiran ratifican que la motivación que tienen es deficiente; por último, 92% de los estudiantes que se retiran son de Cuarto semestre académico. También se observa que los valores de lift son superiores a 1, por lo cual se asume que los indicadores seleccionados se asocian de forma positiva, lo cual indica que la regla hacia el futuro tiene más probabilidades de que se repita.

Para la segmentación de los estudiantes con riesgo de abandono resultan en los algoritmos EM y SOM, el clúster 3 y 1 representa el 35% y 19% de los estudiantes de la Institución Educativa respectivamente, donde se observa que los factores académicos son determinantes para la deserción de alumnos.

RECOMENDACIONES

A. Se recomienda a los Promotores, los Directivos y Docentes de la IE tomar especial interés en las actividades académicas a fin de evitar el fenómeno de la deserción estudiantil.

B. Se recomienda a la Institución, convocar a profesionales en psicología de la educación y afines con el fin de tener en cuenta el factor Académico potencial determinante en la deserción estudiantil.

C. Se recomienda implementar adecuada infraestructura y laboratorios para cada carrera profesional.

D. Se recomienda analizar nuevas variables endógenas con el fin de determinar más patrones de comportamiento.

E. Se recomienda para próximos estudios incluir variables como cuarentena y COVID-19, para identificar al alumnado con riesgo de deserción.

BIBLIOGRAFÍA

Aranciaga, J. & Ccanto, E. (2021), Factores asociados a la deserción de estudiantes en un instituto de educación superior privado de Lima, Recuperado de: https://repositorio.unife.edu.pe/repositorio/bitstream/handle/20.500.11955/870/Aranciaga%20Valladares%2c%20J_Ccanto%20Sayajo%2c%20ER_2021.pdf?sequence=1&isAllowed=y.

Baviera, T. (2016), Técnicas para el análisis del sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength. Revista Dígitos. Recuperado de: https://revistadigitos.com/index.php/digitos/article/view/ 74/39.

Belamate, D., Cassani, M. & Ricci, C. (2016). Aplicación de reglas de asociación para la detección de patrones de comportamiento en sistema académico universitario. Universidad Tecnológica Nacional. Argentina. Recuperado de: http:// cytal.frvm.utn.edu.ar/q/tf/7/62

Beron, E., Mejía, D., Castrillón O. (2020). Principales causas de ausentismo laboral: una aplicación desde la minería de datos. Recuperado de: https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-07642021000200011.

Castillo, P. (2017). Aplicación de Aprendizaje Automático para la Predicción de Clientes Potenciales en Procesos de Mercadotecnia (Tesis de posgrado). Centro de Investigación en Matemáticas, A.C., Guanajuato, México.

Cestero, E., & Caballero, A. (2018), Data Science y Redes Complejas (1 Ed.) Madrid, España: Editorial Universitaria Ramón Areces S.A.

Cifuentes, F. (2016). Clasificación Automática de Tweets Utilizando K-NN y K-Means como Algoritmos de Clasificación Automática, Aplicando TF-IDF y TF-RFL para las Ponderaciones (Tesis de pregrado). Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

De la Cruz, K. (2017). Segmentación de clientes con Inteligencia Analítica para personalizar las Ventas de los Servicios de las Agencias Turísticas (Tesis de posgrado). Universidad Peruana Unión, Lima, Perú.

Díaz P., Ch. (2008). Modelo conceptual para la deserción estudiantil universitaria Chilena, Universidad Católica de la Santísima Concepción – Chile.

Eckert, K. B., & Suénaga, R. (2015). Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos. Formación universitaria, 8(5), 03-12.

Gil, C. (2018). Análisis de componentes principales (PCA). Recuperado de: https://rpubs.com/Cristina_Gil/PCA.

Hall, M. A., y Smith, L. A. (1998). Practical feature subset selection for machine learning. Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Hamilton LC. (1992). Regression With GRAPHICS. A second course in applied statistics. Belmont, Duxbury.

Hernández C., J. (2011). Descubrimiento de conocimiento en la base de datos académica de una institución de educación superior usando redes neuronales. Universidad Santo Tomás, Bucaramanga, Colombia.

Himansu, S., Janmenjoy, N., Bighnaraj N. y Ajith A. (2018), Computational Intelligence in Data Mining (1 Ed.), Singapur: Editorial Springer.

