Técnicas de minería de datos para determinar la deserción escolar

Alejandro Apaza-Tarqui
Walter Borda-Navedos
Noemí Cayo
Jhon Huanca-Suaquita

DOI: 10.35622/inudi.b.053

EDITADA POR
INSTITUTO
UNIVERSITARIO
DE INNOVACIÓN CIENCIA
Y TECNOLOGÍA INUDI PERÚ

INUDI PERÚ
INSTITUTO UNIVERSITARIO DE INNOVACIÓN
CIENCIA Y TECNOLOGÍA

Técnicas de minería de datos para determinar la deserción escolar

DOI: https://doi.org/10.35622/inudi.b.053

Alejandro Apaza-Tarqui
https://orcid.org/0000-0003-1622-8862
apazatarqui@unap.edu.pe

Walter Borda-Navedos
https://orcid.org/0000-0003-1916-3638
wborda28@gmail.com

Noemí Cayo
https://orcid.org/0000-0002-9690-3006
noemicayo@unap.edu.pe

Jhon Huanca-Suaquita
https://orcid.org/0000-0001-6683-8859
jr.huanca@unaj.edu.pe

Instituto Universitario
de Innovación Ciencia y Tecnología Inudi Perú

Técnicas de minería de datos para determinar la deserción escolar

Alejandro Apaza Tarqui
Walter Borda Navedos
Noemí Emperatriz Cayo Velásquez
Jhon Richard Huanca Suaquita
(Autores)

ISBN: 978-612-5069-42-9 (PDF)
Hecho el depósito legal en la Biblioteca Nacional del Perú N° 2022-11406
DOI: https://doi.org/10.35622/inudi.b.053

Editado por Instituto Universitario de Innovación Ciencia y Tecnología Inudi Perú S.A.C
Urb. Ciudad Jardín Mz. B3 Lt. 2, Puno – Perú
RUC: 20608044818
Email: editorial@inudi.edu.pe
Teléfono: +51 973668341
Sitio web: https://editorial.inudi.edu.pe

Primera edición digital
Puno, noviembre de 2022

Libro electrónico disponible en
https://doi.org/10.35622/inudi.b.053

Editores:
Wilson Sucari / Patty Aza / Antonio Flores

Contenido

DEDICATORIA

Dedicamos este trabajo con mucho cariño a todos nuestros familiares,

que siempre son una motivación para seguir adelante,

fuente de inspiración para seguir adelante y

lograr todas las metas que soñamos alcanzar.

AGRADECIMIENTO

A la Universidad Nacional del Altiplano.

Nuestra alma mater,

por ser la fuente y generación de los conocimientos

y experiencias que vamos acumulando

para el desarrollo personal de cada uno de nosotros.

SINOPSIS

La presente investigación tuvo por objetivo determinar las técnicas de minería de datos y los factores asociados que permitan segmentar los alumnos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA, en Andahuaylas (Perú). Para este fin se aplicaron técnicas de Aprendizaje Automático y Minería de Datos implementadas en software WEKA: Se aplicó el método de evaluación *CfsSubsetEval* y el método de búsqueda *BestFirst* para seleccionar los factores de mayor significancia, 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(*Self Organizing Maps*, SOM). Se obtuvo 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, extracción de conocimiento (KDD), deserción estudiantil, segmentación de alumnos.

ABSTRACT

The objective of this research was to determine the data mining techniques and the associated factors that allow the segmentation of students at risk of dropping out at the Instituto Superior Tecnológico Privado ISTEPSA, in Andahuaylas (Peru). For this purpose, Automatic Learning and Data Mining techniques implemented in WEKA software were applied: The CfsSubsetEval evaluation method and the BestFirst search method were applied to select the most significant factors, to establish the patterns the association algorithm A was used. priori and to segment, the Expected Value Maximization algorithm "Expectation Maximissation" (EM) and Kohonen's self-organizing maps (Self Organizing Maps, SOM) were used. The following results were obtained: 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: machine learning, knowledge extraction (KDD), student dropout, student segmentation.

INTRODUCCIÓN

La problemática nacional referente al abandono de alumnos en las entidades de formación profesional en sus diferentes niveles es muy preocupante, para el caso de Institutos técnicos privados este problema es aún más preocupante porque las tazas de deserción son más altos, ello se debe a diversos factores y patrones como; Sociales, económicas, políticas, demográficas y académicas. En vista a lo descrito la investigación aborda la siguiente temática: Segmentación de alumnos con riesgo deserción mediante las técnicas de la minería de datos en el Instituto Superior Tecnológico Privado ISTEPSA de la ciudad de Andahuaylas, Región Apurímac; para ello se utilizó las metodologías y algoritmos de Machine Learning; los cuales son 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 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 los factores que ocasionan la deserción de alumnos, para ello la investigación se enfoca en las características sociales, económicas, demográficas y académicas de los alumnos; específicamente en aquellas características cuantificables puesto que los algoritmos de *clustering* de la minería de datos elegidos trabajan con distancias. Por otro lado, la metodología usada en la presente investigación es Descubrimiento de Conocimiento en Base de Datos (KDD), por sus siglas en inglés significa *Knowledge Discovery* in Databases; 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 fueron 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 de la investigación, asimismo la participación de la Directora del Instituto.

Una vez aplicadas los criterios de pre procesamiento de datos el propósito fue: Segmentar los alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA, mediante las técnicas de minería de datos.

CAPÍTULO I

MARCO TEÓRICO

1.1 Antecedentes de investigación

1.1.1 Antecedentes nacionales

Mollo (2018) desarrolló 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, para minería de datos CRISP-DM, y las 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), tipo de ingreso (factores económicos), 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) realizó la 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 elección de 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) ejecutó 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.

Yamao (2018) desarrolló la investigación "Predicción del rendimiento académico mediante de minería de datos en 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) logró la Segmentación demográfica y relaciones con los clientes en la empresa Hotel Cielo, Distrito de Tarapoto, utilizando la prueba 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 es el trabajo de investigación de Linarez (2019), sobre 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 la 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.

1.1.2. 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 más 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.

1.2 Bases teóricas

1.2.1 Machine Learning

Inteligencia Artificial (IA)

Mathivet (2018), indica que la Inteligencia Artificial es: "Un concepto difícil de definir con precisión, porque adopta muchas formas. Resulta difícil, también, medirla, y las pruebas de C.I. están sesgadas. Se resume 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 este término fue acuñado en 1956 por John McCarthy y desde entonces ha tenido avances muy importantes y también su definición ha evolucionado. Sin embargo, Terrones, (2018) 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). También se define como "la ciencia y la ingeniería de crear máquinas inteligentes, especialmente programas de computadora inteligentes. Está relacionada con la tarea similar de utilizar computadoras para comprender la inteligencia humana" (McCarthy, 2007).

Los autores Russell y Norvig (2010) conceptúan a inteligencia artificial como una aproximación que se tome al enfrentar el problema de construir entes que emulan el comportamiento inteligente. Se agrupan en cuatro grandes grupos, diferenciandose principalmente en si se busca emular el comportamiento o el pensamiento, y de si se busca hacerlo humanamente, o racionalmente.

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).

Aprendizaje Automático No Supervisado

Machine Learning (ML) es una disciplina científica que maneja sistemas inteligentes, es decir, sistemas que aprenden automáticamente al identificar ciertos patrones presentes en los datos, usando algoritmos que se encargan de revisar datos mediante ejemplos o instrucciones predefinidas a fin de predecir comportamientos futuros a partir de información adicional y reajustar el resultado; maneja conocimiento inductivo obteniendo un enunciado general en base a enunciados que describen casos particulares (Mohri et al., 2018), su proceso es similar a la minería de datos, ambos sistemas buscan patrones entre los datos iniciales; mineración extrae los datos para la comprensión humana mientras que ML utiliza esos datos para detectar patrones y reajustar las acciones del programa.

Como menciona Baviera (2016), el aprendizaje automático nació en el campo de la informática cuyo procesamiento está asociado con el aprendizaje, donde 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, en ese sentido, los términos aprendida y asimilada concluyen 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, cuyos procesos son:

- **Seleccionar problema:** Definir el problema real a resolver, analizando ventajas-desventajas, costos y beneficios de alto nivel.
- **Seleccionar atributos:** Establecer los atributos o variables de fácil acceso y extracción para resolver el problema.
- **Seleccionar modelo:** Definir el algoritmo que mejor se adapte a la resolución del problema.

- **Preparar datos:** La extracción, la consolidación y el escalamiento de atributos que conforman el archivo de datos a utilizar.
- **Analizar datos:** Depurar los datos o atributos en base a los algoritmos que resuelvan mejor el problema planteado
- **Evaluar el modelo:** El 60% de los datos recolectados serán datos de entrenamiento; el 20% para realizar una validación cruzada del rendimiento del modelo para depurarlo y reajustarlo; y el 20% restante de los datos recolectados para la prueba del modelo generado y su comprobación.
- **Publicar modelo:** usarlo en un ambiente productivo.
- **Monitorear y ajustar:** Una vez que el modelo esté en producción, monitorear su desempeño y hacer los reajustes para mejorarlo.

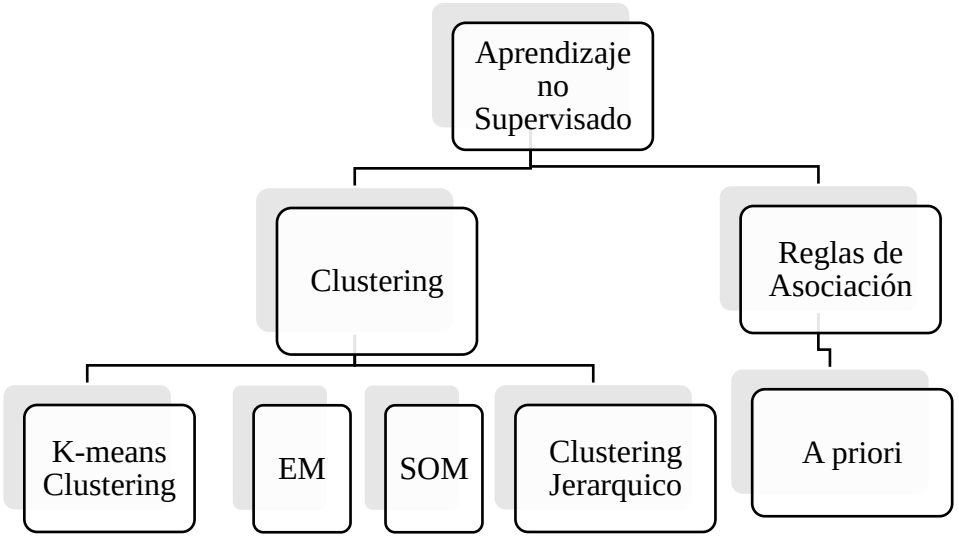
Además, se tiene los siguientes tipos de aprendizaje automático.

Aprendizaje no Supervisado o descriptivo

Conforme a Mativet (2017) utiliza procedimientos inductivos, extrayendo conocimiento sólo de los datos, como en el caso del análisis de clústeres para la agrupación, es menos común por no esperar algún resultado previamente planificado, 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. Sólo se dan los datos finales (inputs) a la máquina para que encuentre patrones interesantes a partir de esos datos (Murphy, 2012). Estos algoritmos se clasifican en:

Figura 1

Principales algoritmos de aprendizaje no supervisado



Los problemas de aprendizaje no supervisado se clasifican en:

- **Clustering:** dividen los datos en grupos según algunas de sus características.
- **Asociación:** manejan reglas o características que definan grandes cantidades de datos por ejemplo en las compras en línea un usuario compra a, también compra b.

Aprendizaje Supervisado o predictivo

Tal como afirma Murphy (2012) que la máquina no sólo aprende de los datos de entrada sino que es posible darle modelos o datos de salida adicionales ya categorizados para que el aprendizaje sea mucho más fiable. El uso de aprendizaje supervisado requiere algoritmos especializados que detecten patrones en los datos, desde una programación avanzada, similar al análisis de contenido automatizado, aplicado a grandes cantidades de datos requieren plataformas distribuidas para el procesamiento en paralelo, lo cual, implica el desarrollo de código y el despliegue de centros de cómputo en la nube, ha prosperado una serie de servicios comerciales que permiten el aprendizaje automático de manera mucho más sencilla (Arcila Calderón et al., 2016).

El aprendizaje supervisado que a partir de una muestra:

$$£ = \{(X_i, Y_i) | i = 1,2, \dots n\} \qquad 1$$

Construida por n realizaciones de un par de variables (X, Y), se tiene una función $f: X \to Y$, para un vector de entrada X, se predice con cierto grado de confianza la variable $Y = f(X)$. Para cada observación (X_i, Y_i) de £, a la variable $X_i \in X$ se le llama variable de entrada, explicativa o input y a $Y_i \in Y$ variable dependiente u output (Bourel, 2012).

Estos aprendizajes se subdividen en:

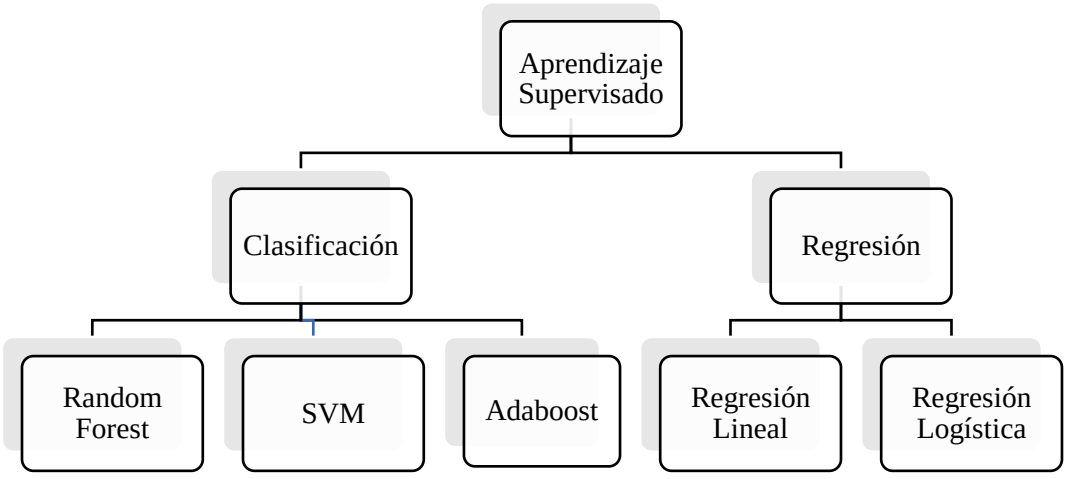
a) **Clasificación:** predice la clase de inputs en un conjunto de categorías prefijadas, ejemplo: determinar que una noticia es de deportes, salud, cine, etc.

b) **Regresión:** predice un valor real en base a los valores almacenados, ejemplo: el ganador de pela de gallos, a partir de peleas anteriores de los mismos.

Los algoritmos más utilizados en el aprendizaje supervisado se muestran en el gráfico siguiente.

Figura 2

Principales algoritmos de aprendizaje supervisado



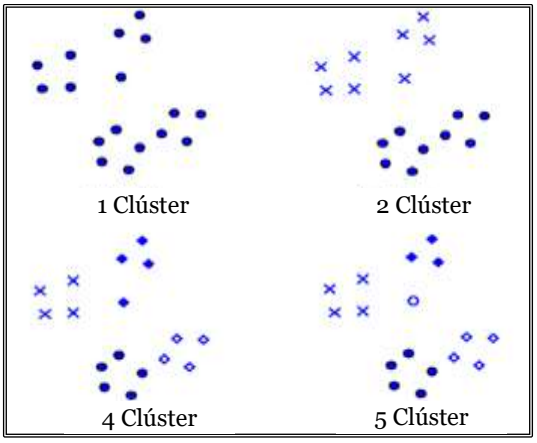
El aprendizaje supervisado consiste en utilizar una muestra de datos para el entrenamiento y prueba de algoritmo, luego se comparan los resultados obtenidos, si presenta márgenes de error se deben aplicar técnicas para disminuir el error y finalmente dar o aproximar con los resultados esperados, los parámetros usados en el entrenamiento serán las variables que el investigador manipule para determinar un modelo apropiado para el caso de estudio (Mativet 2017), en este caso se supervisa el entrenamiento del algoritmo de forma iterativa hasta encontrar el modelo que aproxime los resultados reales.

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 3

Diversos Clustering para datos en observación



Nota. Tomado de 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.

Existen dos tipos básicos de clustering y se distinguen por ser de naturaleza jerárquica o no jerárquica. (Johnson, 2000), dentro de los métodos de aglomeración, el jerárquico (árboles ultramétricos) es más usado por la comunidad científica, se divide en aglomerativo y divisivo.

Los aglomerativos se parte inicialmente de los objetos, que se van progresivamente fusionando para formar particiones sucesivas que contienen todos los objetos originales, mientras que el divisivo parte del conjunto total Ω que se subdivide progresivamente en conglomerados pequeños.

En grupos no jerárquicos se forman grupos homogéneos sin establecer relaciones entre los grupos; mientras que los grupos jerárquicos se van fusionando progresivamente, baja la homogeneidad entre los grupos, cada vez más amplios que se van formando. Se mide la homogeneidad mediante un índice (distancia fenética) siendo una característica de la taxonomía numérica. El análisis de clúster se aplica sobre una matriz de distancias y no sobre una de asociación. Para descriptores cualitativos debe ser transformada en distancias, éstas a su vez, se grafican en diferentes formas, siendo los dendrogramas y la dispersión de puntos en un plano cartesiano fáciles de interpretar (Franco and Hidalgo, 2003).

Métodos jerárquicos aglomerativos: se calcula de la matriz de distancias entre los elementos de la muestra, la cual contiene las distancias existentes entre cada elemento de la muestra y todos los restantes. A continuación, se buscan los dos elementos más próximos (similares en términos de distancias) y se agrupan en un conglomerado (*cluster*). El resultante es indivisible a partir de ese momento. De esta manera se van agrupando los elementos en conglomerados cada vez más grandes y más heterogéneos hasta llegar al último paso, en el que todos los elementos de la muestra quedan agrupados en un único grupo (Ferreira and Hitchcock, 2009). A partir de la primera agrupación de dos casos en un *cluster*, debe quedar claro además cuál es el criterio de agrupación de un *cluster* de 2 o más instancias con un caso individual, o más generalmente, cuál es el

criterio de agrupación de dos *clusters* previamente formados con distintos tipos de medidas para estimar la distancia existente entre los casos y la posibilidad de seleccionar uno entre una gran variedad de métodos de agrupamiento establecidos. Pero no existe ninguna combinación de estas posibilidades que optimice la solución obtenida. En general, será conveniente valorar distintas soluciones para elegir la más consistente.

El método de agrupación de enlace simple (*Single linkage clustering*) o del vecino más próximo (*nearest neighbor cluster*ing), inicia seleccionando y uniendo los dos elementos de la matriz de distancias que se encuentran más próximos. La distancia del nuevo clúster respecto a los restantes elementos de la matriz se calcula como la menor de las distancias entre cada elemento del clúster y el resto de los elementos de la matriz. En los pasos sucesivos, la distancia entre dos clústeres se calcula como la distancia entre sus dos elementos más próximos. Así, la distancia d_{AB} entre clúster *A* y *B* se calcula mediante: $d_{AB} = \min (d_{ij})$ donde d_{ij} es la distancia entre los elementos *i* y *j*, el primero perteneciente al clúster *A* y el segundo al clúster *B*. Esta técnica, por ejemplo, contribuye a crear *clusters* "longanizas" donde el último *cluster* añadido está adicionado al *cluster* inicial por asociaciones reiteradas, pero muy distante de éste.

El método de agrupación de enlace completo (*Complete linkage clustering*) o del vecino más lejano (*furthest neighbor clustering*), se comporta de manera opuesta al anterior. La distancia entre cluster se calcula como la distancia entre sus dos elementos más alejados. Es decir, la distancia entre dos clústeres A y B se calcula como:

$$d_{AB} = \max (d_{ij})$$

Este método, a diferencia del anterior, forma *clústeres* más "compactos". El método de agrupación de vinculación promedio (*Average linkage clustering*) o vinculación inter-grupo (*unweighted Pair-group arithmetic averages)* presenta la ventaja, sobre los dos anteriores, de aprovechar la información de todos los miembros de los dos conglomerados que se comparan. La distancia entre dos conglomerados se calcula como la distancia promedio existente entre todos los pares de elementos de ambos conglomerados: nA nB

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} dij \qquad 2$$

El método de agrupación de Ward o de varianza mínima, debían constituirse de tal manera que, al fundirse dos elementos, la pérdida de información resultante de la fusión fuera mínima, cuantifica la cantidad de información como la suma de las distancias al cuadrado de cada elemento respecto al centroide del conglomerado que pertenece (SCE = Suma de Cuadrados del Error). Para ello, inicia calculando, en cada clúster, el vector de medias de todas las variables, es decir, el centroide multivariante. A continuación, se calculan las distancias euclidianas al cuadrado entre cada elemento y los centroides (vector de medias) de todos los clústers, además, se suman las distancias correspondientes a todos los elementos.

En cada paso se unen aquellos clústeres que dan lugar a un menor incremento de la SCE, intra clústeres.

$$SCE = \sum_{j=1}^{k} \left(\sum_{i=1}^{n_i} X_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{n_i} X_{ij} \right)^2 \right) \qquad 3$$

El método de enlace medio dentro de los grupos, o de vinculación Intra-grupos, como en el caso anterior, aprovecha la información de todos los miembros de los conglomerados que se comparan uniéndolos previamente. La distancia entre dos conglomerados se calcula como la distancia promedio existente entre todos los miembros del conglomerado unión de ambos:

$$d_{AB} = \frac{1}{C^2_{n_A + n_B}} \sum_{i,j \in A \cup B} d_{ij} \qquad 4$$

Los métodos de conglomerados jerárquicos proporcionan un dendrograma (Zhao y Karypis, 2005) que es una gráfica para visualizar las relaciones entre las distintas características, la matriz de similitud involucra a veces cientos de características que resultarían imposible de analizar en conjunto.

Métodos jerárquicos divisivos: son algoritmos que inician con todas las observaciones unidas en un solo clúster, en pasos sucesivos el clúster se va dividiendo, y el resultado de la división conforma dos nuevos clústeres, que se vuelven a dividir.

Métodos de aglomeración particionales: son algoritmos que inician repartiendo las *n* observaciones en *K* grupos. La primera asignación se hace aleatoriamente. En cada uno de los grupos se obtiene el vector de medias (centro del grupo) y se asigna secuencialmente cada observación al grupo cuyo centro esté más cercano. En cada etapa se re-calcula el centro del grupo al que se añade una observación y el centro del grupo del que se elimina esa observación.

Número de clústeres óptimos y validación del agrupamiento: el método de aglomeración jerárquica se representa con un dendrograma o un gráfico similar a un árbol. La distancia cofenética entre dos objetos en un dendrograma es la distancia con la cual los dos objetos se unen a un mismo grupo. Una matriz cofenética es una matriz que representa las distancias cofenéticas entre todos los pares de objetos. Se calcula el coeficiente de correlación *r* de Pearson entre la matriz de disimilaridad original y la matriz cofenética, a este coeficiente se le denomina Coeficiente de Correlación Cofenético (CCC). El método con mayor CCC es visto como el mejor modelo para una determinada matriz de distancia (Borcard *et al.*, 2011). Además, tiene una significación estadística aplicando el *test* de Mantel, como una matriz de correlación entre dos matrices de disimilaridad, en este caso la ultramétrica del árbol y la matriz de distancia original de los datos. Sin embargo, la significación no es evaluada directamente porque hay *N(N-1)/2* entradas para sólo *N* observaciones. Para esto Mantel desarrolló un *test* asintótico que utiliza la permutación de las filas y las columnas de la primera matriz de disimilaridad, repetidas veces.

El coeficiente aglomerativo (CA) es una medida de la calidad de la estructura encontrada por el algoritmo de clasificación y varían en un rango de 0 (no estructura) y 1 (estructura completamente definida).

Cuando se emplea técnicas de grupos jerárquicos, el investigador no está interesado en la jerarquía completa sino en un subconjunto de particiones obtenidas a partir de ella. Las particiones se obtienen cortando el dendograma o seleccionando a través de algunos métodos más utilizados como el Índice de Dunn que es la razón entre la distancia menor entre los individuos que no pertenecen al mismo clúster y la mayor distancia intra clúster. La idea es identificar los conglomerados que están bien compactos y bien separados de los

demás. Dada una partición de conglomerados donde c_i representa el *i-ésimo* conglomerado de la partición, se define el Índice de Dunn por:

$$D_n = \min_{1<i<n}\left\{\min_{1<j\neq i\leq n}\left\{\frac{d(c_i,c_j)}{\max_{1<k<n}\left(d'(c_k)\right)}\right\}\right\} \qquad 5$$

donde *d(ci,cj)* – es la distancia entre los conglomerados c_i, *y* c_j y *d'(ck)* representa una distancia intra conglomerado c_k, es la distancia máxima entre los elementos de clústeres. Si el conjunto de datos tiene clústeres bien compactos y definidos entonces D_n es grande. El número óptimo de clústeres es aquel que maximiza *D*.

Otra medida del número óptimo de clústres es el ancho de silueta. El ancho de la silueta de la i-ésima observación es definida por:

sil_i = (b_i - a_i)/ max(a_i, b_i), donde, a_i denota la distancia promedio entre la observación *i* y todas las otras que están en el mismo clúster de *i*, y *bi* denota la distancia promedio mínima de *i* a las observaciones que están en otros clústeres.

El índice muy utilizado en la evaluación de los clústeres es el Índice de Calinski-Harabasz (G1) (1974), que es un índice definido en término de las trazas de las matrices inter e intra conglomerado. (Gan *et al.*, 2007)

Sea *n* el número de instancias (filas) de la matriz de datos y *k* el número de conglomerados, entonces el índice G1 es definido cómo:

$$G1(u) = \frac{Traza(B_u)/(u-1)}{Traza(W_u)/(n-u)} \qquad 6$$

donde, $X = \{Xij\}, i = 1,\ldots,n;\ j = 1,\ldots,m$ – matriz de datos,

n – número de objetos,

m – número de variables,

u – número de conglomerados ($u = 2,\ldots,n-1$),

$W_u = \sum_r \sum_{i\in Cr}\left(X_{ri} - \overline{X_r}\right)\left(X_{ri} - \overline{X_r}\right)^T$ – matriz de dispersión intra conglomerado para los datos agrupados en el conglomerado *u*,

$B_u = \sum_r n_r \left(\overline{X_r} - \overline{X}\right)\left(\overline{X_r} - \overline{X}\right)^T$ – matriz de dispersión inter conglomerado para los datos agrupados en el conglomerado *u*,

$r = 1,\ldots,u$ – número del conglomerado,

$\overline{X_r}$ – centroide o mediode del conglomerado *r*,

$\overline{X}$ – centroide o mediode la matriz de datos,

C_r – los índices de objetos en el conglomerado *r*,

n_r – número de objetos en el conglomerado *r*.

El valor de *u*, que maximiza a G1(*u*), es considerado como el número de conglomerados óptimo.

Existen dos índices muy útiles para medir la calidad interna de un conglomerado, estos son:

Índice de Baker & Hubert (G2) que es una adaptación del estadístico Gamma de Goodman & Kruskal's. (Everitt *et al.*, 2001)

$$G2(u) = \frac{S(+) - S(-)}{S(+) + S(-)} \qquad G2(u) \in [-1,1] \qquad 7$$

dónde: S (+) – número de comparaciones concordantes (la disimilaridad dentro del conglomerado es estrictamente menor que la disimilaridad entre conglomerados), S(-) – número de comparaciones discordantes (la disimilaridad dentro del conglomerado es estrictamente mayor que la disimilaridad entre conglomerados), *u* – número de conglomerados (*u* = 2, ..., *n*-1), *n* – número de objetos.

El valor de *u*, que maximiza a G2(*u*), es considerado como el número de conglomerados óptimo.

Índice de Hubert & Levine (G3) (Gatnar and Walesiak, 2004)

$$G3(u) = \frac{D(u) - r.D_{\min}}{r.D_{\max} - r.D_{\min}}, D_{\min} \neq D_{\max} \qquad G3(u) \in (0,1), \qquad 8$$

dónde: *D(u)* – todas las disimilaridades intra conglomerado,

r – número de disimilaridades intra conglomerado,

D_{min} – menor disimilaridad intra conglomerado,

D_{max} – mayor disimilaridad intra conglomerado,

El valor de u, que maximiza a G3(u), es considerado como el número de conglomerados óptimo.

Combinación de agrupamientos: No existe un algoritmo de agrupación óptimo para un problema determinado, es difícil seleccionar cual será el método de aglomeración que logre encontrar una mejor estructura para separar las accesiones.

En la búsqueda de mejores algoritmos de clasificación aparece una tendencia a combinar varios algoritmos de agrupamiento en el mismo problema. La base de estos algoritmos está en la lógica de utilizar el criterio de varios expertos y combinarlos en aras de lograr un mejor rendimiento.

Dada N diferentes particiones de los datos $X = \{x_1, x_2, \ldots, x_n\}$ de n objetos se define una combinación de agrupamiento $P=(P^1,P^2,\ldots,P^N)$ donde $P^i = (C_1^i, C_2^i, \ldots C_{k_i}^i)$ tiene k_i conglomerados, el problema consiste obtener una partición P^*, la cual es el resultado de combinar toda la información existente de las N particiones en P. Se profundizará en estas ideas en el capítulo III.

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.

Herramientas de Minería de Datos

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

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.

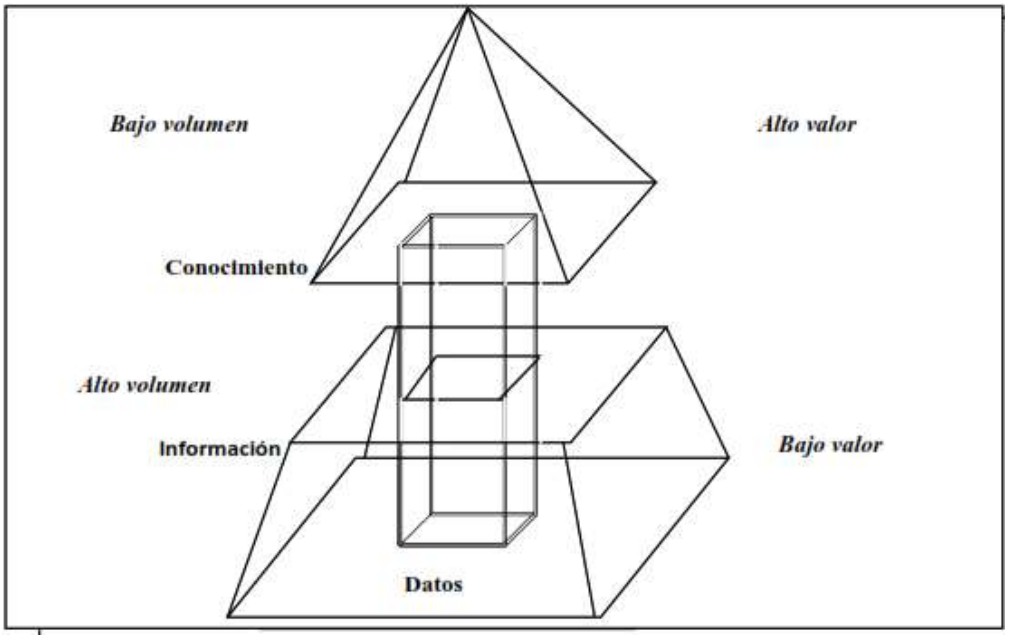
Metodología KDD

Periera et al. (2013) enfatizan que el proceso KDD en la educación no es término nuevo su aplicación es muy importante y relevante en los últimos años, pues permite analizar grandes volúmenes de datos encontrando relaciones y patrones no triviales sobre una cantidad extensa de información y conocimiento

extrayendo tendencias y modelos, donde su interpretación representa un valor agregado, en ese contexto, KDD consta de una jerarquía que existe entre datos, información y conocimiento.

Figura 4

Jerarquía de la base de datos; entre datos, información y conocimiento



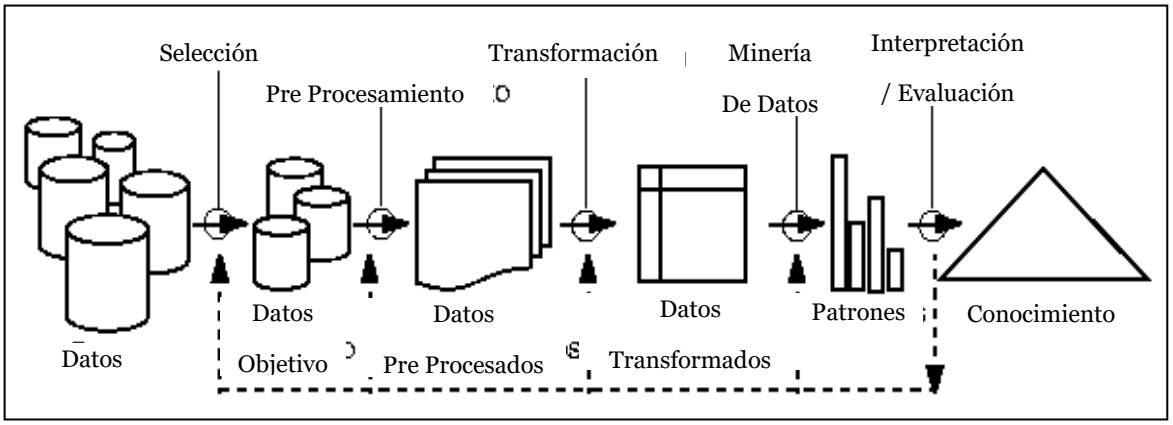
Descubrimiento de Conocimiento en Bases de Datos

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 5

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:

Determinación de las fuentes de información.

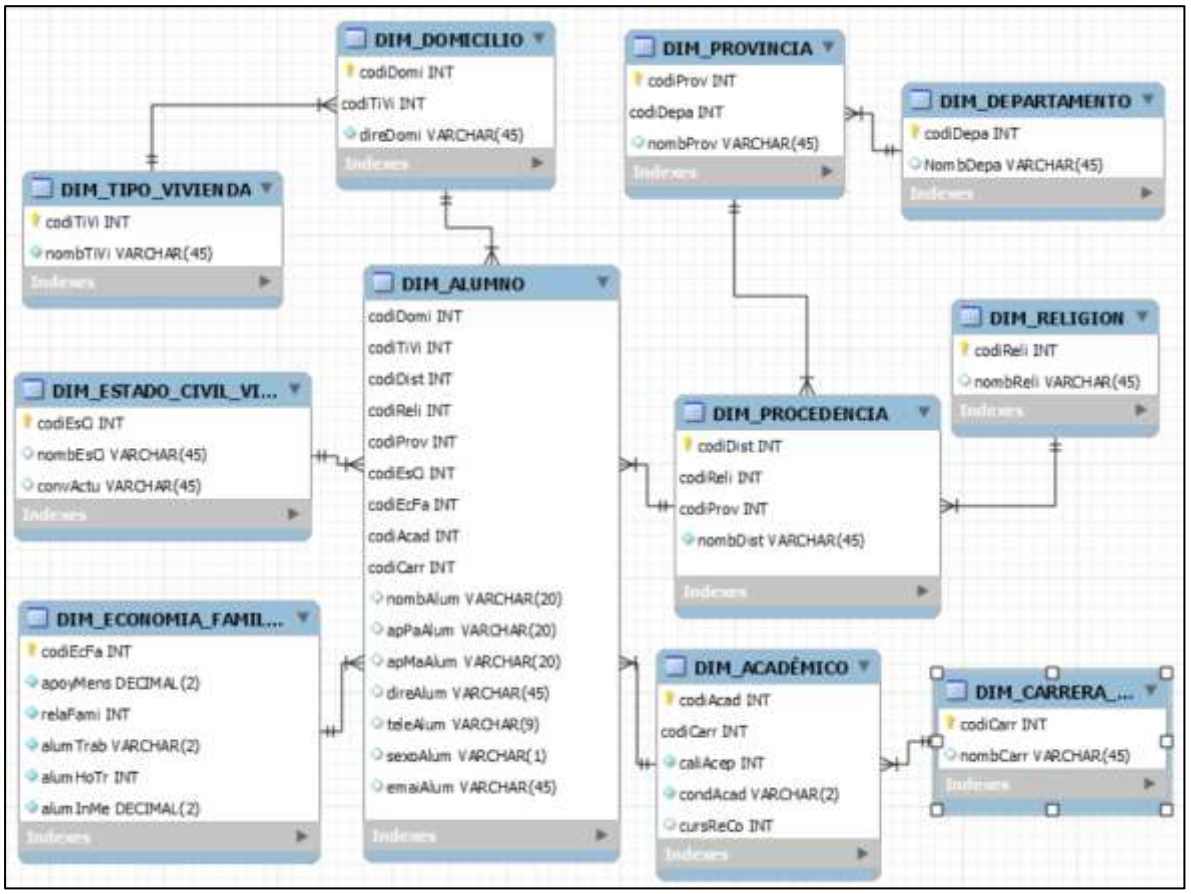
Este es la primera fase del proceso de investigación y para lo cual existen diversos instrumentos que el investigador puede usar, para nuestro caso utilizaremos el instrumento cuestionario estructurado el cual permitirá recoger datos cuantitativos de los alumnos del Instituto de Educación Superior Tecnológico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

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, este esquema está 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.

Figura 6

Esquema de copo de nieve



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.

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),

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.

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).

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.

Representación de patrones

Se distinguen dos técnicas de representación no simbólicas y simbólicas.

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 estas técnicas es poca (o nula) inteligibilidad, destacan algoritmos basadas en: Redes Neuronales Artificiales, Lógica Difusa, Algoritmos Genéticos y combinaciones entre ellos.

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.

Selección de subconjunto de Atributos

Según Hall y Smith (1998) 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 y 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).