Holgado, L. (2018), Detección de Patrones de Bajo Rendimiento Académico Mediante Técnicas de Minería De Datos de los Estudiantes de la Universidad Nacional Amazónica de Madre de Dios 2018. Recuperado de: http://repositorio.unap.edu.pe/handle/UNAP/9815.

Hoyos J. G. & Aponte F. A. (2019), Caracterización de los estudiantes de una Institución de Educación Superior Mediante Big Data. Recuperado de: https://www.redalyc.org/journal/852/85263724001/85263724001.pdf.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. Neural computation, 6(2), 181-214.

Kira, K., Renedell, L. (1992). A practical approach to feature selection. Proceedings of the Ninth International Conference on Machine Learning. Aberdeen Scotland. Morgan Kaufmann. pp. 249–256.

Koller & Sahami (1996), Toward Optimal Feature Selection. Recuperado de: http://ilpubs.stanford.edu:8090/208/1/1996-77.pdf.

Ley Peruana N° 30512 (2016), Ley de Institutos y Escuelas de Educación Superior y de la Carrera Pública de sus Docentes, Recuperado el 28 de Abril del 2019 de, https://www.gob.pe/institucion/minedu/normas-legales/118500-30512.

Linares, A. (2019). Predicción de Renuncia de Socios de una Cooperativa Utilizando Técnicas Supervisadas de Aprendizaje Automático, Recuperado de: http://tesis.ucsm.edu.pe/repositorio/bitstream/handle/UCSM/9742/71.0633.IS.pdf?sequence=1&isAllowed=y.

Mathivet, V. (2018). Inteligencia Artificial para Desarrolladores (2 Ed.) Barcelona, España: Editorial ENI.

Miranda, M. & Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos, Recuperado de: https://www.redalyc.org/articulo.oa?id=373551306007.

Moerland, P. (1997). Some methods for training mixtures of experts. Informe técnico, Dalle Molle Institute for Perceptive Artificial Intelligence.

Mollo, N. (2018). Análisis Predictivo de la Deserción Estudiantil Utilizando Data Warehouse y Minería de Datos en la Universidad Nacional Jorge Basadre Grohmann – Tacna, 2012-2018. Recuperado de: http://repositorio.unjbg.edu.pe/handle/UNJBG/3506.

Ochoa, L. (2016). Estudio Comparativo de Técnicas no Supervisadas de Minería de Datos Para Segmentación de Alumnos. (Tesis de Pregrado). Universidad Católica de Santa María, Arequipa, Perú.

Pacco, R. (2015). Análisis Predictivo Basado en Redes Neuronales no Supervisadas Aplicando Algoritmo de K-Medias y CRISP-DM para Pronóstico de Riesgo de Morosidad de los Alumnos en la Universidad Nacional Peruana Unión. (Tesis de Posgrado). Universidad Peruana Unión, Lima, Perú.

Pavón, F. (2016). Generación de Conocimiento Basado en Aprendizaje Automático y Aplicación en Diferentes Sectores. (Tesis de Posgrado). Escuela Técnica

Superior de Ingeniería Informática (ETSI) Universidad Nacional de Educación a Distancia (UNED), Madrid, España.

Pérez G. (2020), Comparación de Técnicas de Minería de Datos Para Identificar Indicios de Deserción Estudiantil, a Partir del Desempeño Académico, Recuperado de: https://www.redalyc.org/journal/5537/553768131019/553768131019.pdf.

Pérez, M., Norma, P., Aguilar, C., Jorge, R., Zamora, R., Rosa, A., & Miguel, J. (2018). Diseño de un modelo predictivo aplicando minería de datos para identificar causas de deserción estudiantil universitaria. México.

Quezada, N. (2017). K-vecinos más Próximos en una Aplicación de Clasificación y Predicción en el Poder Judicial del Perú. (Tesis de Posgrado). Universidad Nacional Mayor de San Marcos, Lima, Perú.

Redondo, M. (2016), Simulación de Redes Neuronales como Herramienta Big Data en el Ámbito Sanitario. Recuperado de: https://books.google.com.pe/books?id=9vSBDgAAQBAJ&pg=PA17&dq=redes+neuronales&hl=es&sa=X&ved=0ahUKEwiF486J26_hAhVCxVkKHYVpC6gQ6AEIKDAA#v=onepage&q=redes%20neuronales&f=false.

Riquelme S., J. C., Ruiz, R., y Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, 10 (29), 11-18.

Rivera, M. (2016), Los Factores determinantes y su relación con la deserción escolar en los alumnos del primero al sexto grado del nivel primaria de la x, de Monzón, 2010 al 2015. Recuperado de: https://renati.sunedu.gob.pe/handle/sunedu/1799018.

Sancho, Q. (2000). Sistemas Modulares, Mezcla de Expertos y Sistemas Híbridos. Informe Técnico DI-2000-001 Departamento de Informática, Universidad de Valladolid, España.

Tan, Steinbach & Kumar (2006), Introducción a la minería de datos. Recuperado de: http://didawiki.di.unipi.it/lib/exe/fetch.php/dm/2.2018-dm-introduction.pdf

Terrones, A. (2018), Inteligencia Artificial y Ética de la Responsabilidad. Cuestiones de Filosofía, 4(22), 141-170. Páginas.

Timaran, R., Jiménez, J. (2014). Detección de patrones de deserción estudiantil en programas de pregrado de instituciones de educación superior con CRISP-DM. Congreso Iberoamericano de Ciencia, Tecnología, Innovación y Educación.

Torres, M. (2018). Segmentación demográfica y relación con los clientes en la empresa Hotel Cielo, Distrito de Tarapoto, 2018. Recuperado de: https://repositorio.ucv.edu.pe/handle/20.500.12692/51256.

Urbina, A.B., Camino, J.C. & Cruz, R. (2020). Deserción Escolar Universitaria: Patrones Para Prevenirla Aplicando Minería de Datos Educativa. Recuperado de: https://www.redalyc.org/journal/916/91664838013/91664838013.pdf.

Villamarín V., J. H. (2017). Análisis de la deserción estudiantil en la FCECEP utilizando Machine Learning específicamente Mapas Auto Organizados de Kohonen. Universidad Autónoma de Occidente Posgrado de la Facultad de Ingeniería-Santiago de Cali, Colombia.

WEKA3 (2019), WEKA3, Recuperado de: https://www.cs.waikato.ac.nz/~ml/weka/

Yamao, E. (2018). Predicción del Rendimiento Académico Mediante Minería de Datos en Estudiantes del Primer Ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres, Lima-Perú. Recuperado de: https://repositorio.usmp.edu.pe/handle/20.500.12727/3555.

Zavala, J. (2017). Pronóstico de la Exportación Pesquera por Redes Neuronales y Modelos Arima (Tesis de pregrado). Universidad Nacional de Trujillo, Trujillo, Perú.