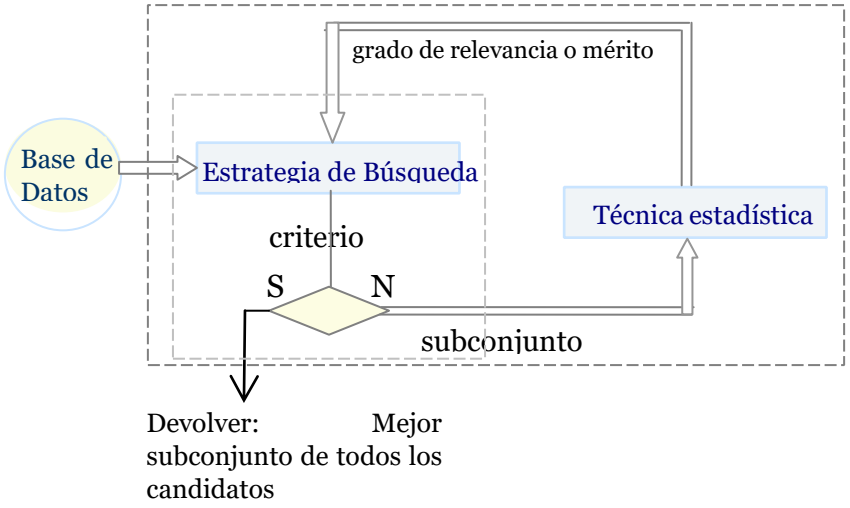
Clasificación de los evaluadores de Factores

Existen varios algoritmos para seleccionar factores (atributos), uno constituye el modo de aplicación, según el cual es visto como de filtro (filter), envolvente (wrapper) y orden (ranking):

Los selectores con estrategia de filtro son las que se apoyan en la extracción de información útil de la base de datos, utilizando técnicas estadísticas para evaluar los subconjuntos. Son independientes de los algoritmos de *machine learning*, cuyo proceso es:

Figura 7

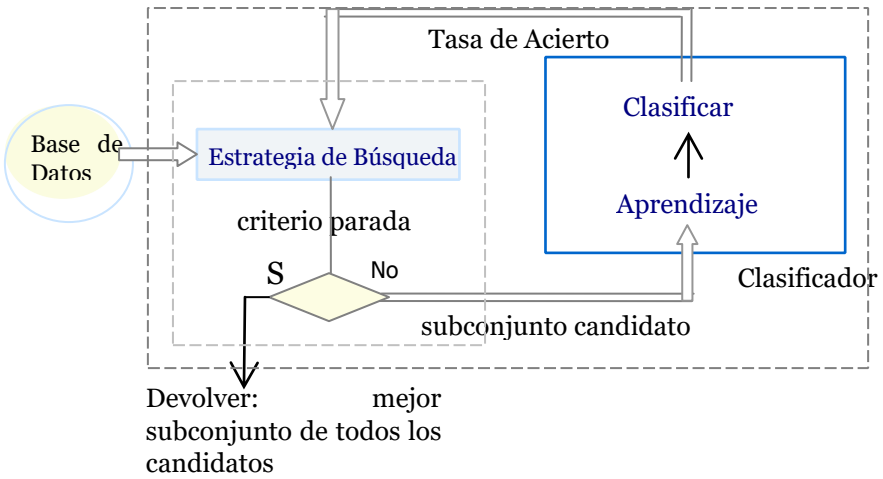
Selección de Atributos utilizando una estrategia de filtro



Los envolventes son las que se auxilian de la precisión del clasificador para evaluar a los subconjuntos del espacio, esta estrategia ofrece mejores resultados, ya que en un paso previo a la clasificación del Algoritmo de Aprendizaje escoge a los atributos que mejor representen el conocimiento para su construcción; pero, es altamente costosa

Figura 8

Selección de Atributos utilizando una estrategia envolvente



De acuerdo a la forma de evaluación de los factores, se dividen en evaluadores individuales y de subconjuntos de factores. Los primeros realizan un ranking en

orden descendente según el grado de relevancia de estos para la clase de factor, sin tener en cuenta la redundancia entre ellos, mientras que los subconjuntos evalúan el grado de predicción de una serie de atributos hacia la clase, teniendo en cuenta la repetición de ellos con dependencia de la técnica empleada, considerando las categorías: distancia, información, dependencia y basados en la tasa de error del clasificador (Dash et al., 2003).

Medidas de Distancia: halla la relevancia de un subconjunto de atributos a partir del cálculo de distancias entre poblaciones. Es decir, dado dos conjuntos de prototipos pertenecientes a diferentes clases, se calcula el grado de separabilidad estadística entre ambas poblaciones teniendo en cuenta el espacio vectorial perteneciente al subconjunto de atributos a evaluar, basada en la premisa de que un subconjunto óptimo es aquel que logra la mayor separabilidad o diferencia entre las funciones de densidad de probabilidad de todas las clases. En la ecuación (1.1) se muestra el cálculo de este valor para el subconjunto S dada las clases C_i y C_j.

$$J_{IJ} = \int_x \left\{ \sqrt{P(S \mid C_i)} - \sqrt{P(S \mid C_J)} \right\}^2 \qquad 9$$

CfsSubsetEval: Los métodos de filtro implementadas en el software Weka realizan la búsqueda que permite la selección de un subconjunto de atributos representativo del problema original, denominada CfsSubsetEval mide la "bondad" de los subconjuntos de atributos teniendo 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 predictivos correlacionados a la clase:

$$G_S = \frac{k\bar{r}_{ci}}{\sqrt{k + k(k-1)\bar{r}_{ii}}} \qquad 10$$

Dónde: G_S es el mérito heurístico del subconjunto S conteniendo k características, $\bar{r}_{ci}$ 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. Asume que los atributos son independientes condicionalmente dada la clase, siendo una

simplificación aceptable en algunos casos, pero si existe una fuerte interacción entre distintos atributos, entonces 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.

$$H(Y) = \sum_{y=1} p(y)\log_2(p(y)) \qquad 11$$

$$H(Y/X) = \sum_{X=1} p(x)\sum_{y=1} p(y \mid x)\log_2(p(y \mid x)) \qquad 12$$

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

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

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

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

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

$$P(C_i \mid v_1, v_2, ..., v_n) = \frac{P(C_i)\prod P(v_j \mid C_i)}{P(v_1, v_2, ..., v_n)} \qquad 13$$

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:

$$s(X \Rightarrow Y) = s(X \cup Y) \qquad 14$$

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: $P(X / Y) = P(X \cup Y) / P(X) = c(X \Rightarrow Y)$ 15

Lift (levantamiento): cuantifica la relación existente entre X e Y: se define como:

$$\text{lift}(\, s(X \Rightarrow Y) = s(X \Rightarrow Y) / s(Y) \qquad 16$$

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):

$$(X \Rightarrow Y) = s(X \Rightarrow Y) - s(X)s(Y) = P(X \cap Y) - P(X)P(Y) \qquad 17$$

Conviction (convicción):

$$(X \Rightarrow Y) = 1 - s(Y)/(1 - conf(Y \Rightarrow Y)) = P(X)P(Y^{'})/P(X \cap Y^{'}) \qquad 18$$

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.

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).

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 este 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:

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

$$L = \ln\left(\prod_{t=1}^{N} P(y^{(t)} / x^{(t)}, \theta)\right) = \sum_{t=1}^{N} \ln\left(\sum_{i=1}^{K} g_i^{(t)} \sum_{J=1}^{L} g_{j/i}^{(t)} P_{ji}(y^{(t)})\right) \qquad 19$$

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 sus 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:

$$P(y / x) = \sum_{i=1}^{K} g_i P_i(y / x) \qquad 20$$

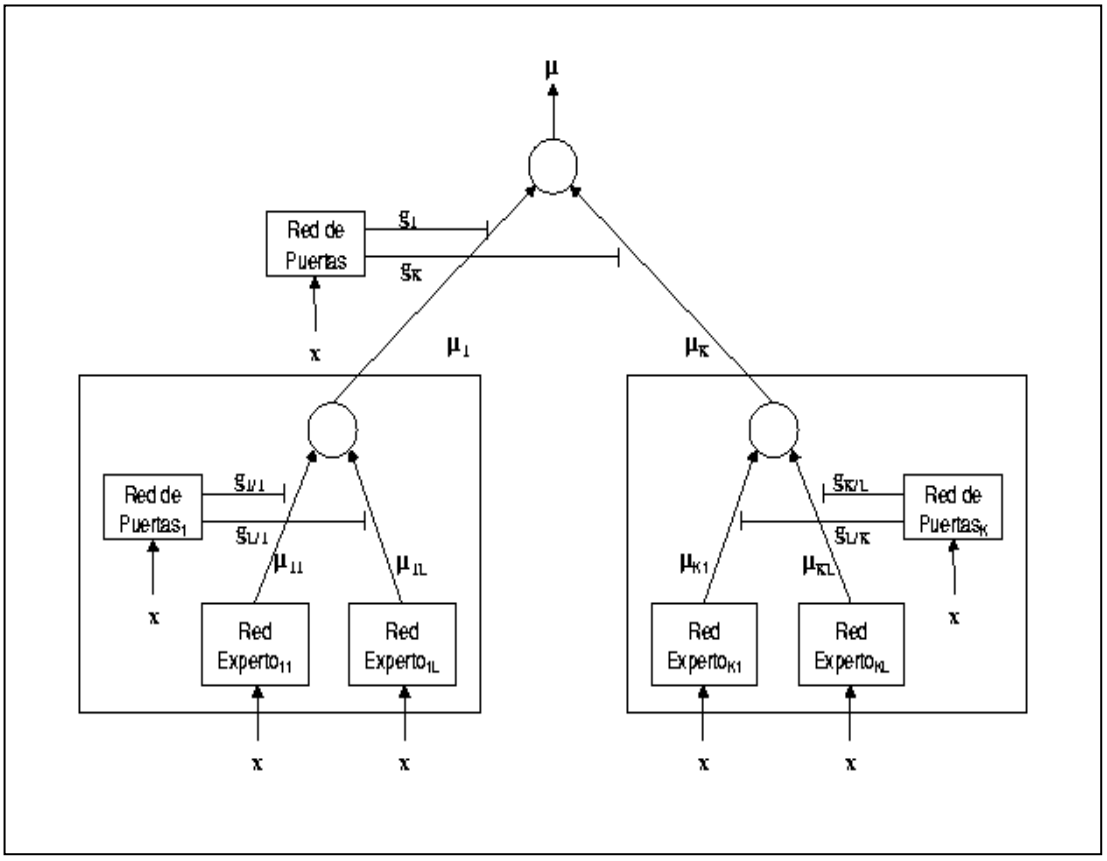
Siendo g_i las correspondientes salidas de las redes de puertas.

$$P(y / x) = \frac{1}{(2\pi)^{k/2}} \sum_{i=1}^{K} g_i \exp\left(-\frac{1}{2}\|y_i - \mu_i\|^2\right) \qquad 21$$

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

Figura 9

El modelo de Jacobs-Jordan de dos niveles de módulos expertos, y las redes de puertas (gating networks)



Nota. Tomado de 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:

$$g_i = \frac{\exp \xi_i}{\sum_{j=1}^{k} \exp \xi_j} \qquad 22$$

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

$$P(y / x, \Theta) = \sum_{i=1}^{K} g_i(x, v_i) \sum_{j=1}^{L} g_{j/i}(x, v_{j/i}) P_{ji}(y / x, w_{ji}) \qquad 23$$

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:

$$h_i = \frac{g_i \sum_{j=1}^{L} g_{j/i} P_{ji}(y)}{\sum_{i=1}^{k} g_i \sum_{j=1}^{L} g_{j/i} P_{ji}(y)} \qquad 24$$

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**:

$$h_{j/i} = \frac{g_{j/i} P_{ji}(y)}{\sum_{j=1}^{L} g_{j/i} P_{ji}(y)} \qquad 25)$$

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:

$$Q = \prod_{t=1}^{N} P(y^{(t)} / x^{(t)}, \theta) \qquad 26$$

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:

$$\text{CC}\, L = \sum_{t=1}^{N} \ln\left(\sum_{i=1}^{K} g_i^{(t)} \sum_{J=1}^{L} g_{j/i}^{(t)} P_{ji}(y^{(t)}) \right) = \sum_{t=1}^{N} \ln\left(\prod_{i=1}^{K} \prod_{j=1}^{L} (g_i^{(t)} g_{j/i}^{(t)} P_{ji}(y^{(t)}))^{z_{ij}^{(t)}} \right) \quad 27$$

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:

$$L = \sum_{t=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{L} z_{ij}^{(t)} \left(\ln g_i^{(t)} + \ln g_{j/i}^{(t)} + \ln P_{ji}(y^{(t)}) \right) \quad 28$$

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:

$$E(\theta, \theta^{(p)}) = E(L(\theta, Z) / X) = \sum_{i=1}^{K} \sum_{j=1}^{L} z_{ij}^{(t)} \left(\ln g_i^{(t)} + \ln g_{j/i}^{(t)} + \ln P_{ji}(y^{(t)}) \right) \quad 29$$

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

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:

$$w_{ji}^{(p+1)} = \arg\max_{w_{ji}} \sum_{t=1}^{N} h_{ji}^{(t)} \ln P_{ij}(y^{(t)})$$

$$v_i^{(p+1)} = \arg\max_{v_i} \sum_{t=1}^{N} \sum_{k=1}^{K} h_k^{(t)} \ln g_k^{(t)}$$

$$v_{ji}^{(p+1)} = \arg\max_{v_{ji}} \sum_{t=1}^{N} \sum_{k=1}^{K} h_k^{(t)} \sum_{l=1}^{L} h_{l/k}^{(t)} \ln g_{l/k}^{(t)}$$

$$\theta^{(p+1)} = \operatorname{argmax}_{\theta} E(\theta, \theta^{(p)})$$

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).

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, η_c, como sigue: $\| x(t) - w_c \| = \min \min_i^N \{\| x(t) - w_i \|\}$ 30

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

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

$$h_{ci}(t) = \exp\left(\frac{\left\|r_c - r_i\right\|^2}{2\sigma^2(t)}\right) \qquad 31$$

Donde:

$h_{ci}(t)$ 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:

$$\sigma(t) = \left\{ \begin{array}{ll} R & para\ t < t_g \\ R(1 - \frac{t}{t_{\max}}) & para\ t \geq t_g \end{array} \right\}, \qquad 32$$

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.

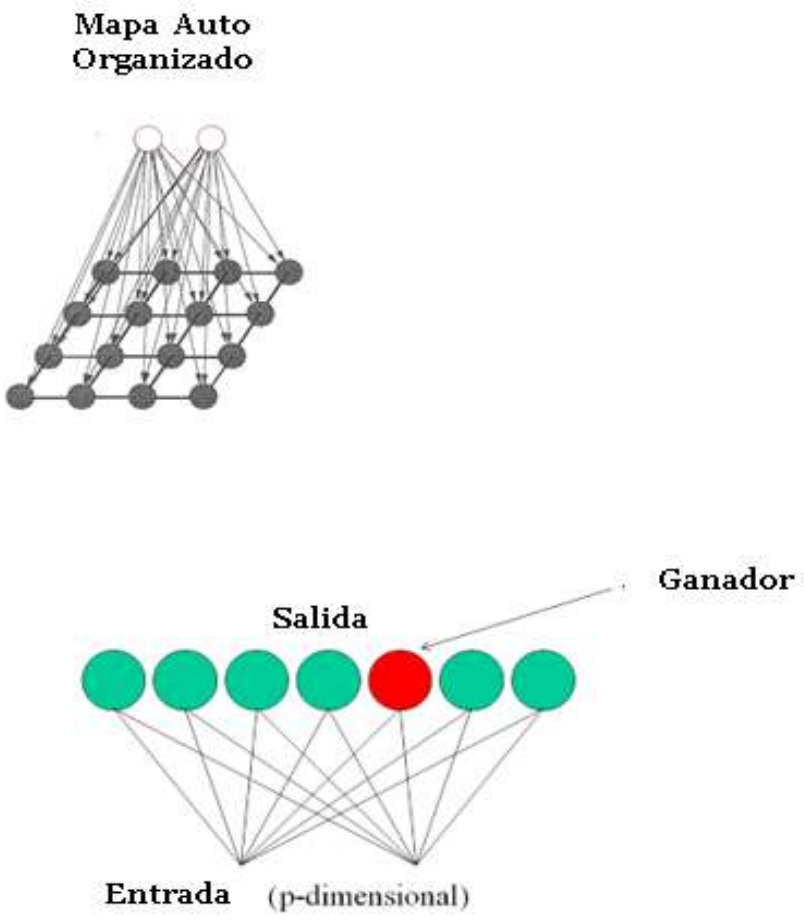
Factor de Aprendizaje. La función alfa es la responsable de garantizar la convergencia del proceso de entrenamiento:

$$\alpha(t) = \begin{Bmatrix} \alpha_{\max}, t < t_g \\ \alpha_{\min}, t \geq t_g \end{Bmatrix} \quad 33$$

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 10

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 11

Arquitectura de las redes de Kohonen

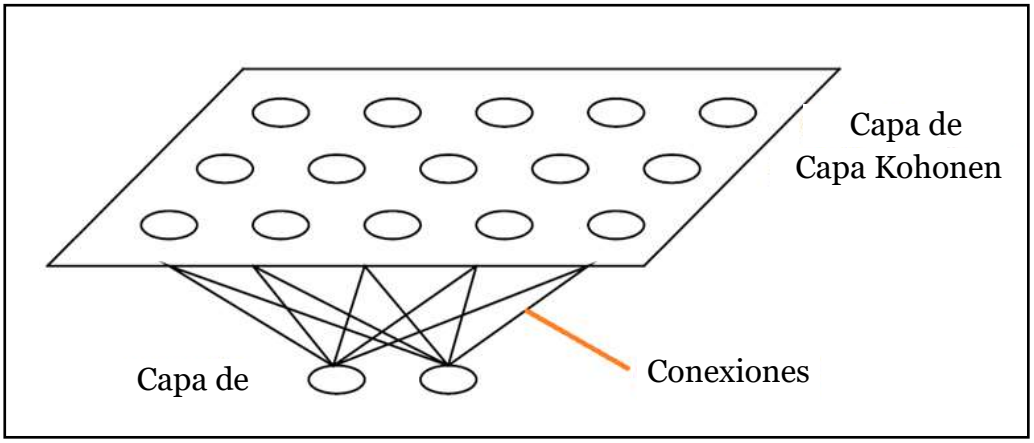
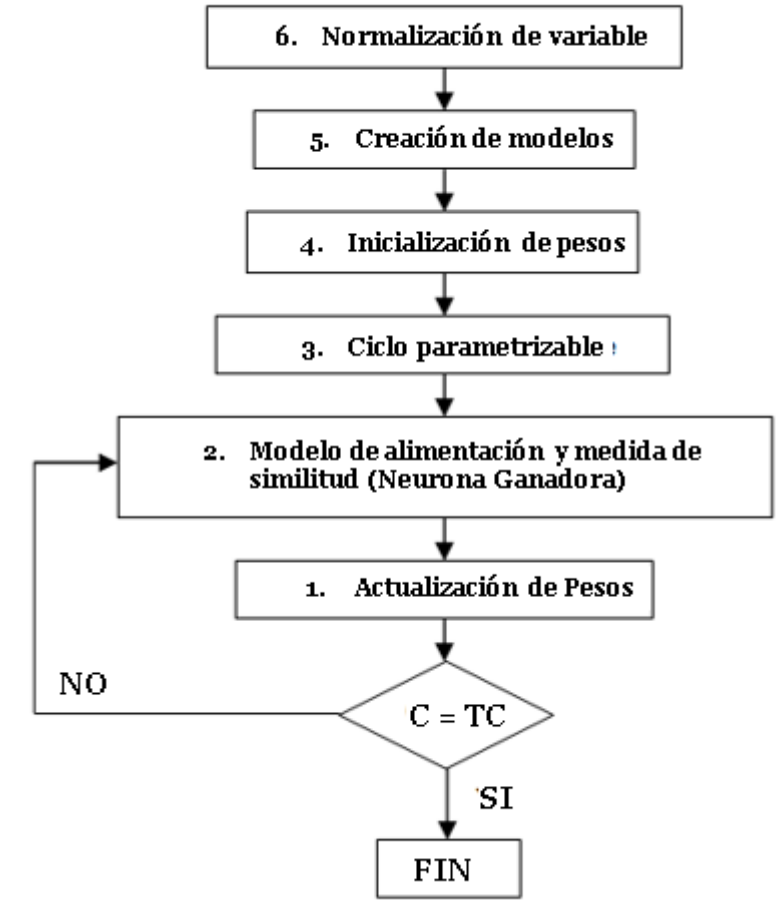


Figura 12

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.

$$W_j(t+1) = W_j(t) + \eta(t)h_j(t)(X - W_j(t)) \qquad 34$$

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:

$$h_j = \exp\left(-\frac{\left\|u_j - u_j^*\right\|^2}{2\sigma^2}\right) \qquad 35$$

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.

Análisis clúster k-means.

El análisis clúster es un método utilizado para la detección de grupos homogéneos dentro de una muestra de estudio. Los algoritmos clúster *k-means* convencionales permiten clasificar cada elemento de un conjunto de *n* elementos exclusivamente en un grupo de los *K* preestablecidos, considerándose la exclusividad de pertenencia. Una vez se ha fijado el número de clústeres deseados, la determinación de centroides busca minimizar la dispersión de los elementos dentro de un grupo como:

$$\underset{c_k, u_{i,k}}{Min} \quad \sum_{k=1}^{K} \sum_{i=1}^{n} u_{i,k} ||x_i - c_k||^2 \qquad 36$$

sujeto a $\sum_{k=1}^{K} u_{i,k} = 1, u_{i,k} \in \{0,1\}$

Siendo $||\cdot||$ una norma normalmente euclídea, x_i la observación del i-ésimo elemento sobre un conjunto de características, c_k el *prototipo* o *centroide* del clúster k y $u_{i,k}$ el nivel de pertenencia del elemento i al clúster k, $u_{i,k} \in \{0,1\}$. Así, $u_{i,k} = 1$ si el elemento i-ésimo es clasificado en el grupo k y $u_{i,k} = 0$ en caso contrario.

La modelización fuzzy *k-means* relaja la condición de exclusividad de pertenencia a un grupo y supone que cualquier elemento puede pertenecer en un cierto grado delimitado en el intervalo [0, 1], a más de uno de los grupos prefijados. Así, el programa de optimización anterior queda re-expresado como (37):

$$\underset{c_k,u_{i,k}}{Min} \quad J = \sum_{k=1}^{K}\sum_{i=1}^{n}\left(u_{i,k}\right)^m ||x_i - c_k||^2 \quad 37$$

$$\text{sujeto a } \sum_{k=1}^{K} u_{i,k} = 1, 0 \leq u_{i,k} \leq 1$$

donde el parámetro m es denominado en la literatura como *fuzzificador* y puede tomar valores $1 \leq m < \infty$. Cuando $m \to 1$, la partición resultante es una convencional (nítida) mientras que si $m \to \infty$ el análisis clúster no añade valor ya que $u_{i,k} \to \frac{1}{K}$. Tal como indican Klawonn et al. (2015), el análisis clúster convencional puede ser entendido como un caso particular del análisis clúster borroso ya que los grupos borrosos pueden ser transformados en grupos nítidos considerando que cualquier observación x_i pertenece exclusivamente al clúster s si $u_{i,s} = max_{k=1}^{K} u_{i,k}$.

En problemas de clasificación relativos a las Ciencias Sociales, la definición de las clases suele ser difusa, de tal forma que muchos de los elementos a clasificar pueden participar de características de más de un grupo. Así, si se establecen dos grupos en función de la eficiencia en *RAP* "alta eficiencia en RAP" y "baja eficiencia en RAP", la misma naturaleza con la que hemos etiquetado estas clases económicas nos lleva a considerar el fuzzy clustering como una alternativa interesante al hard clustering. Así, aceptaríamos que la pertenencia de Luxemburgo al primer grupo y Grecia el segundo es inequívoca, pero Dinamarca, con un índice de eficiencia en *RAP* de 0,99, muy cercano a 1, participaría, en cierto grado, de ambos grupos. El análisis fuzzy clustering ha sido ampliamente utilizado en problemas de clasificación de tipo económico y social. Así, Derrig y Ostaszewski (1995) clasifican varios municipios del estado de Massachussets en

función de variables relacionadas con el fraude en seguros de automóvil; Yu et al. (2012) y Yu et al. (2014) clasifican las provincias chinas en función del cumplimiento de los objetivos de emisiones en CO^2 y Wu et al. (2013) categorizan las regiones chinas en función de la vulnerabilidad de su producción agrícola. En Irán, en el ámbito de la gestión empresarial, clustering para discriminar instituciones financieras en Rumania. Se introduce una nueva razón de tipo algorítmico para emplear métodos de fuzzy clustering. Indican que el clustering borroso en muchas ocasiones evita los problemas algorítmicos de los métodos convencionales clúster k-means como, por ejemplo, que sus resultados dependen fuertemente de la inicialización del algoritmo. Además, fuzzy clustering no es únicamente una simple mejora de los algoritmos convencional k-clustering sino que abre la posibilidad de realizar análisis clúster más flexibles y sofisticados que los que se pueden hacer con el análisis clúster nítido. La obtención de los centroides y los niveles de pertenencia asociados a los grupos puede ser resuelto con el algoritmo propuesto en Bezdek (1981). Éste se basa en la aplicación recurrente de las siguientes ecuaciones. La primera estipula que el centroides del clúster k-ésimo es la media ponderada:

$$c_k = \frac{\sum_{i=1}^{n} (u_{i,k})^m x_i}{\sum_{i=1}^{n} (u_{i,k})^m} \qquad 38$$

La segunda ecuación estipula que el nivel de pertenencia del elemento *i*-ésimo al k-ésimo clúster, $u_{i,k}$, se encuentra como:

$$u_{i,k} = \left[\sum_{s=1}^{K} \left(\frac{||x_i - c_k||^2}{||x_i - c_s||^2} \right)^{\frac{1}{m-1}} \right]^{-1} \qquad 39$$

Así, los pasos que sigue el algoritmo de Bezdeck son los siguientes:

Paso 1: A partir de un número predefinido del número de clústeres, K y de m, escoge un nivel inicial de pertenencia de cada elemento a cada clúster $u_{i,k}^{(0)}$, de tal manera que $\sum_{k=1}^{K} u_{i,k}^{(0)} = 1$.

Paso 2: Con (8c) calcula el centroide del clúster k-ésimo de la primera iteración, $c_k^{(0)}$, k=1,2,...,K.

Paso 3: Con (8d) calcula el nivel de pertenencia de cada elemento a cada clúster. En esta primera iteración obtenemos $u_{i,k}^{(1)}$ $i = 1,2, \ldots, n$; $k = 1,2, \ldots, K$.

Paso 4: Calcula el valor de la función de coste en (38) con $u_{i,k}^{(1)}$. Si se produce en su valor una mejora respecto al asociado a $u_{i,k}^{(0)}$ en una cuantía superior a un valor predefinido ε, repite los tres pasos anteriores a partir del valor $u_{i,k}^{(1)}$ para encontrar $u_{i,k}^{(2)}$ e implementa este paso. Así, en la iteración i-ésima, si a partir de $u_{i,k}^{(i)}$ se produce una mejora en la función objetiva respecto a $u_{i,k}^{(i-1)}$ superior al valor predefinido ε, se vuelven a implementar los cuatro pasos descritos. El algoritmo finaliza en el momento en que la iteración no proporciona una mejora en la función de coste superior a ε.

Una cuestión relevante en el análisis fuzzy k-means es la elección del número óptimo de clústeres *K*. Esto ha generado una extensa literatura. Aquí se utiliza dos índices de validación del número de clústeres que parten de una filosofía diferente. El primero únicamente utiliza los niveles de pertenencia y es un refinamiento denominado índice como *IDB* y se calcula como:

$$IDB = 1 - \frac{K}{K-1}\left(1 - \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(u_{i,k}\right)^2\right) \qquad 40$$

El número óptimo de clústeres es aquel valor de que *K* maximice *IDB* y, básicamente, será aquella partición que se acerca a una partición *hard*.

El segundo índice que se emplea es más completo ya que, aparte de los niveles de pertenencia, utiliza el valor de las observaciones. El valor óptimo de *K* debe minimizar la dispersión dentro de las clases y, al mismo tiempo, maximizar la dispersión entre clases. Si denominado índice como *IXB*, el *K* óptimo deberá minimizarlo, ya que:

$$IXB = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n}\left(u_{i,k}\right)^m ||x_i - c_k||^2}{n \cdot min_{j,k}||c_j - c_k||^2} \qquad 41$$

1.2.2 Criterios de desempeño de los modelos de predicción de Minería de Datos

Las medidas de asociación global y local se obtienen de la tabla de contingencia con presencia de asociación y su intensidad, evalúan el desempeño de las técnicas de predicción de minería de datos se obtienen a partir de una matriz de confusión (MC) la cual, describe el conteo de los verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN). Las filas representan el número de muestras en la clase observada y las columnas el número de predicciones de cada clase. La diagonal de MC corresponde al número de muestras que el algoritmo clasifica correctamente en cada clase. Si MC solo tiene valores positivos en la diagonal, indica que el clasificador clasifica correctamente todas las muestras. La métrica precisión global de clasificación (PG) mide la proporción global de muestras bien clasificadas en cada clase y se calcula como:

Verdadero positivo	Falso positivo (FP)
Falso negativo (FN)	Verdadero negativo

Figura 13: Matriz de confusión

$$Precisión\ (PG) = \frac{VP+VN}{FP+FN+VP+VN} \qquad 42$$

Las métricas para medir el desempeño del clasificador en cada clase son precisión (P), sensibilidad (S), especificidad (E) y puntaje $F1$. Se definen con las siguientes expresiones:

a) **Valor predictivo positivo**: conocido también como Valor Predictivo Positivo, es definida como la proporción de verdaderos positivos respecto al total de pruebas con resultado positivo, es la probabilidad de retirase si el caso es expuesto:

$$\text{VP+} = \frac{VP}{VP+FP} \qquad 43$$

b) **Sensibilidad:** es la probabilidad de que la prueba dé positiva si el caso está presente. También se define como la proporción de verdaderos positivos respecto al total de casos

$$Sensibilidad\ (S) = \frac{VP}{FN+VP} \qquad 44$$

c) **Especificidad:** es la probabilidad de que la prueba dé negativa si retirado está ausente. También se define como la proporción de verdaderos negativos respecto al total de sujetos no retirados:

$$\text{Especificidad } (E) = \frac{VN}{VN+FP} \qquad 45$$

d) Odds Ratio (OR): define la posibilidad de que una condición del caso (deserción) se presente en un grupo de población frente al riesgo de que ocurra en otro:

$$\text{Odds Ratio } (OR) = \frac{VN+FP}{VP+FN} \qquad 46$$

e) Valor Predictivo Negativo (VPN): Es la probabilidad de no retirarse si el caso es no expuesto. También definida como la proporción de verdaderos negativos respecto al total de pruebas con resultado negativo:

$$\text{VP}-= \frac{VN}{VN+FN} \qquad 47$$

f) Razón de Verosimilitud: se define como la razón entre la posibilidad de observar un resultado en los pacientes con la enfermedad en cuestión versus la posibilidad de ese resultado en pacientes sin la patología.

g) Índice Kappa: La estadística Kappa compara el nivel de concordancia esperado por azar.

h) Riesgo Relativo: es la probabilidad de que ocurra un evento en el grupo expuesto y la probabilidad de que el mismo evento ocurra en el grupo no expuesto

$$Riesgo\ Relativo\ (RR) = \frac{VP(VN+FN)}{VN(VP+FP)} \qquad 48$$

i) Puntaje: Se resume como precisión y sensibilidad en una sola métrica, es un estimador apropiado en clases desbalanceadas y varía entre cero y uno.

$$Puntaje\ (F1) = 2\frac{P\times S}{P+S} \qquad 49$$

j) Prevalencia

$$Prevalencia\ (Pr) = \frac{VP+FN}{VP+FP+FN+VN} \qquad 50$$

La curva *receiver operating characteristics* (ROC) es una curva que relaciona valores de *S* versus 1-*E*. Los diferentes puntos en la curva corresponden a los puntos de corte utilizados para determinar si los resultados de la prueba son positivos. El valor de AUC_{ROC} (área bajo la curva ROC) se interpreta como la probabilidad de que, en dos muestras, una positiva y una negativa, la prueba asigne una probabilidad más alta a la muestra positiva, clasificación correcta (Mandrekar, 2010). Su valor oscila entre cero y uno; cuanto mayor es AUC_{ROC} mejor es la clasificación, un valor cercano a 0.50 indica una mala clasificación. La curva *P-S* es el resultado de graficar *P versus S*. Ésta permite observar a partir de qué *S* se tiene una degradación de *P* y viceversa. El resultado ideal es una curva que se acerque a la esquina superior derecha (alta *P* y S), lo que genera un área bajo la curva $AUC_{P\text{-}S}$, que, entre más cercano a uno, es mejor el modelo (Saito y Rehmsmeier, 2015).

1.2.3 Factores Asociados a la Deserción Estudiantil

Deserción estudiantil es un término comúnmente utilizado en todo el mundo para referirse al abandono escolar. Se trata de aquella situación en la que el alumno después de un proceso acumulativo de separación o retiro, comienza a retirarse antes de la hora establecida por el sistema educativo sin obtener una nota o un certificado escolar (Lyche, 2010).

La deserción estudiantil en las Instituciones Educativas de nivel Superior es un problema que desde los inicios de funcionamiento se ha presentado y que a la fecha no se ha abordado 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 estos índices.

En un caso se ha evaluado la cantidad de alumnos matriculados en los 6 semestres académicos de las 02 carreras profesionales durante los último 3 años y se conoce que actualmente se tiene un 34 % de deserción estudiantil 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.

Jara Tuesta (2017) plantea diversas teorías y definiciones de factores asociados a la deserción estudiantil, como:

Teoría humanista: enfoca al hombre en el tránsito de su vida busca la realización y la felicidad, basado en el empirismo como principio filosófico que sostiene lo aprendido a partir de sus propias experiencias del existencialismo, "desde su interior", siendo independiente e íntegro el hombre un ser universal, que va desarrollándose a través de la toma de decisiones, como ser activo y persistente en su aprendizaje y evolución.

El coaching ontológico: es una técnica de cambio del pensamiento y comportamiento interno, brinda la posibilidad de explicar la forma de conseguir los objetivos en el ámbito social y/o empresarial como "obstáculos cuándo pensamos que nuestras captaciones sobre la realidad son propias y no pertenecen a los elementos perturbadores del entorno", el coaching ontológico tiene tres premisas básicas: "Los seres humanos seres lingüísticos, el lenguaje es generativo de realidades en nuestro entorno y los seres humanos se crean así mismo en el lenguaje" (Echevarría, 2003). Las personas se comportan de acuerdo al tipo de sociedad en la que viven, pretender cambiar dicho tipo de sociedad depende de las acciones que se realicen y de las decisiones que tomen en su vida, no es persuasión de una persona a que cambie su forma de pensar y/o actuar solo porque otro individuo con más jerarquía considere que sus pensamientos o acciones son erróneos.

La psicología positiva: busca rescatar el aspecto más positivo del ser humano, desde el punto de vista emocional, plantea que el dinero ayuda a cubrir las necesidades básicas, brinda tranquilidad, pero no ayuda a alcanzar el gran anhelo que tiene el ser humano de alcanzar "la felicidad", la tranquilidad y satisfacción, alejando de su vida la incertidumbre y la insatisfacción, la realización personal no es una percepción económica, familiar y laboral o académico, sino, su auto realización y la motivación más grande radica en su voluntad para cumplir proyectos en cada etapa de la vida.

Teorías sobre la elección de carrera y la deserción: En el análisis de Jara Tuesta (2017), los jóvenes alcanzan mayor madurez en la última etapa de su adolescencia es decir entre los 17 y 18 años, donde son responsables para enfrentar su futuro profesional. Además, Ginzberg et al. (1951) estableció que

muchos jóvenes ingresan a la universidad sin haber definido bien lo que desean estudiar, posiblemente por influencias externas ya sean por los amigos o familiares, conllevándoles a obtener grandes dudas sobre la elección hecha y el temor de haberse equivocado de carrera; por lo que se recomienda hacer test vocacionales y asesoría constante a los estudiantes que les permitan elegir correctamente la carrera profesional que desean estudiar.

Teorías sobre la elección de carrera: es importante la decisión que va a tomar un joven sobre su futuro profesional, pues es la actividad que va realizar por el resto de su vida, por consiguiente, es necesario que haga una buena elección, sin embargo, eso no se logra en un momento determinado, sino se va formando a la persona desde muy pequeño, se le va perfilando de acuerdo a las características que presente o las áreas académicas que más le agraden y pudiera sentirse satisfecho realizarlas y así, ser un profesional de éxito (Grinder, 2001). La orientación vocacional tiene un proceso largo de maduración y va de acuerdo con la edad de la persona, incluso cuando se hallan hecho profesional, todavía existe un periodo de afianzamiento o determinación sobre lo que han escogido para consolidarse como profesionales de vocación y con un alto sentido ético y amor a su carrera (Super, 1953). La teoría de Ginzberg et al. (1951) hace énfasis a "la idea de que el desarrollo de la personalidad está ligado a las experiencias de la infancia y también describe el desarrollo de las carreras como un evento cuya secuencia es predecible y cada evento enfrenta al individuo con un conjunto de problemas que debe resolver" el proceso de desarrollo vocacional se da en dos etapas, la primera se caracteriza por las diversas oportunidades que se le presenta al joven y la segunda por ser una decisión concreta más individualista.

La teoría de la influencia familiar en las decisiones del adolescente: la naturaleza de inseguridades y confusiones debido a los cambios fisiológicos y psicológicos que experimenta el adolescente, los padres determinan muchas decisiones sobre sus hijos, en muchos casos ingresan a estudios superiores por encargo de los padres, sin haber hecho una adecuada elección, luego abandonan sus estudios, éstos factores afectan más que los factores económicos, sociales, culturales, etc., mientras un estudiante ingresa a la universidad inmediatamente saliendo del colegio, sin haber pasado por un test y orientación vocacional, se confunde y está inmerso en un mundo totalmente diferente, decide abandonar

los estudios, en los primeros ciclos académicos, entonces la elección de una carrera profesional es una de las decisiones importantes que un adolescente o joven toma en su vida, si no supo elegir bien la carrera a estudiar, en el futuro trabajará en algún oficio que no sea de su agrado. El joven encontrándose aún en la época de rebeldía optará en ocasiones rechazar las reglas establecidas por las instituciones educativa, faltando reglas de convivencia o al no prestar atención a las clases, desaprueba los cursos y por ende tiene que desertar. Por otro lado, las habilidades y conocimientos que consideran indispensables para realizar estudios superiores con éxito, no se desarrollan de manera homogénea para todos los estudiantes, en algunos casos se discriminan causando la deserción.