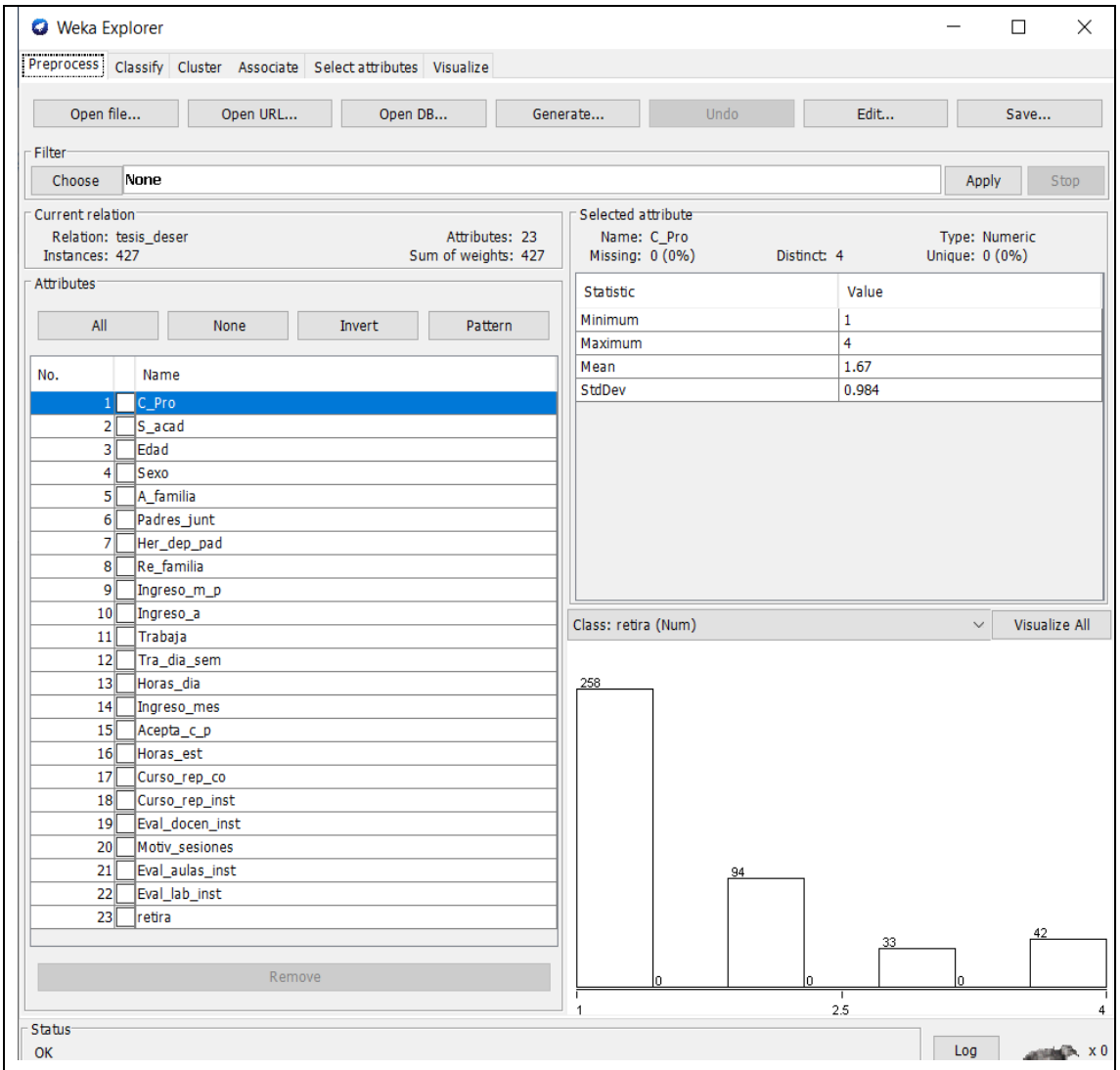
ANEXOS

Anexo 1: Matriz de consistencia

PROBLEMA	OBJETIVOS	HIPOTESIS	VARIABLES	INDICADORES	INDICES	MÉTODO
PROBLEMA PRINCIPAL ¿Cuáles son los factores y patrones que permiten segmentar los alumnos con riesgo de deserción del Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019? **PROBLEMAS ESPECÍFICOS** a. ¿Cuáles son los factores que afectan en la deserción de alumnos del Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019?	**OBJETIVO GENERAL** Determinar los factores y patrones para segmentar los alumnos con riesgo deserción del Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019. **OBJETIVOS ESPECIFICOS** a. Identificar los factores de deserción en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019. b. Establecer los patrones de deserción en el Instituto Superior Tecnológico	**HIPOTESIS GENERAL** Los factores y patrones permiten segmentar significativamente los alumnos con riesgo de deserción del Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019. **HIPOTESIS ESPECIFICAS** a. El Segmento de los alumnos con riesgo de deserción al estudio corresponde al 34% del total en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.	**VARIABLE 1** Factores y patrones de deserción.	Procedencia	Ciudad	**POBLACIÓN** La población es igual a total de alumnos matriculados en las 04 facultades durante el periodo académico 2019 – II, el cual asciende a un total de 427 alumnos. **MUESTRA** La muestra será igual a la población, el cual asciende a un total de 427 alumnos. **NIVEL DE INVESTIGACIÓN** Estudio Explicativo. **DISEÑO DE INVESTIGACIÓN** Correlacional.
				Sexo	Género M/F	
				edad	Número (1-100)	
				Estado Civil	Casado/Soltero	
				Número de Hijos	Número (0-10)	
				Vive con sus padres	SI/NO	
				Percepción relación Familiar	Número (0-10)	
				Apoyo Económico Familiar mensual	Moneda (0-3000)	
				Días de trabajo/semanal	Número (0-7)	

b. ¿Cuáles son los patrones significativos en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019? c. ¿Cuál es el segmento de alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019?	Privado ISTEPSA, durante el periodo 2019. c. Segmentar los alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.			Horas trabajo/día	Número (0-15)	**METODOLOGÍA PARA LA SEGMENTACIÓN DE ALUMNOS CON RIESGO DE DESERCIÓN** Mediante la aplicación técnicas de minería de datos.
				Ingreso mensual propio	Moneda (0-3000)	
				Aceptación por la carrera escogida	Número (0-10)	
				Cursos reprobados en el colegio	Número (0-10)	
			VARIABLE 2 Segmentación de alumnos con riesgo de deserción.	Tamaño del segmento de alumnos con riesgo de deserción.	Porcentaje (0-100)	