Las teorías sociales del abandono estudiantil: refieren a los factores socioeconómicos asociados a la educación superior, donde el factor económico determina más que los demás, los jóvenes alcanzan ser profesionales y cumplen su sueño, solventando el costo que implica los estudios sin problemas, también es importante el factor social, orientado a centros de estudio para diversos estratos sociales, donde la calidad de enseñanza es diferenciada y hasta discriminada, también influye el nivel cultural de la familia, si en nivel cultural de padres es bajo, las aspiraciones de los hijos a desarrollarse como profesional también son mínimas, donde la motivación o paradigma de seguir estudios superiores, depende de la formación, profesión y desarrollo social de los padres, la historia de la familia, la historia escolar de los miembros de la misma y perfil de las instituciones educativas, responden al interés de estratificación social, según las posibilidades económicas de los padres de familia direccionados ya a continuar sus estudios superiores en universidades e institutos con las mismas características. En cuanto a los factores externos de deserción son los diferentes niveles de preparación durante la secundaria y condición de estudiante en la misma, marca perfiles de estudiantes que si fueron a una buena escuela con todas las oportunidades para desarrollar sus capacidades y los que no tuvieron esa oportunidad. el origen social ha sido estudiado ampliamente. Numerosos teóricos debaten sobre los factores como el sistema educativo dominante, el origen social, cultural, la familia y su historia escolar, la actitud hacia el éxito y entre otros influyen a la deserción estudiantil ya sean de mayor o menor grado, y difíciles de controlar directamente por las instituciones educativas.

Teorías pedagógicas y la deserción: es importante lo que plantea Yopo (1997) de los cuatro tipos de escuela y explica la característica e influencia de cada una de ellas en el estudiante y la sociedad, pasando desde la educación tradicional hasta la educación actual que es más horizontal, busca el desarrollo de capacidades que le permitan al estudiante no solo adquirir conocimientos, sino estar en la capacidad de resolver situaciones problemáticas de su vida diaria, estar preparado para la vida, la vida real es resolver situaciones del día a día, esa es la función de la nueva escuela, de la nueva universidad. Adicionalmente, considera tres corrientes tales como "la escuela conservadora (tradicional), la escuela progresiva (modernista o desarrollista) y la escuela reconstruccionista (revolucionaria, de dónde se origina la nueva universidad)". Desde este punto de vista se afirma que el alumno se encuentra con menor intención de abandonar los estudios, por la empatía que encuentra en el maestro y ve las asignaturas académicas con mayor agrado, estando pre dispuesto a aprender. Por su parte, Álvarez (1995) señala que las teorías pedagógicas actividad comunicación y proceso, donde, la comunicación es fundamental en el proceso de enseñanza-aprendizaje y el estudiante es el actor principal de este proceso que logra desarrollar sus capacidades incluso desconocidas, sobre una base de la teoría de los procesos, que es un método de actuación que asume el individuo hasta lograr sus metas importantes de su propio aprendizaje, en el sentido que "hace", es protagonista de la transformación, sin perder el objetivo final de una clase ni completado, siendo realizado y satisfecho, es trascendental la propuesta del cambio educativo, siendo ahora más práctica que teórica, el estudiante debe estar en la capacidad de desarrollar habilidades para sociedad sin problemas. Por lo tanto, es responsabilidad de los docentes encontrar la metodología adecuada con una currícula flexible que se adapte a las nuevas expectativas de formar profesionales activos, probos y útiles a la sociedad.

Modelos Psicológicos: en este ámbito el abandono estudiantil se enlaza con los preceptos psicológicos de la persistencia educativa, es clave realizar la motivación del estudio, generar el interés y sobretodo la personalidad del individuo para no desertar. Ya que, la deserción estudiantil es básicamente la inmadurez y la rebeldía formada en la casa, es la dificultad manifiesta de los individuos en su capacidad de afrontar nuevos retos y metas en la vida estudiantil o el fracaso personal, el cual se debe a dos razones: primero por una falta de

preparación en la educación secundaria y la otra por la falta de una adecuada prueba o test psicológicos, En este marco, la psicología considera a la clase social como un generador de problemas. Ya que, el estar en una determinada clase social predispone al individuo en su conducta de vida (como sus modales, formas de hablar y de vestir), así también sus opiniones. Reconocer que a pesar que se imparte los mismos lineamientos de educación, lamentablemente el modo como se imparte marca la diferencia. Obviamente una educación impartida a un grupo que cuenta con todos los recursos económicos que dan acceso a una mejor alimentación y materiales, no será igual que una educación brindada a un grupo con escasos recursos. Todas estas diferencias se asocian a la deserción. Al respecto Contasi y Vidal (1970), señalan que la diferencia es marcada entre estudiar en una institución privada o pública, así la política educativa sea la misma, los factores como social, económico, cultural e incluso psicológico determinan dichas diferencias. Muchas instituciones públicas carecen de la infraestructura y los recursos básicos para brindar un servicio de calidad o al menos aceptable a su comunidad, mientras las instituciones privadas brindan todo lo necesario para cumplir con los objetivos plasmados en su visión y misión institucional.

Equidad en la educación: De acuerdo a lo establecido en el artículo 100.7 de la Ley universitaria N° 30220, explica "Tener en las universidades privadas, la posibilidad de acceder a escalas de pago diferenciadas, previo estudio de la situación económica y del rendimiento académico del alumno". Según la ley mencionada todas las universidades deben crear disposiciones con una escala de pensiones que favorezcan al estudiante para que pueda continuar sus estudios universitarios y disminuir la deserción provocada por los factores económicos. Asimismo, en la Política de aseguramiento de la calidad de la Educación Superior Universitaria (D.S. N° 016-2015-MINEDU), considera "Inclusión y equidad. Todos los actores involucrados en el Sistema Universitario promueven y garantizan el acceso, permanencia y culminación satisfactoria de los estudios universitarios a todos los jóvenes del país, sin distinción de lengua, etnia, religión, sexo u otra causa de discriminación; y con especial énfasis en las personas con discapacidad, grupos sociales excluidos, marginados y vulnerables, especialmente en el ámbito rural" (pág. 31). La equidad es el derecho de todo ser humano de contar con una educación igualitaria en todos los aspectos, tratando de permanecer en ella y poder concluirla satisfactoriamente. Este concepto está

respaldado por la “Declaración Universal de Derechos Humanos” de 1948 (Artículo 26), en su manifiesto propone lograr reponer la igualdad, a través de acciones que modifiquen o mejores las actitudes de las personas. Por otro lado, la “Declaración Mundial sobre Educación para todos”, aprobada por los Ministros de Educación de todos los Estados miembros de la UNESCO en Jomtien, Tailandia, en 1990, motivada por la Declaración de Derechos Universales confirma que la educación “es un derecho para todas las personas, hombres y mujeres, de todas las edades, a través de todo el mundo”.

Concepto de deserción estudiantil: Los argumentos teóricos que aclaren deserción estudiantil son varios para contar con una definición acertada y consolidada sobre los modelos de deserción existentes hasta el momento de la publicación, el glosario de la Red Iberoamericana para la acreditación de la calidad de la Educación Superior (RIACES) plantea la idea de deserción como un sinónimo de abandono, al que se refiere como una mortalidad escolar, la misma que precisa como un estudiante que suspende, repite, cambia de carrera, o abandona antes de obtener el título . asimismo, la deserción es concebida como la interrupción definitiva o temporal, voluntaria o forzada, las cuales son totalmente diferentes de otras formas de deserción tales como: abandono de la carrera, abandono de la institución y el abandono del sistema de educación superior (Romo y Hernández, 2005). De acuerdo al autor existe otra terminología acerca de la deserción estudiantil denominada “Mortalidad escolar”, quienes se encuentran en este nivel, son aquellas personas que no terminaron sus estudios sea de educación básica o estudios superiores. Así mismo, otros autores coinciden en relacionar la deserción con el abandono de la vida académica por diversos factores.

Definición de tres autores acerca de la deserción estudiantil: Castaño, (2004), en su investigación sobre la deserción “considera como conjuntos de factores que van a ser concluyentes en la deserción estudiantil. Siendo el primero el individual, que va a agrupar las características demográficas del estudiante. En seguida es el factor académico, el que tiene relación con la educación que a la larga recibe el estudiante y sobre todo con la orientación que se le brinda, además de su grado de aprovechamiento en su desarrollo académico”. A continuación, propone el factor socioeconómico, el cual tiene por objeto de observación la vida

laboral del estudiante y su familia, teniendo en cuenta el grado de dependencia económica y el incremento del poder adquisitivo de la familia. El último factor considerado por el investigador es el institucional, que tiene relación con los procesos de socialización del individuo con su nuevo entorno.

De acuerdo a lo planteado por el autor existen cuatro factores que determinan la deserción estudiantil que son: Individual; se sustenta en la indecisión o falta de madurez por parte del estudiante, en poder decidirse qué carrera profesional deberá continuar. Académico; que se sustenta en la preparación que haya tenido durante su educación básica regular, donde no ha tenido una adecuada base académica por lo tanto no se encuentra en el nivel de competencia frente a los demás estudiantes. Socioeconómico; donde el estudiante tiene la capacidad y predisposición para seguir estudiando con éxito, pero, debido a las circunstancias económicas se ve obligado a desertar causando una gran frustración en el individuo y finalmente, Institucional; donde muchas veces el individuo no se adapta a nueva etapa de formación profesional y deserta. Por la diversidad de grupos socioculturales de los jóvenes, el bajo nivel adquisitivo y la frustración por no poder alcanzar la carrera de su interés, llegar a traer consigo una crisis vocacional que a la larga llevara al estudiante a una deserción. Asociado a las características ya mencionadas se considera también que los estudiantes tendrán un bajo rendimiento académico, no se sentirán identificados con lo que están estudiando, desencadenando con el tiempo una deserción.

Efectos de la deserción escolar: En su gran mayoría los países latinoamericanos son los que sufren más los efectos de la deserción estudiantil, tanto en el aspecto social como en el aspecto individual, lo que agrava su situación económica. Como se sabe que los individuos dejan de estudiar son poco productivos, menos competentes y con mayores problemas evaluados en distintas áreas del mundo laboral. Se afirma que aquellas personas que no terminan los estudios causan un déficit en las economías de sus países. Es también muy marcada la diferencia de clases cuando existen grupos grandes de deserción, a consecuencia de esto, se observa aquellos individuos que no concluyen sus estudios solo tienen acceso a empleos sobrevalorados, mientras que aquellos que concluyeron sus estudios tienen mejores oportunidades de trabajo. Estas deserciones van a marcar el rumbo de una sociedad, que tendrá un alto grado de

desigualdad que se reflejará en la sociedad y la economía. Pero también la deserción causa en el individuo una sobrevaloración de su trabajo, es así, que existen trabajadores que no tienen un poder adquisitivo bueno, sino más bien se crean marcadas diferencias. Lo que con el tiempo se traduce en un alto índice de delincuencia, drogadicción y miseria con personas sin el propósito de superarse. Según Cárdenas (2000), la deserción estudiantil trae como consecuencia una serie de problemas sociales, en principio por carecer de mano de obra calificada por falta de preparación académica, donde el nivel cultural y laboral del país es pobre rezagado en comparación a otros países. También dice el autor que una persona con mejor y mayor preparación académica tendrá mayores oportunidades de trabajo y con una buena remuneración permitiéndole escalar socialmente. En caso contrario, es vulnerable a los peligros de la sociedad con muchos problemas sociales, siendo el mayor impacto la corrupción y la falta de empleo, la drogadicción y la delincuencia.

Factores de la deserción estudiantil:

Para Jara Tuesta (2017), la clasificación de los factores según los tipos de variables es, individual, el poder adquisitivo y el nivel social considerado como factores socioeconómicos, las características de la vida escolar del estudiante y su desempeño, son los factores académicos, y el centro de estudios refiere a los factores institucionales. Otra clasificación es cinco los factores que influyen en la deserción del estudiante como son: Psicológico, que tiene que ver prácticamente con el aspecto personal del individuo y su situación emocional; Sociológico el contexto en el que se desenvuelve le individuo y la influencia que esta ejerza sobre él. Económico no contar con una fuente de ingreso permanente y la falta de apoyo económico por parte de los familiares influye en que el individuo se vea en la obligación de dejar de estudiar. Organizacional, el no mantener un orden o disciplina en sus actividades particulares hace que se sature de trabajo y/o actividades que conllevan a que deje de estudiar. En el aspecto. Interaccionista, si el individuo es introvertido y no interactúa con las personas de su entorno, no comparte los mismos objetivos, entonces se verá fuera de contexto y por ende abandonará los estudios.

Consecuencias de la deserción escolar: Para Ruiz, García, Ruiz-Ramírez, García-Cué y Pérez-Olvera (2014) la deserción escolar tiene como efecto

importante en el embarazo prematuro, donde tienen que afrontar nuevas responsabilidades para las cuales no están preparados. También influye a la violencia familiar que hace que los hijos abandonen su casa cansados de tantos maltratos y se exponen a los peligros de la sociedad, carente de valores y peligros constantes. En consecuencia, es una sociedad con un nivel cultural muy bajo que impide salir del sub desarrollo.

Una de las razones por las cuales los ingresos económicos son bajos es por la falta de preparación académica. En muchos casos cuando se va a solicitar empleo lo mínimo que piden es tu nivel de estudios, si no tiene la secundaria completa, el trabajo que encuentre es mínimo, por ende, la remuneración es baja; ante la necesidad de que los padres piden apoyo a los hijos y los envían a trabajar dando prioridad a las necesidades básicas del hogar, dejando de lado los estudios, dándose a sí la deserción y el nivel cultural de la sociedad no cambia.

Es preciso mencionar lo que dice el autor que el peligro más latente de los desertores es caer en la drogadicción y la delincuencia, esto debido a los factores antes mencionados coincidiendo con lo que determinó la Organización

Los factores institucionales: se define como factores institucionales a las características estructurales y funcionales que difieren en cada institución, son las propiedades que hacen de una universidad única y diferente del resto y su grado de influencia confiere a la universidad peculiaridades propias. Dentro de los factores institucionales se menciona la infraestructura, los horarios, la información referente a las escuelas, la matricula, entre otros.

Los factores personales: son el conjunto de actitudes que cambian de forma consciente y que están íntimamente ligados al aspecto emocional, familiar y social. en el plano educativo considerar la situación familiar de los estudiantes será determinante, la situación económica es determinante en la continuidad de los estudios.

Los factores académicos: implica tener en consideración el rendimiento académico, que es la confluencia de diversos criterios tales como pedagógico – didáctico que con lleva a relacionar el logro de los estudiantes con los planes de estudios a adecuados a su realidad y a la carrera elegida, los estilos de enseñanza de los docentes, identificación con la carrera elegida. Todos los antes

mencionados determinaran el rendimiento académico de los estudiantes universitarios.

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.

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.

A. 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 estos 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.

CAPITULO II

CARACTERIZACIÓN DEL PROBLEMA Y MARCO METODOLÓGICO

2.1 Descripción del problema

Los Institutos técnicos a nivel nacional son alternativas 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.

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.

Por lo expuesto la presente investigación tiene como objetivo resolver la siguiente problemática: ¿Cuáles son los Segmentación de alumnos con riesgo deserción mediante las técnicas de la minería de datos?

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).

Enunciado general

¿Cuáles son las técnicas de Machine Learning y Minería de Datos que determinan los factores asociados para segmentar los alumnos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019?

Enunciados específicos

- ¿Cuáles son las técnicas de la minería de datos que reducen la dimensionalidad de los factores asociados a la deserción estudiantil en el Instituto Superior Tecnológico Privado ISTEPSA?
- ¿Cuáles son las técnicas no supervisadas de aprendizaje automático que segmentan mejor los alumnos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA?
- ¿Cuál es la técnica no supervisada de Asociación para encontrar la forma de relaciones de implicación asociados a la deserción estudiantil en el Instituto Superior Tecnológico Privado ISTEPSA?
- ¿Cuáles son los parámetros de los factores asociados en los segmentos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA?

Hipótesis general

Las técnicas de selección de factores, asociación y clustering de *Machine Learning* y Minería de Datos determinan los factores académicos e institucionales como más significativos para segmentar los alumnos con riesgo de deserción del Instituto Superior Tecnológico Privado ISTEPSA.

Hipótesis específicas

- Las técnicas de selección de atributos como filtro, envoltorio y ranking reducen la dimensionalidad de los factores asociados a la deserción estudiantil en el Instituto Superior Tecnológico Privado ISTEPSA significativamente.
- Las técnicas de Clustering EM. SOM y K medias determinan mejor los grupos de alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA.

— Las técnicas de A priori permiten encontrar las asociaciones interesantes en forma de relaciones de implicación entre los factores asociados a la deserción estudiantil en el Instituto Superior Tecnológico Privado ISTEPSA.
— La técnica K medias de la minería de datos determina los parámetros de los segmentos con riesgo deserción del Instituto Superior Tecnológico Privado ISTEPSA.

2.2. Objetivos de la Investigación

Objetivo general

Determinar los factores asociados para segmentar los alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA de Andahuaylas, con las técnicas de Machine Learning y Minería de Datos.

Objetivos específicos

— Reducir los factores que influyen significativamente en la deserción estudiantil en el Instituto Superior Tecnológico Privado ISTEPSA, mediante las técnicas de selección de atributos de Weka
— Determinar las técnicas no supervisadas de aprendizaje automático que segmentan mejor los alumnos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA
— Establecer los patrones de deserción en el Instituto Superior Tecnológico Privado ISTEPSA, mediante la técnica A priori de la asociación de la minería de datos.
— Determinar los parámetros de segmento de alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.

2.3 Método, diseño y tipo de investigación

El diseño o enfoque de investigación es cuantitativo, diseño no experimental por no manipularse las variables independientes intencionalmente, es de nivel o tipo descriptivo y correlacional, es un estudio de corte transversal por aplicarse una encuesta en una sola oportunidad, es observacional de acontecimientos y

fenómenos en su estado natural sin ninguna alteración (Hernández y Mendoza, 2018).

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.

2.4 Técnicas e instrumentos de investigación

La técnica de recolección de datos: Encuesta

El instrumento de recolección de datos: Cuestionario

Sobre el total de estudiantes matriculado al segundo semestre del año lectivo 2019, con el fin de obtener valores de las variables numéricas y varibles ordinales de una variedad de características objetivas y subjetivas de la población (García, 1993, como se citó en Chiner, 2005, p. 2).

El instrumento que se empleó fue el cuestionario organizado en la ficha de la Figura 14, el cual contiene los factores asociados a la deserción estudiantil.

Normas de aplicación: la persona encuestada responde las preguntas de cada ítem, de acuerdo con su percepción de la realidad, con distintos índices de valoración, en cuanto a la evaluación o calificación se utilizó las escalas cualitativas oficiales.

- Deficiente (1)
- En proceso (2)
- Logrado (3)
- Logro destacado (4)

2.5 Procedimientos de investigación

Para la realizar la investigación se desarrollaron las siguientes actividades:

1. Sistematización de información secundaria para conocer el estado del arte de la temática a partir de la lectura y evaluación de bibliografía de consulta y de referencia.
2. Conocimiento y caracterización de experiencias relevantes sobre deserción estudiantil y el uso de la técnicas y algoritmos de Machine Learning y Minería de Datos.

3. Definición y estructura de una base de datos integrando las dimensiones, componentes, fenómenos e indicadores requerida para el análisis y evaluación de las diferentes técnicas de *Machine Learning* y Minería de Datos
4. Evaluación, selección y consenso sobre los indicadores de la base de datos del caso propuesto.
5. recolección y pre procesamiento de datos sobre los indicadores propuestos para el análisis para a investigación.
6. Uso estricto del proceso de *Knowledge Dicover Database* KDD para la mineración de Datos.
7. Análisis de resultados
8. Discusión de los resultados con antecedentes
9. Redacción de conclusiones y recomendaciones.