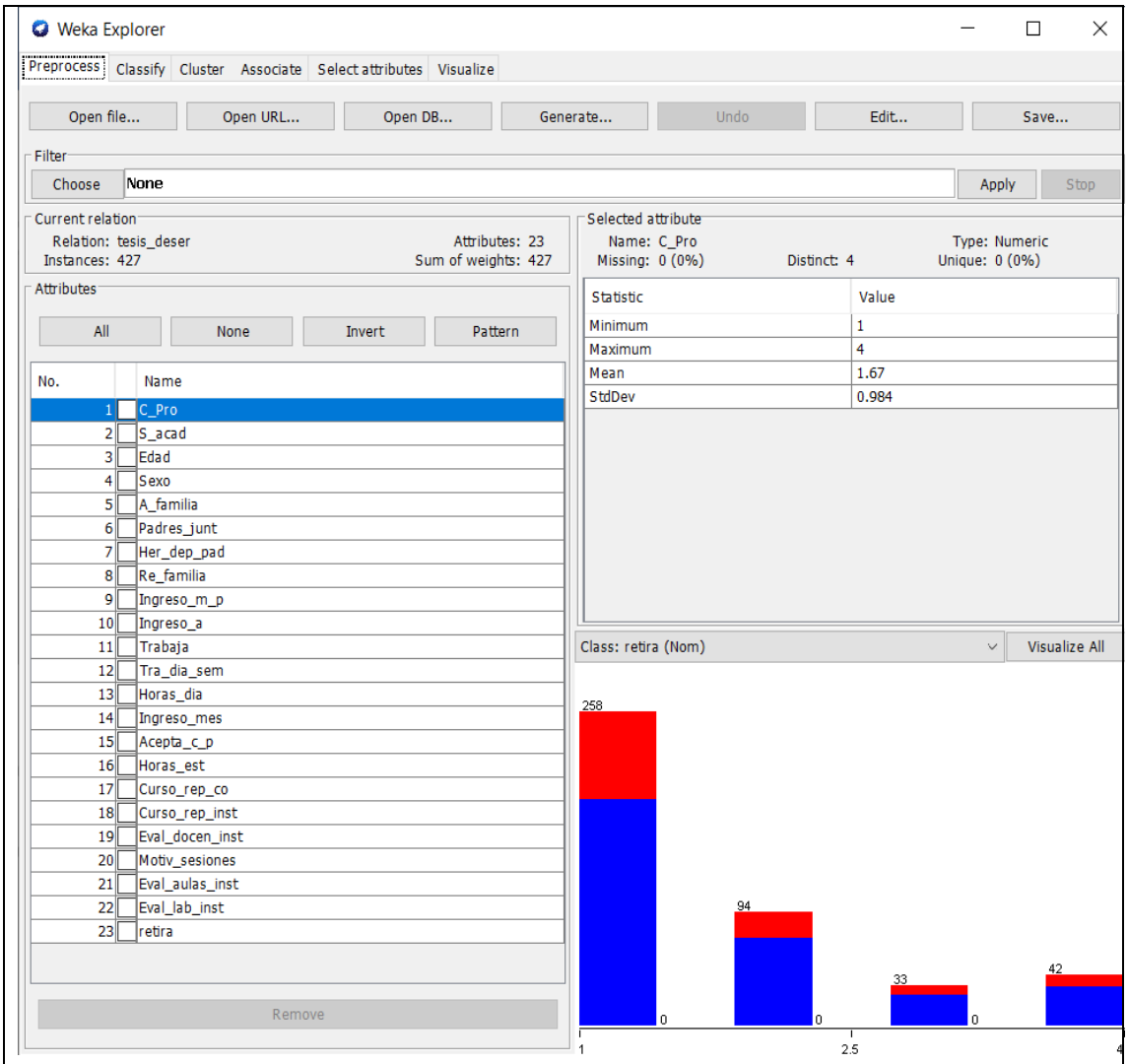
Anexo 2: Proyección del total de atributos recogidos



DESCRIPCIÓN: Proyección de todos los atributos recogidos de los alumnos del Instituto de Educación Superior Tecnológico Privado ISTEPSA.