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

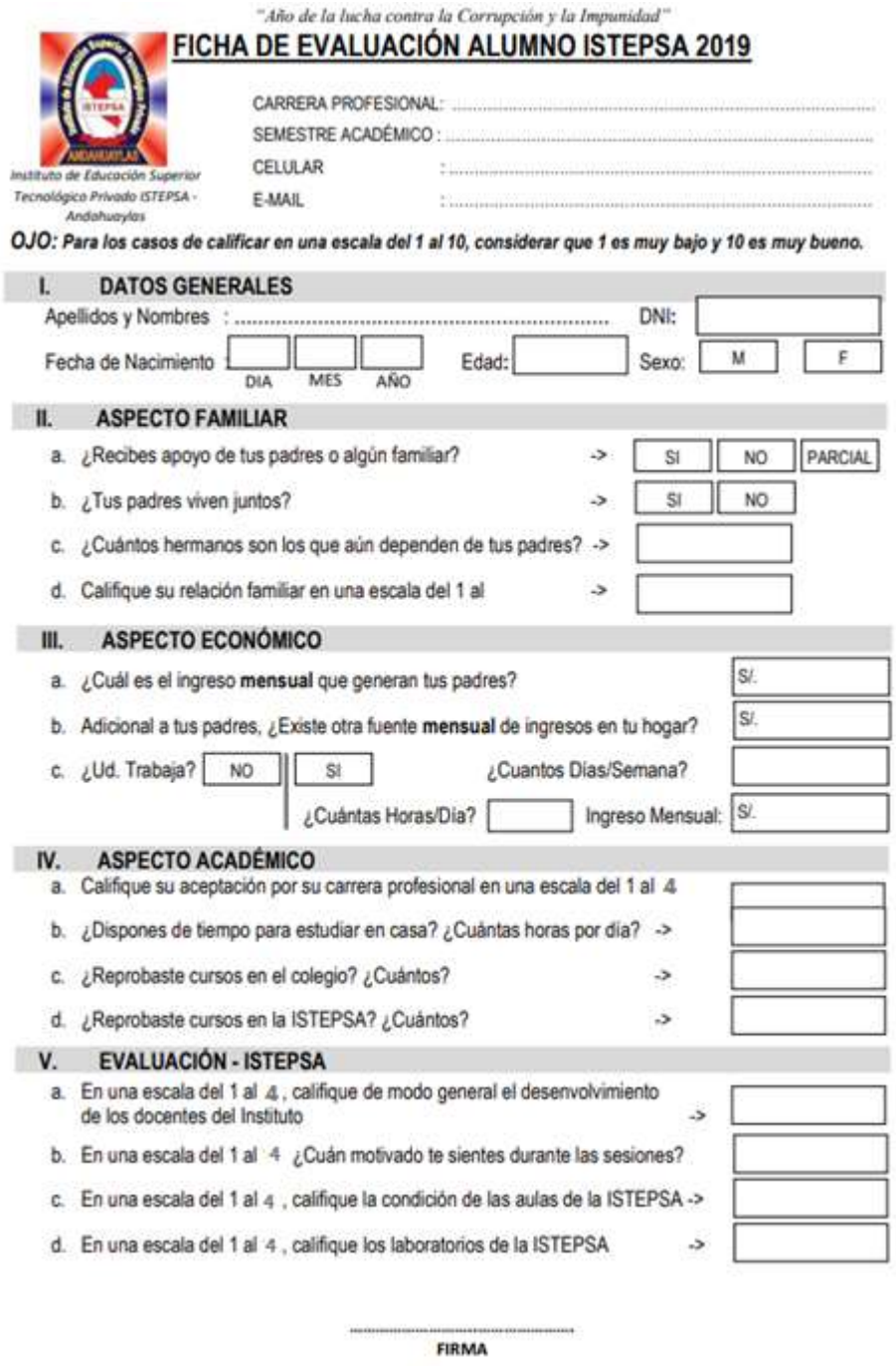
Dominio de aplicación: Incluye el conocimiento relevante previo y las metas de la aplicación. Se identifican en el Instituto Tecnológico Privado ISTEPSA la cantidad variable de estudiantes ingresantes en cada proceso de admisión, donde en los últimos semestres se observa que durante el proceso de formación 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.

Extracción de 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:

Figura 14

Ficha de cuestionario de aplicación al alumno – ISTEPSA



"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 ->

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 4

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 4, califique de modo general el desenvolvimiento de los docentes del Instituto ->

b. En una escala del 1 al 4 ¿Cuán motivado te sientes durante las sesiones?

c. En una escala del 1 al 4, califique la condición de las aulas de la ISTEPSA ->

d. En una escala del 1 al 4, califique los laboratorios de la ISTEPSA ->

.....

FIRMA

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*.

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:

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.

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.

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).

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 *CfsSubsetEva*l 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.

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.

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.

2.6 Consideraciones éticas

En cada una de las etapas de la investigación existen cuestiones éticas correspondientes a los participantes. Con respecto a los autores, estos cumplen con los requisitos de haber contribuido en la concepción, el diseño, análisis e interpretación de los datos, así como haber participado en la redacción del libro con una revisión detallada, y haber aprobado la versión final del manuscrito.

Los que ayudaron únicamente en la captura de los datos o en la supervisión general del grupo de investigación, no son considerados como autores del libro.

Los profesionales que participan de la elaboración del texto, pero no aparecen sus nombres por ser responsables de la editorial.

El presente texto también ha sido administrado el nivel de similitud aceptable por el software turnitin.

2.7 Operacionalización de variables

Tabla 1

Operacionalización de variables

PROBLEMA	OBJETIVOS	HIPOTESIS	VARIABLES	INDICADORES	INDICES	MÉTODO
PROBLEMA PRINCIPAL ¿Cuáles son las técnicas de Machine Learning y Minería de Datos que determinan los factores asociados para segmentar los alumnos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019? **PROBLEMAS ESPECÍFICOS** a. ¿Cuáles son las técnicas de la minería de datos que reducen la dimensionalidad de los factores asociados a la deserción estudiantil en el Instituto Superior	**OBJETIVO GENERAL** Determinar los factores asociados para segmentar los alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA de Andahuaylas, con las técnicas de Machine Learning y Minería de Datos. **OBJETIVOS ESPECIFICOS** a. Reducir los factores que influyen significativamente en la deserción estudiantil en el Instituto Superior Tecnológico Privado ISTEPSA, mediante las técnicas de	**HIPOTESIS GENERAL** Las técnicas de selección de factores, asociación y clustering de Machine Learning y Minería de Datos determinan los factores académicos e institucionales como más significativos para segmentar los alumnos con riesgo de deserción del Instituto Superior Tecnológico Privado ISTEPSA. **HIPOTESIS ESPECIFICAS** a. Las técnicas de selección de atributos como filtro, envoltorio y ranking reducen la dimensionalidad de los factores asociados a la deserción estudiantil en el Instituto Superior Tecnológico Privado	**VARIABLE 1** Factores y asociados a la 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** No experimental Transversal. **METODOLOGÍA PARA LA**
				Sexo	Género M/F	
				Edad	Número (1-100)	
				Estado Civil	Casado/Soltero	
				Número de Hijos	Número (0-5)	
				Vive con sus padres	SI/NO	
				Percepción relación Familiar	Número (1-4)	
				Apoyo Económico Familiar mensual	Moneda (0-3000)	
				Días de trabajo/semanal	Número (0-7)	
				Horas trabajo/día	Número (0-15)	
				Ingreso mensual propio	Moneda (0-3000)	
				Aceptación por la carrera escogida	Número (1-4)	
				Cursos reprobados en el colegio	Número (1-4)	
				Motivación de sesiones	Número (1-4)	
				Calificación de aulas ISTEPSA	Número (1-4)	

Tecnológico Privado ISTEPSA? b. ¿Cuáles son las técnicas no supervisadas de aprendizaje automático segmentan mejor los alumnos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA? c. ¿Cuál es la técnica no supervisada de Asociación determina la forma de relaciones de implicación asociados a la deserción estudiantil en el Instituto Superior Tecnológico Privado ISTEPSA? d. ¿Cuáles son los parametros de los factores asociados en los segmentos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA?	selección de atributos de Weka b. Determinar las técnicas no supervisadas de aprendizaje automático que segmentan mejor los alumnos con riesgo de deserción en el Instituto Superior Tecnológico Privado ISTEPSA c. Establecer los patrones de deserción en el Instituto Superior Tecnológico Privado ISTEPSA, mediante la técnica A priori de la asociación de la minería de datos. d. Determinar los parámetros de segmento de alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA, durante el periodo 2019.	ISTEPSA significativamente. b. Las técnicas de Clustering EM. SOM y K medias determinan mejor los grupos de alumnos con riesgo de abandono de estudios en el Instituto Superior Tecnológico Privado ISTEPSA. c. Las técnicas de A priori permiten encontrar las asociaciones interesantes en forma de relaciones de implicación entre los factores asociados a la deserción estudiantil en el Instituto Superior Tecnológico Privado ISTEPSA. d. La técnica K medias de la minería de datos determina los parámetros de los segmentos con riesgo deserción del Instituto Superior Tecnológico Privado ISTEPSA.		Calificación de laboratorios ISTEPSA	Número (1-4)	**SEGMENTACIÓN DE ALUMNOS CON RIESGO DE DESERCIÓN** Mediante la aplicación técnicas de minería de datos.
			VARIABLE 2 Técnicas de Machine Learning y Minería de Datos.	Técnicas de Minería de Datos: Método evaluador del subconjunto de atributos por filtro: CfsSubsetEval Best first EvolutionarySearch GreedyStepwise LinearForwardSelection SubsetSizeForwardSelection Método evaluador del subconjunto de atributos por envoltorio: WrapperSubsetEval Best first EvolutionarySearch GreedyStepwise LinearForwardSelection Método evaluador individual de atributos:	Validación interna y externa de las técnicas de Minería de Datos y Machine Learning		

				ChiSquaredAttributeEval Ranker Técnicas de No Supervisadas de Machine Learning: EM (Expectation Maximization) SOM (Mapas auot organizados) K medias Técnica A Priori de Asociación		

CAPÍTULO III

RESULTADOS, DISCUSIONES Y CONCLUSIONES

3.1 Exposición de los resultados

3.1.1. Resultados de los factores asociados a la deserción estudiantil de ISTEPSA

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 2

Indicadores para el análisis de deserción estudiantil

N°	Indicadores	Descripción	Valores
1	C_Pro	Carrera Profesional	1-4
2	S_acad	Semestre Académico	1-6
3	Edad	Edad	Numérico
4	Sexo	Genero 0: mujer; 1: varón	0-1
5	A_familia	Apoyo familiar	Si, No
6	Padres_junt	Padres Viven Juntos	Si, No
7	Her_dep_pad	Hermanos dependen de padres	Numérico
8	Re_familia	Relación familiar	1-4
9	Ingreso_m_p	Ingreso mensual padres	Numérico
10	Ingreso_a	Ingreso adicional	Numérico
11	Trabaja	Trabaja el estudiante	Si, No
12	Tra_dia_sem	¿Cuántos días a la semana?	Numérico
13	Horas_dia	¿Cuántas horas/día?	Numérico
14	Ingreso_mes	Ingreso mensual	Numérico
15	Acepta_c_p	Acepta a su carrera profesional	1-4
16	Horas_est	Cuántas horas dispones para estudio	Numérico
17	Curso_rep_co	Cursos reprobados en colegio	Numérico
18	Curso_rep_inst	Cursos reprobados ISTEPSA	Numérico

19	Cal_docen_inst	Calidad de docentes de ISTEPSA	1-4
20	Motiv_sesiones	Motivación de sesiones	1-4
21	Cal_aulas_inst	Calificación de aulas de ISTEPSA	1-4
22	Cal_lab_inst	Calificación de laboratorios de ISTEPSA	1-4
23	Retirado	Se retira de la institución	0-1

De los 23 indicadores codificados para el análisis en la Tabla 2, no todos tienen la misma importancia en la explicación de la clase deserción estudiantil (valor de Se retira = 1). En este estudio, el uso de las estrategias de selección automática para determinar los factores con mayor poder discriminante, se utilizó software Weka que tiene una variedad de técnicas de selección de atributos que reduce el subconjunto de indicadores que explican la clase retirado. 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 3 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 3

La valuación del subconjunto de factores seleccionados con el método evaluador de atributos por Filtro CfsSubsetEval

Método de Búsqueda	Nº 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
SubsetSizeForwardSelection	6	S_acad, Acepta_c_p, Curso_rep_co, Motiv_sesiones, Cal_aulas_inst, Cal_lab_inst

Los métodos evaluadores del subconjunto de atributos por Filtro son implementados con técnicas estadísticas mientras que los métodos evaluadores del subconjunto de atributos por Envoltorio son implementados con técnicas avanzadas de *Machine Learning* que demanda bastante recurso computacional, por lo que en este trabajo no se pudo abordar.

Los factores asociados a la deserción estudiantil son: Semestre académico, Acepta a la carrera profesional que cursa, cursos reprobados en secundaria, motivación de sesiones de aprendizaje, calificación de las aulas y la calificación de laboratorio.

Tabla 4

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 métodos evaluadores individuales conocidos como prorrateadores de atributos (*AtributeEval*) que no seleccionan indicadores, sino que estable un orden por relevancia descendente. El caso utilizó el prorrateador *ChiSquaredAttributeEval* que evalúa el valor de un atributo mediante el cálculo del estadístico chi-cuadrado con respecto a la clase, de los 22 factores independientes son ordenados por método de búsqueda Ranker, se ordenó tal como se muestra en la Tabla 4, los 6 atributos que son seleccionados por todos los métodos de selección de atributos se encuentran entre los 7 primeros del ranking.

3.1.2. Resultados de los patrones asociados a la deserción estudiantil en ISTEPSA

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 5

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 Cal_lab_inst=deficiente 55 ==> Cal_aulas_inst=deficiente 52	0.95	1.89	0.06	6.86
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 a las reglas producidas mediante el algoritmo A priori se identifican un papel relevante del factor Institucional y Académico. Calificación de laboratorios del instituto, es considerado 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,; además 96% de los estudiantes que se retiran, consideran deficiente a la carrera profesional que estudian a diferencia de la primera regla, es más 92% de estudiantes que se retiran consideran que las aulas de la institución son deficientes; Por otro lado tres reglas señalan que el 92% de los estudiantes que se retiran, ratifican que la motivación de sesiones de aprendizaje 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.

3.1.3. Resultado del segmento de los alumnos con riesgo de abandono de estudios de ISTEPSA.

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, la tabla de relación: tesis_deserción implementada con el software Weka se obtiene resultados de frecuencias probabilísticas esperadas con el algorimo EM para los parámetros 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

Del total 427 instancias: 427, los atributos seleccionados, considerando 22 Atributos: 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.

Modo de prueba: evaluación de clases a grupos en datos de entrenamiento (conjunto de entrenamiento completo), número de conglomerados seleccionados mediante validación cruzada: 5, número de umbral iteraciones realizadas: 26 se obtiene a continuación:

Tabla 6

Frecuencias esperadas de los segmentos de instancias por Clúster a partir de algoritmo EM

Clúster	0 (0.09)	1 (0.2)	2 (0.29)	3 (0.42)
		S_acad		
1	12.1642	7.12	9.6485	14.0672
2	9.0017	29.1877	21.3377	47.473
3	3.0517	3.0644	11.2246	15.6594
4	9.1569	36.8402	51.5463	77.4566
5	7.5706	11.53	14.7332	26.1662
6	4.9406	2.652	22.9986	2.4089
[total]	45.8856	90.3943	131.4888	183.2313
		Acepta_c_p		
1	7.7595	3.4933	4.1203	75.6269
2	3.4748	2.8473	8.9067	17.7713
3	8.8356	17.5175	90.5636	61.0833
4	23.8157	64.5362	25.8983	26.7498
[Total]	43.8856	88.3943	129.4888	181.2313
		Motiv_sesiones		
1	5.7605	2.7221	6.6676	120.8498
2	5.0858	4.2862	7.221	27.407
3	13.0484	54.0485	100.3715	29.5316
4	19.9908	27.3376	15.2286	3.443
[total]	43.8856	88.3943	129.4888	181.2313
		Eval_aulas_inst		
1	2.863	50.6502	9.5156	154.9713
2	1.7919	8.689	20.6575	13.8616
3	12.7925	25.2276	93.188	10.7919
4	26.4382	3.8275	6.1278	1.6065
[total]	43.8856	88.3943	129.4888	181.2313
		Eval_lab_inst		
1	2.9032	57.2307	4.7551	157.1109
2	2.2775	9.8512	18.5513	8.32
3	5.4441	15.3291	101.1245	14.1023
4	33.2607	5.9833	5.0579	1.6981
[total]	43.8856	88.3943	129.4888	181.2313

Tiempo necesario para crear el modelo (datos de entrenamiento completos): 3.17 segundos, de acuerdo al modelo y evaluación se obtiene 71 estudiantes que se retiran agrupados en el segmento 3, tabla 6.

Tabla 7

Instancias agrupadas con clase se_retira

Clúster	Instancias	No se retiran	Se retiran	Asignación
0	37 (9%)	31	6	Sin Clase
1	76 (18%)	62	14	Sin Clase
2	133 (31%)	113	20	No se retiran
3	181 (42%)	110	71	Se retiran

Probabilidad de registro: -5.6649, considerando el atributo de clase: segmento de estudiantes que se retira representado por el clúster 3, segmento de los estudiantes que no se retiran están en el clúster 2, mientras los clústeres 0 y 1 son segmentos de los estudiantes que no se sabe. Sin embargo, las Instancias agrupadas incorrectamente son: 243 (56.9087 %)

Segmentación con el algoritmo de 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 completa de los 23 factores, sin considerar la reducción de la dimensionalidad de factores, es decir con las técnicas de selección de atributos de minería de datos para la clase "se retira". Los resultados obtenidos en Weka se ilustran a continuación:

Esquema: weka. clusterers. SelfOrganizingMap, para los parámetros -L 1.0 -O 2000 -C 1000 -H 2 -W 2, sobre la tabla de la relación: tesis_deserción de 427 instancias, 22 indicadores, modo de prueba: evaluación de clase se retira, los grupos para datos de entrenamiento, para los atributos seleccionado resulta no significativo, sin embargo, se aplica para la totalidad de los indicadores con una validación cruzada: 4; con número de iteraciones realizadas: 1; los parámetros de los segmentos de instancias por Clúster se presenta a continuación:

Tabla 8

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

Clúster/Atributo	0 (110)	1 (48)	2 (186)	3 (83)
		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
		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
		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
		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 respectivo.

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 %).

3.1.4. Resultados de parámetros de los segmentos con riesgo de deserción mediante el algoritmo K medias

En la Tabla de datos: tesis_deser1, para 427 instancias: 427, 6 atributos como son: S_acad, Acepta_c_p, Motiv_sesiones, Eval_aulas_inst, Eval_lab_inst, para el atributo clase se retira, con evaluación de los datos de entrenamiento del algoritmo k Media, número de iteraciones: 4, dentro de la suma del conglomerado de errores al cuadrado: 973.0 y puntos de partida iniciales (aleatorios), se obtienen dos segmentos cuyos parámetros son:

Tabla 11

Parámetros de los segmentos de estudiantes con riesgo de deserción mediante algoritmo K medias

Clústeres	0	1	2	3	4	5
Segmento1	2	1	2	1	3	0
Segmento 2	4	3	3	4	4	0

Valores faltantes reemplazados globalmente por media/moda:

Tabla 12

Centroides de los segmentos finales

Factores	Completo	Clúster 1	Clúster 2
S_acad	4	4	4
Acepta_c_p	3	4	3
Motiv_sesiones	3	1	3
Cal_aulas_inst	1	1	3
Cal_lab_inst	1	1	3
Se retira	0	0	0

Tiempo necesario para construir el modelo (datos de entrenamiento completos): 0 segundos, modelo y evaluación en conjunto de entrenamiento de Clúster 0 compuesto de 270 (63%) instancias, mientras que el Clúster 1 está compuesto de 157 (37%) instancias.

Clúster 1 segmento de los estudiantes con riesgo de deserción resultan de 4to semestre, acepta su carrera profesional como logro destacado, motivación de sesiones de aprendizaje es deficiente, deficiente la calificación de aulas y laboratorios de la Institución.

3.2 Discusión

Mollo (2018) desarrolló análisis predictivo de la deserción estudiantil utilizando data warehouse con la metodología Ralph Kimball y minería de datos con la metodología CRISP-DM concluye que los factores económicos destacan como causantes de deserción, al respecto, la investigación de manera similar utiliza la metodología Ralph Kimball para construir data warehouse, en este caso los que destacan son factores institucionales y factores académicos.

Las similitudes alcanzadas concordante con Holgado (2018), determinó que la elección de la carrera profesional fue influyente en el rendimiento académico; mientras que, en esta investigación resulta que la calificación de la preferencia de la carrera profesional se relaciona inversamente a mayor preferencia menos es el riesgo de la deserción.

Como destaca Rivera (2016), los factores determinantes en la deserción escolar de nivel primario son ligeramente diferentes a nivel de educación superior tecnológica, por lo que, se debe asumir con bastante prudencia en estos estudios.

Yamao (2018) predijo el rendimiento académico mediante de minería de datos, logrando identificar tempranamente a los alumnos que tuvieron dificultades académicas, tomando acciones para mitigar el riesgo de deserción, en este caso, los resultados son parecidos son académicos e institucionales, una vez identificadas la institución tiene responsabilidad de implementar acciones correctivas.

Torres (2018), segmentó las relaciones con los clientes en la empresa optimizando los grupos homogéneos y así enfocar apropiadamente las estrategias comerciales, en el caso de esta investigación aplicado a las instituciones del sector privado debe buscar la fidelización apropiada de los estudiantes.

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 (De la Cruz, 2017), para tales fines el presente trabajo también utilizó la técnica K-medias para conocer los perfiles de estudiantes.

La identificación de los factores o atributos que explican significativamente la deserción estudiantil de ISTEPSA, durante el periodo 2019 resultan seis obtenida mediante las técnicas de filtro, envoltorio y ranking de acuerdo al orden son: motiva sesiones de aprendizaje (Motiv_sesiones), califica las aulas de la Institución (Cal_aulas_inst), califica los laboratorios de la Institución (Cal_lab_inst), acepta a la carrera profesional (Acepta_c_p), cursos reprobado en colegio (Curso_rep_co) y semestre académico (S_acad), son factores asociados a la dimensión académica.

En el contexto académico, resulta que “motiva sesiones de aprendizaje” como un factor de mayor importancia para el desarrollo de las clases, es decir realizar una alta motivación al inicio de las sesiones de aprendizaje, implica generar en los estudiantes disposición psicológica para concentrarse y atender las clases desarrolladas por el docente, para ello, debe complementar condiciones pedagógicas de infraestructura adecuada en aulas y laboratorios para generar el aprendizaje significativo teórica y práctico, sin perder de vista que la vocación de la formación profesional es fundamental, algunos estudiantes están por la

obligación de padres o necesidad de continuar sus estudios superiores en carreras profesionales que no responden a los perfiles profesionales ideales del estudiante, lo cual se manifiesta, que en el segundo semestre del año lectivo las matrículas son reducidos.

Ésta realidad comparada frente a otros trabajos como sostienen Pérez et al. (2018) determinaron el subconjunto de atributos mediante 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 y Suénaga (2015) en el estudio de "Análisis de Deserción-Permanencia de Estudiantes Universitarios utilizó la técnica de Clasificación de Minería de Datos", también redujo los factores aplicando el método de evaluación CfsSubsetEval y el de búsqueda BestFirst, lo que resultó 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 los resultados de este método 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 sugieren realizar reuniones (al menos una por semestre) con los padres de los estudiantes para que ellos entiendan y conozcan las actividades educativas y presten 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 dicha institución.

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:

- El abandono voluntario ocurre durante los primeros meses posteriores al ingreso a la institución;
- Cinco de cada diez estudiantes desertan al inicio del segundo año;
- Cuatro de cada diez estudiantes que comienzan el cuarto año, no obtienen el título profesional correspondiente; y
- 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.

En este trabajo el propósito es detectar rápidamente el fenómeno de deserción y conociendo los factores académicos y otras que corresponden a la situación socioeconómica se reporta las tasas de abandono que oscilan entre el 15% y el 40%.

Los resultados son oportunos para que las autoridades académicas implementen políticas de retención de los estudiantes que han sido detectados con riesgo de deserción, lo que redundaría en una mejora del sistema académico institucional.

En cuanto al uso de los métodos no supervisadas de *Machine Learning* como son las técnicas de clustering y asociación, son muchas las técnicas, pero algunos están orientados para análisis de variables numéricas mientras que el presente trabajo utiliza variables categóricas, para los cuales se adapta mejor los algoritmos de clustering EM y SOM.

En primer lugar, se utilizó los métodos de selección de características basados en correlación para la extracción de variables. Como era de esperar, la variable

Por lo tanto, sería interesante para futuras investigaciones estudiar la deserción universitaria después del primer año, evaluar más grados en las diferentes áreas del programa y comparar los datos obtenidos de otras técnicas de aprendizaje automático, como *Random Forest* o *Gradient Boosting*. Además, aunque los

resultados no encontraron que el género fuera un predictor significativo de la deserción universitaria en el contexto específico de este estudio, sí lo es

De los 40 estudios estudiados sobre la predicción del rendimiento académico de estudiantes universitarios aplicando técnicas de minería de datos, 26 autores utilizan la metodología KDD, esto resalta la importancia de esta metodología, siendo la más utilizada a la hora de predecir el rendimiento académico. Así mismo, con respecto a los atributos se consideraron para la predicción del rendimiento académico: Género, CGPA y trabajo del padre/madre son los factores más utilizados y corresponden 20 al factor personal, 21 al factor académico y 5 al socioeconómico. factor respectivamente. Aunque el total de los factores es múltiple, sus características cambian de uno a otro. Así, en relación a los algoritmos de selección de variables, 3 autores utilizan tanto correlación *AttributeEval* como *InfoGain-AttributeEval*, ya que estos modelos pueden ofrecer mejores resultados que se acerquen a la realidad. Por otro lado, con respecto a las técnicas utilizadas en la predicción y su precisión, existen muchos algoritmos para la realización de cada tarea, siendo que 17 utilizan Naives Bayes y 14 utilizan J48, demostrando que estos son los más eficientes a la hora de predecir resultados académicos. actuación. En cuanto a la precisión, 16 artículos hablan de Naives bayes, siendo este el más utilizado por la capacidad de registrar grandes grupos de datos, con un 94% de precisión. Asimismo, la herramienta más utilizada para el desarrollo y prueba del modelo es Weka, la cual fue utilizada por 22 de los investigadores considerados para el estudio, esto se puede justificar por la frecuencia y la mayor recepción que tiene a la hora de desarrollar minería de datos. Finalmente, 27 investigadores utilizan la métrica de precisión y 13 de Recall, ya que arrojaron mejores resultados a la hora de predecir el rendimiento y son los más efectivos a la hora de predecir.

El algoritmo k-means no produce un agrupamiento adecuado debido al uso de datos categóricos.

El algoritmo EM, ha producido una clasificación de baja calidad en comparación SOM, se encontró que las características de los estudiantes no estaban bien separadas.

El algoritmo DBSCAN es una técnica de agrupamiento basada en la densidad que asume que los agrupamientos son áreas de alta densidad en un solo espacio y están separados por áreas de baja densidad. Entonces, este algoritmo identifica grupos en un conjunto de datos simplemente observando la densidad local de los puntos de datos. En este caso, el algoritmo DBSCAN no agrupa bien porque el conjunto de datos de los estudiantes utilizado es de gran dimensión y no de naturaleza espacial, y el ruido hace que la complejidad sea alta un análisis no supervisado para recopilar información sobre el comportamiento de los estudiantes para un Sistema de Gestión del Aprendizaje.

Al analizar la correlación de los atributos con la clase, verificamos que los datos socioeconómicos hacen que los atributos sean pobres para identificar el desempeño de los estudiantes. Esto puede explicarse en parte por el hecho de que falta una gran parte de los datos, como los ingresos familiares, que los estudiantes a menudo se niegan a proporcionar, y la situación laboral. Por otro lado, confirmamos que algunos datos de Moodle son atributos muy útiles. Un ejemplo es el puntaje general bajo en las actividades de Moodle, lo que indica que el estudiante es propenso a fallar o abandonar. El poco compromiso con las actividades de Moodle es una señal de advertencia temprana y se convierte en un indicador importante de fallas en M2 y M3.

También vale la pena señalar que nuestros estudios se realizaron con datos de clases completadas a fines de 2019, por lo tanto, antes del inicio de la pandemia por SARS-CoV-2. Después de eso, sin embargo, se podrán realizar nuevos estudios y se podría realizar un análisis similar para considerar el impacto del aislamiento social a lo largo del curso y cómo la adopción de aulas remotas ha afectado la presencial o semipresencial.

3.3 Conclusiones

Mediante el método de evaluación *CfsSubsetEval* y el método de búsqueda *Best first* de minería de datos se determinó los factores académicos e institucionales que influyen a la deserción de los estudiantes de ISTEPSA como son 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 determinar las técnicas no supervisadas de aprendizaje automático EM para encontrar las frecuencias esperadas de las probabilidades de ocurrencia de segmentos de alumnos con riesgo de deserción, mientras que, para representar los parámetros las variables numéricas se usaron SOM y para las características de las variables categóricas se usó K medias.

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, donde se observa que los factores académicos e institucionales son determinantes para la Institución Educativa Privado.

REFERENCIAS

Aranciaga, J. & Ccanto, E. (2021). *Factores asociados a la deserción de estudiantes en un instituto de educación superior privado de Lima*. https://cutt.ly/zMq4ybq

Arcila Calderón, C., Barbosa Caro, E., & Cabezuelo Lorenzo, F. (2016). *Técnicas de Big Data: Análisis de textos a gran escala para la investigación científica y periodística*. https://cutt.ly/WMq4pcF

Baviera, T. (2016), Técnicas para el análisis del sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength. *Revista Dígitos*. 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. 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*. 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*. 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.

Dash, M., Liu, H., & Motoda, H. (2000, April). *Consistency based feature selection. In Pacific-Asia conference on knowledge discovery and data mining* (pp. 98-109). Springer, Berlin, Heidelberg.

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

Echevarría, R. (2003). *El coaching ontológico posee tres premisas básicas: Los seres humanos seres lingüísticos, el lenguaje es generativo de realidades en nuestro entorno y los seres humanos se crean así mismo en el lenguaje*. Chile: Lom Ediciones S.A.

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)*. https://rpubs.com/Cristina_Gil/PCA.

Ginzberg, Axelrad y Hermán. (1951). *Teoría sobre la elección de carrera y su relación con la deserción*. https://cutt.ly/GMq4ZkL

Grinder, R. (2001). Adolescencia. Editorial Limusa. México D.F. https://cutt.ly/YMq4N8f

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*. 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*. 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*. https://www.redalyc.org/journal/852/85263724001/85263724001.pdf.

Jara Tuesta, B. A. (2017). *Factores que conducen a la deserción en estudiantes de una universidad privada de Lima Norte* (Tesis de maestría). Universidad Cesar Vallejo. https://cutt.ly/2Mq7iXs

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*. 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. 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*. https://cutt.ly/bMq7mwG

Mandrekar, J. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, 5(9), 1315-1316. https://doi.org/10.1097/JTO.0b013e3181ec173d.

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

McCarthy, J. (2007). *What is artificial intelligence?* Standford University, California, EE.UU.

Miranda, M. & Guzmán, J. (2017). *Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos*. 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.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. Cambridge, MA : The MIT Press.

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*. 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*. 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*. https://cutt.ly/2Mq5a0S

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*. https://renati.sunedu.gob.pe/handle/sunedu/1799018.

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

Russell, S., y Norvig, P. (2010). *Artificial Intelligence a Modern Approach*. New Jersey: Pearson Education.

Saito, T. and Rehmsmeier, M. (2015). The Precision-Recall Plot is More Informative than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, 10(3). 1-21. https://doi.org/10.1371/journal.pone.0118432

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

Super, D. (1953). *Teoría de Super. Blog Orientación vocacional y educativa*. http://orientacion-morelos.blogspot.pe/2010/02/teoria-desuper.html

Tan, Steinbach & Kumar (2006). *Introducción a la minería de datos*. https://cutt.ly/bMq5TUy

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*. 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*. 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. 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. 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. Data en formato Weka

```
@relation tesis_deser
@attribute C_Pro {1,2,3,4}
@attribute S_acad {1,2,3,4,5,6}
@attribute Edad numeric
@attribute Sexo {0,1}
@attribute A_familia {1,2,3}
@attribute Padres_junt {1,2}
@attribute Her_dep_pad numeric
@attribute Re_familia numeric
@attribute Ingreso_m_p numeric
@attribute Ingreso_a numeric
@attribute Trabaja numeric
@attribute Tra_dia_sem numeric
@attribute Horas_dia numeric
@attribute Ingreso_mes numeric
@attribute Acepta_c_p {1,2,3,4}
@attribute Horas_est numeric
@attribute Curso_rep_co numeric
@attribute Curso_rep_inst {1,2,3,4}
@attribute Eval_docen_inst {1,2,3,4}
@attribute Motiv_sesiones {1,2,3,4}
@attribute Eval_aulas_inst {1,2,3,4}
@attribute Eval_lab_inst {1,2,3,4}
@attribute retira {0,1}
@data
1,4,20,0,2,1,2,1,500,0,1,2,5,300,4,2,0,0,4,3,3,1,1
1,4,21,0,3,1,5,1,300,0,1,4,3,200,1,1,0,0,1,3,1,1,1
1,4,22,0,2,2,5,1,200,0,1,7,4,600,1,1,0,0,1,1,1,1,0
1,4,19,1,2,1,0,1,1000,0,0,0,0,0,4,1,0,0,3,3,1,1,1
1,4,20,0,2,2,3,4,200,0,1,3,6,300,4,1,0,0,4,4,1,1,1
1,4,19,1,1,1,2,3,300,0,0,0,0,0,1,2,0,2,3,2,1,2,0
1,4,21,1,1,1,2,4,400,0,0,0,0,0,4,2,0,3,4,1,3,3,0
1,4,23,0,2,1,5,4,1500,0,1,3,6,300,2,2,0,2,4,1,1,1,
0
1,4,22,0,1,1,1,4,300,0,0,0,0,0,2,1,0,0,2,2,2,3,0
1,4,27,1,2,1,1,3,200,0,0,0,0,0,4,1,0,0,4,3,3,3,0
1,4,19,0,1,1,4,4,500,0,0,0,0,0,3,1,0,0,3,1,1,1,0
1,4,24,0,1,2,2,1,300,0,0,0,0,0,3,2,0,2,4,4,1,1,1
1,4,20,0,1,1,2,3,200,0,0,0,0,0,4,2,0,0,4,4,1,2,1
1,4,21,0,1,1,3,4,400,0,0,0,0,0,4,2,0,0,3,2,3,4,1
1,4,22,0,3,1,1,3,400,0,1,5,6,400,1,1,0,0,3,1,1,1,1
1,4,21,1,2,1,3,3,600,0,0,0,0,0,4,1,1,1,3,1,1,1,0
1,4,22,1,3,1,1,3,400,0,1,5,6,900,1,1,0,0,3,1,1,1,1
3,4,20,0,1,1,3,4,150,0,0,0,0,0,4,1,0,1,1,3,1,1,1
3,4,21,0,2,2,3,1,300,0,1,7,4,250,2,1,0,3,4,3,1,1,0
3,4,20,0,2,2,0,1,200,0,0,0,0,0,1,1,0,1,4,1,1,1,0
3,4,19,0,2,2,2,1,150,0,1,3,4,250,1,1,4,1,2,1,1,1,0
3,4,21,0,1,2,2,1,900,0,0,0,0,0,1,1,0,1,1,1,1,1,0
3,4,19,1,1,2,1,1,400,0,1,5,3,250,2,1,0,0,1,1,1,1,0
2,4,21,1,2,2,2,2,200,0,1,7,8,800,4,1,0,2,2,3,1,1,0
2,4,21,1,3,2,0,3,700,0,0,0,0,0,3,1,0,1,2,1,3,3,0
2,4,23,0,2,1,2,2,300,0,1,1,8,200,4,1,0,0,1,1,1,1,0
2,4,20,1,1,1,2,4,200,0,1,5,5,150,2,1,0,1,3,1,1,3,0
2,4,21,1,1,1,4,3,250,0,0,0,0,0,3,1,4,4,3,3,3,3,1
2,4,31,0,2,2,0,3,300,0,1,5,5,300,4,1,0,0,3,3,3,3,0
2,4,24,0,1,1,4,4,300,100,1,1,5,800,4,1,0,3,4,3,4,3
,0
2,4,22,1,1,2,5,4,500,100,1,7,6,1500,4,0,0,4,3,3,1,
4,0
2,4,20,0,2,1,1,3,500,0,1,5,8,800,4,1,0,0,3,4,1,4,0
2,4,20,1,2,1,0,3,300,0,1,7,8,800,1,1,0,4,1,1,1,1,0
2,4,20,0,1,1,2,4,1000,0,1,3,4,1000,2,1,0,0,1,1,1,1
,1
2,4,21,1,1,1,3,1,200,0,1,2,6,80,3,1,0,0,3,1,3,1,0
2,4,24,1,1,1,4,4,400,600,1,3,4,400,3,2,1,1,1,1,3,1
,0
2,4,21,0,1,1,5,3,250,300,0,0,0,0,3,2,1,1,2,1,1,1,1
2,4,20,0,1,2,4,4,400,0,0,0,0,0,2,1,0,1,3,3,1,1,0
2,4,19,0,1,1,1,4,250,400,1,5,8,300,4,1,0,1,3,4,3,1
,0
2,4,23,1,1,1,2,4,1500,200,1,2,8,100,4,1,1,1,3,4,4,
4,0
2,4,30,1,3,2,5,4,100,150,0,0,0,0,4,2,0,0,1,3,3,1,0
2,2,20,0,1,1,5,2,200,100,0,0,0,0,1,2,1,2,3,1,1,1,1
2,2,20,0,1,2,4,4,200,0,0,0,0,0,4,2,0,0,1,1,1,1,0
2,2,26,1,1,1,0,3,300,0,0,0,0,0,3,1,0,2,3,1,1,1,1
2,2,24,0,1,2,0,3,300,0,1,6,8,500,4,2,1,2,3,3,2,1,1
2,2,19,0,1,2,3,4,500,10,0,0,0,0,4,2,0,0,4,4,4,3,0
2,2,22,1,1,1,4,3,300,0,1,5,6,500,2,1,2,2,3,2,1,1,1
2,2,18,1,1,2,3,3,10000,500,0,0,0,0,1,1,2,1,1,1,1,1
,1
2,2,20,1,2,1,4,3,180,20,1,5,6,300,3,2,4,2,1,3,1,3,
0
2,2,19,0,1,1,4,4,200,100,0,0,0,0,4,1,1,0,3,4,3,3,0
2,2,19,1,1,1,1,4,200,100,0,0,0,0,4,2,0,0,1,2,1,1,0
2,2,19,1,1,1,3,2,1000,0,1,5,6,300,1,1,1,3,3,2,1,3,
0
2,2,22,1,2,2,3,3,300,0,1,7,8,300,4,1,1,3,3,3,4,3,0
2,2,17,0,1,1,3,4,100,0,0,0,0,0,4,1,0,0,3,4,1,1,0
2,2,21,0,3,1,1,3,800,0,0,0,0,0,4,2,0,0,3,3,1,1,0
2,2,18,0,1,1,4,4,100,0,0,0,0,0,3,1,2,1,3,4,1,1,0
2,2,18,0,1,1,2,4,500,20,0,0,0,0,4,1,0,0,4,4,1,4,0
2,2,20,1,1,1,3,4,200,0,0,0,0,0,3,2,1,1,2,1,1,2,0
2,2,19,1,3,2,3,3,120,30,0,0,0,0,3,2,0,0,3,3,1,1,0
2,2,20,1,1,1,2,4,50,0,0,0,0,0,2,1,1,1,1,1,3,1,0
1,2,20,1,1,1,0,3,200,300,1,6,8,300,2,1,0,0,4,1,1,1
,0
1,2,23,1,1,2,2,1,1500,0,1,4,6,400,3,1,0,4,3,2,1,3,
0
1,2,22,0,2,2,2,1,2000,0,1,7,7,400,2,1,1,1,3,3,3,3,
0
1,2,20,1,3,1,2,3,50,0,1,6,8,930,1,1,0,3,3,4,3,3,0
1,2,21,0,2,1,1,1,500,0,1,2,7,200,3,1,4,1,2,1,1,1,1
1,2,20,0,1,1,2,4,100,60,0,0,0,0,4,2,3,2,3,4,3,4,0
1,2,18,0,1,1,3,4,3000,100,1,1,8,200,4,1,0,4,3,3,3,
3,0
1,2,25,0,1,1,3,3,200,100,1,6,4,200,3,1,0,0,3,4,3,1
,0
1,2,22,0,1,1,2,4,2000,300,1,6,8,300,4,1,0,0,3,3,1,
1,1
1,2,22,1,2,2,0,1,200,0,1,6,7,480,3,1,4,2,1,1,1,1,1
1,2,19,0,1,1,2,4,1000,0,0,0,0,0,4,1,0,0,3,4,4,2,0
1,2,21,1,1,2,0,4,700,200,1,6,4,250,3,2,2,1,2,3,1,1
,0
1,2,22,0,3,1,0,2,300,0,1,7,6,250,3,1,1,2,4,3,4,4,0
1,2,19,1,2,2,2,1,100,50,0,0,0,0,4,1,0,1,2,3,1,4,0
1,2,28,1,2,2,1,4,1200,200,1,6,4,930,4,2,0,0,3,4,3,
1,0
1,2,18,0,1,2,4,1,350,0,1,1,8,120,4,1,0,3,3,3,3,3,0
1,2,18,1,1,1,2,1,500,0,0,0,0,0,1,2,4,0,3,1,2,4,0
1,2,26,1,3,1,4,4,750,0,1,2,8,100,3,1,0,2,3,4,3,2,0
1,2,26,0,2,2,1,3,500,200,1,6,8,430,4,1,0,0,3,3,3,3
,0
1,2,24,0,2,2,2,1,600,0,1,2,8,400,4,1,2,2,3,3,1,1,0
4,4,20,1,1,1,1,4,500,0,0,0,0,0,4,2,3,3,3,3,1,1,0
4,4,19,0,1,1,3,4,500,0,0,0,0,0,4,2,0,0,1,3,1,1,0
4,4,19,1,1,2,1,4,600,0,0,0,0,0,4,2,1,2,1,1,1,1,0
4,4,22,0,2,2,0,2,500,0,0,0,0,0,4,1,0,3,3,3,1,1,0
4,4,23,0,1,1,3,3,200,100,0,0,0,0,3,2,1,1,4,4,1,3,0
4,4,33,1,2,2,0,3,200,0,1,6,6,1000,4,2,1,2,3,4,3,3,
0
4,4,21,0,2,1,0,4,400,0,1,5,6,250,4,1,0,3,4,3,2,3,0
4,4,27,1,1,1,3,3,500,0,1,2,8,300,3,1,1,4,4,3,3,3,0
4,4,24,0,2,1,4,4,500,50,1,3,8,300,3,1,0,2,3,3,3,3,
0
4,4,21,1,3,1,0,3,200,0,1,1,8,200,3,1,1,2,4,4,3,1,0
```