FECHA: Octubre 2020.

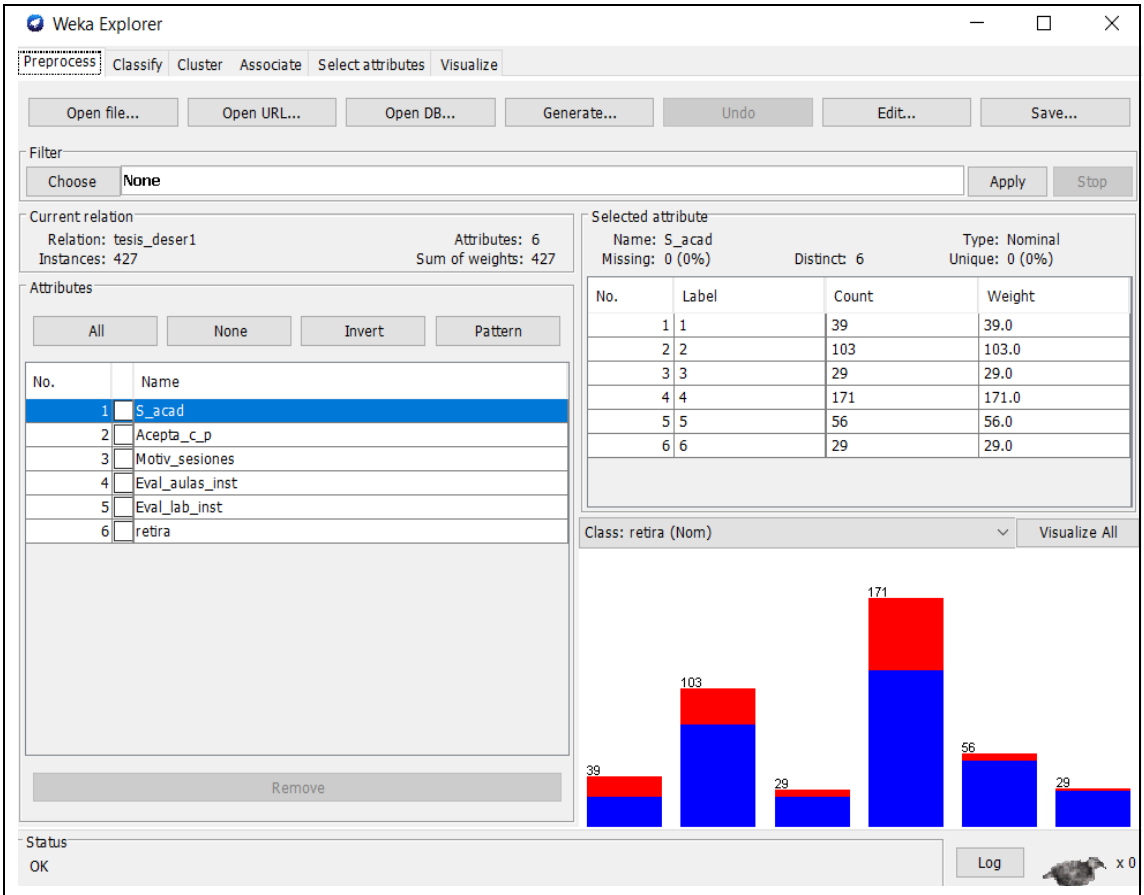
Anexo 3: Proyección de atributos codificados



DESCRIPCIÓN: Proyección los atributos codificados de los alumnos del Instituto de Educación Superior Tecnológico Privado ISTEPSA.

FECHA: Octubre 2020.

Anexo 4: Proyección de atributos seleccionados



DESCRIPCIÓN: Proyección los atributos seleccionados de los alumnos del Instituto de Educación Superior Tecnológico Privado ISTEPSA.

FECHA: Octubre 2020.