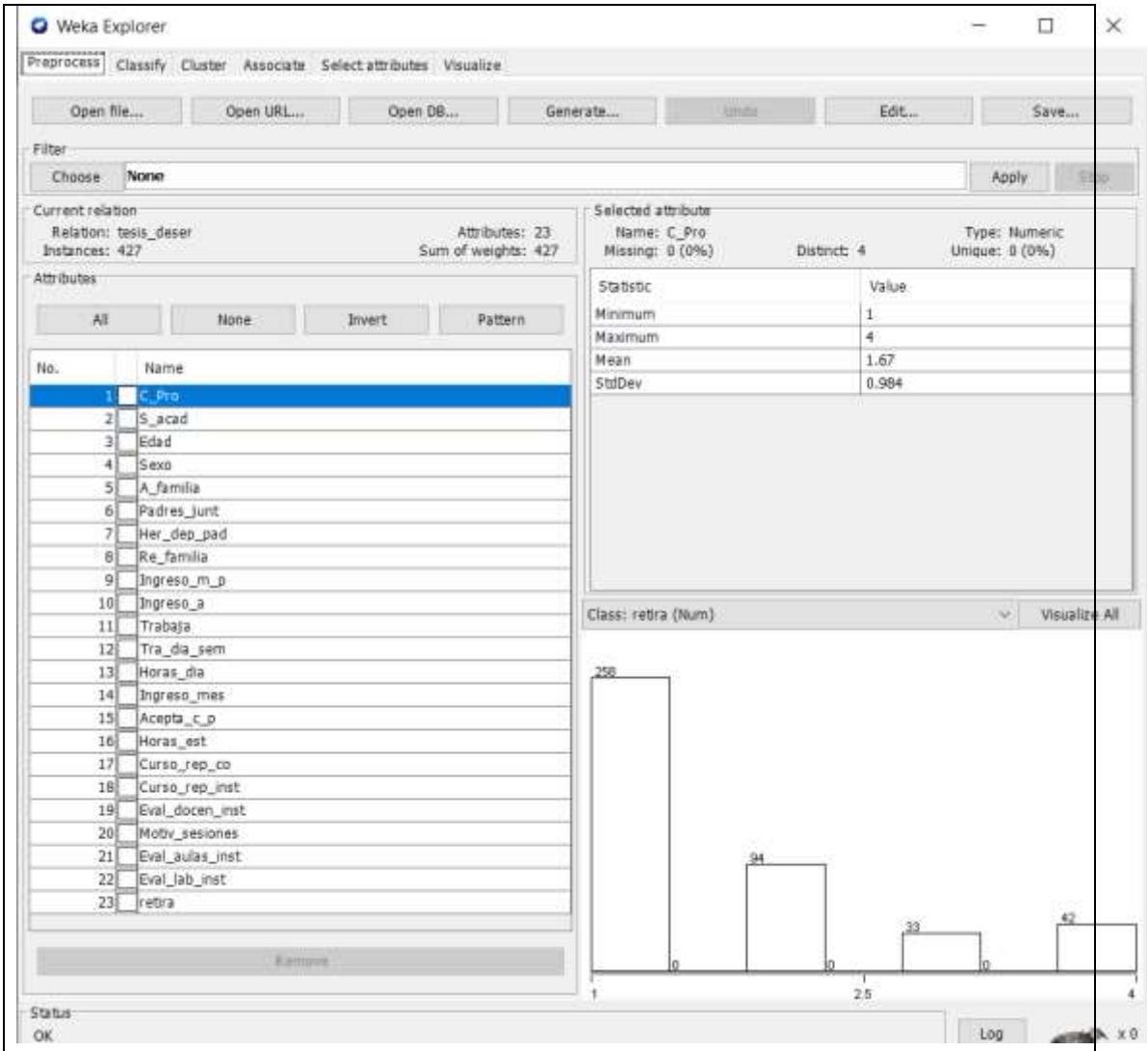
2,3,19,0,3,2,4,3,60,0,1,5,8,300,4,1,0,2,4,3,1,1,0
2,3,18,0,3,1,3,4,200,0,1,6,6,350,4,1,0,0,1,2,1,1,0
2,3,19,1,1,1,1,1,300,0,1,2,5,300,4,1,0,0,1,1,1,1,0
2,3,21,1,3,1,2,3,200,0,1,1,8,100,4,1,4,0,2,2,1,1,1
2,3,20,1,1,1,1,1,100,20,0,0,0,0,1,1,2,1,3,1,3,1,0
2,3,20,0,1,1,1,4,100,0,1,2,6,200,1,1,0,0,2,1,1,1,1
2,3,20,0,1,1,2,4,600,0,1,7,4,300,4,2,0,0,3,3,1,3,0
4,2,30,1,3,1,0,1,750,0,1,3,8,300,3,1,0,2,4,4,3,3,0
4,2,18,1,2,2,1,1,200,0,1,6,8,250,4,1,0,1,4,3,4,4,1
4,2,19,1,1,1,0,3,1000,0,1,4,4,400,3,1,0,1,3,3,1,1,
0
4,2,32,1,1,2,0,1,1500,0,0,0,0,0,0,3,1,0,1,4,3,1,2,0
4,2,24,1,3,1,1,1,500,0,0,0,0,0,3,1,2,1,3,3,1,1,1
4,2,29,0,2,2,4,4,1500,0,1,3,3,500,4,0,0,0,4,3,3,1,
0
4,2,20,0,2,2,3,1,50,0,1,7,8,100,4,1,0,0,3,3,2,1,0
4,2,20,1,1,1,3,4,1500,0,0,0,0,0,4,2,0,0,4,4,1,2,0
4,2,22,0,1,1,1,3,1000,0,0,0,0,0,4,1,1,1,4,3,3,3,0
4,2,18,0,1,1,5,3,200,0,1,2,8,50,4,1,0,2,4,3,3,1,0
4,2,23,1,2,2,1,1,200,0,1,6,8,700,1,1,0,2,1,3,1,1,0
4,2,21,1,1,2,1,1,1000,0,1,2,8,400,1,2,0,0,1,1,1,1,
0
4,2,21,1,1,1,0,4,100,200,0,0,0,0,1,1,4,0,1,1,1,1,0
4,2,19,0,1,2,4,3,1000,300,0,0,0,0,3,2,0,3,1,1,1,1,
0
4,2,22,0,1,2,4,4,300,0,0,0,0,0,3,1,1,3,3,2,1,1,1
4,2,23,0,1,2,3,1,1200,0,0,0,0,0,2,3,0,2,4,4,3,3,0
1,2,28,1,2,2,0,4,300,0,0,0,0,0,4,2,2,0,3,4,3,4,0
1,2,20,0,2,2,3,3,4500,0,1,5,8,500,3,1,0,0,1,1,1,1,
1
1,2,26,0,2,2,0,3,500,0,1,4,8,400,4,1,0,0,1,4,2,2,0
1,2,24,0,2,2,1,4,500,0,1,2,5,360,4,1,0,2,3,3,1,1,0
1,2,19,1,1,2,4,1,300,0,1,1,8,50,2,1,0,1,1,1,1,1,0
1,2,19,0,1,2,5,4,0,0,1,7,8,500,2,1,0,2,3,2,2,1,0
1,2,18,0,1,1,3,1,300,0,0,0,0,0,2,3,3,0,1,3,4,3,0
1,2,23,0,3,1,5,4,300,0,1,6,4,100,3,1,0,0,1,3,1,1,0
1,2,21,0,3,2,1,1,600,0,1,4,5,500,2,2,2,0,1,1,1,1,1
1,2,23,1,1,1,2,4,400,0,1,3,6,200,4,1,0,1,4,4,3,1,0
1,2,20,0,3,2,2,3,200,100,1,2,8,200,4,1,0,2,2,3,1,1
,1
1,2,19,0,3,2,1,3,250,200,1,2,8,200,4,1,0,2,2,3,1,1
,0
1,2,33,0,2,2,5,1,0,0,1,2,7,450,3,1,4,4,3,1,1,1,0
1,2,23,1,1,1,2,1,2,0,0,0,0,0,1,1,0,2,1,1,1,1,0
1,2,23,1,2,2,5,3,0,1000,1,2,7,400,4,1,0,1,1,3,1,2,
1
1,2,20,1,2,2,4,3,300,100,1,7,6,300,1,2,2,0,1,1,1,1
,0
1,2,21,0,1,1,0,1,150,0,0,0,0,0,4,1,0,3,2,2,1,1,0
1,2,21,0,1,2,3,2,100,0,0,0,0,0,3,1,0,1,2,2,1,1,1
1,2,20,1,2,1,0,3,400,200,1,7,6,600,3,1,0,0,3,3,2,2
,0
1,4,21,0,1,2,4,1,820,0,1,2,4,100,1,1,2,2,4,1,2,1,1
1,4,20,0,2,1,2,1,600,100,0,0,0,0,1,1,0,0,1,1,1,1,1
1,4,24,0,1,1,3,3,500,0,0,0,0,0,1,1,0,1,1,2,1,1,1
1,4,25,0,2,2,1,2,150,0,0,0,0,0,3,1,0,0,2,2,3,3,0
1,4,24,0,3,1,3,3,100,0,1,7,3,200,3,1,0,1,2,1,1,1,1
1,4,22,1,1,1,4,1,930,0,0,0,0,0,1,1,1,0,1,1,1,1,1
1,4,25,0,3,1,3,1,200,0,1,6,6,600,3,1,0,4,3,3,1,3,0
1,4,20,0,1,1,1,4,400,0,0,0,0,0,3,1,4,4,1,1,1,3,0
1,4,21,0,1,1,5,4,200,0,0,0,0,0,3,1,0,0,4,3,3,1,0
1,4,25,0,2,1,1,4,2800,0,1,6,4,500,4,1,0,0,1,1,1,1,
1
1,4,24,0,2,1,1,1,200,100,1,7,8,430,1,1,0,2,1,1,1,1
,1
1,4,24,0,1,1,2,1,200,0,0,0,0,0,3,2,2,3,1,1,1,1,1
1,4,19,0,1,1,2,3,200,0,0,0,0,0,4,1,0,1,1,1,1,1,1
1,4,19,1,1,1,2,4,1200,0,0,0,0,0,3,1,0,1,3,3,1,1,1
1,4,22,0,1,1,2,4,300,0,1,7,4,300,4,1,1,1,1,1,1,1,1
1,4,21,1,1,1,0,3,200,0,0,0,0,0,4,1,0,0,1,1,1,1,1
1,4,23,0,1,2,3,4,500,0,1,2,4,120,4,1,0,2,1,2,1,1,1
1,4,24,1,3,1,5,3,500,0,0,0,0,0,4,1,0,2,4,4,3,3,0
1,4,24,0,2,1,0,4,200,0,1,6,8,1200,4,1,0,0,4,4,1,2,
0
1,4,21,0,2,2,4,1,200,0,1,6,8,800,4,1,0,0,3,3,1,3,0
1,4,21,0,2,2,5,4,200,0,0,0,0,0,1,2,0,4,1,1,1,1,0
1,1,19,0,1,1,5,2,200,0,0,0,0,0,3,1,0,0,2,1,1,1,0
1,1,19,0,1,2,2,4,50,30,1,6,8,50,4,2,0,0,3,4,4,4,0
1,1,18,0,3,2,1,1,200,0,1,6,7,400,3,1,0,0,1,1,1,1,0
1,1,23,0,1,2,1,4,100,0,0,0,0,0,0,1,1,0,0,3,3,4,4,0
1,1,20,1,1,1,3,3,800,500,1,6,1,500,3,1,1,0,3,3,2,1
,0
1,1,21,0,3,1,4,2,600,100,1,4,8,320,3,2,0,0,3,2,3,1
,0
1,1,17,0,2,2,5,3,1000,0,1,3,4,250,4,2,0,0,3,4,1,2,
0
1,1,19,0,1,1,1,2,500,120,0,0,0,0,0,1,1,0,0,1,2,3,4,0
1,1,37,0,2,2,0,1,1200,0,1,6,7,1200,3,0,0,0,3,2,3,2
,0
1,1,20,1,1,1,4,3,1000,500,1,1,5,300,4,2,1,0,3,3,1,
1,1
1,1,22,1,3,2,4,3,1000,0,0,0,0,0,0,3,2,1,0,3,3,3,4,1
1,1,22,1,1,2,5,4,600,600,1,6,6,800,1,1,0,0,1,1,4,4
,1
1,1,21,0,3,2,4,2,1500,200,1,6,8,500,3,2,0,0,3,3,2,
2,1
1,1,19,0,1,1,4,1,800,0,0,0,0,0,0,1,1,0,0,1,1,1,3,0
1,1,20,1,1,1,4,4,650,0,1,2,8,200,4,1,0,0,1,3,1,1,0
1,1,19,1,1,2,1,4,300,0,1,1,4,20,4,2,0,0,3,1,1,1,0
1,1,26,0,2,1,3,4,800,200,1,7,5,600,4,1,1,0,1,3,1,1
,0
1,1,20,1,1,1,5,3,300,0,1,3,3,50,4,1,2,0,3,2,3,4,0
1,1,26,1,2,2,0,1,1200,0,1,3,3,60,4,2,1,0,4,3,3,4,0
1,1,19,1,1,2,1,3,600,150,1,2,6,150,1,2,4,4,3,2,1,3
,1
1,1,20,0,1,1,0,1,900,0,0,0,0,0,3,3,0,0,1,1,1,3,0
1,1,18,0,1,1,2,4,2000,0,0,0,0,0,0,4,1,0,0,1,1,4,4,0
1,1,23,1,1,1,5,4,200,0,0,0,0,0,0,4,1,0,0,1,2,4,4,1
1,1,18,1,1,2,4,3,1500,0,0,0,0,0,0,3,1,0,0,3,3,3,3,1
1,1,20,0,1,1,0,3,400,0,0,0,0,0,0,1,1,0,0,2,3,3,4,0
1,1,19,0,1,1,3,4,500,0,0,0,0,0,0,1,1,0,0,3,4,3,3,0
1,1,19,0,1,2,1,4,1300,0,0,0,0,0,0,4,2,0,0,4,3,3,3,0
1,1,18,0,1,1,1,2,1500,0,0,0,0,0,0,3,1,0,0,3,4,3,4,0
1,1,18,1,1,1,3,4,1000,0,0,0,0,0,0,3,1,0,0,3,3,1,1,1
4,4,20,0,3,1,1,4,200,0,1,5,5,350,1,1,0,2,4,2,2,1,1
4,4,20,0,3,1,2,4,200,0,1,6,6,400,3,1,0,1,3,3,3,3,1
4,4,26,0,1,1,0,4,200,0,1,2,6,600,4,1,0,0,1,1,2,1,1
4,4,23,1,3,1,5,4,600,100,0,0,0,0,3,1,4,1,3,4,2,2,1
4,4,27,1,3,2,2,3,100,0,0,0,0,0,3,1,4,0,3,3,3,3,1
4,4,24,1,2,2,0,1,0,0,1,5,2,600,3,1,0,1,3,3,2,2,0
4,4,20,1,1,1,1,4,200,0,1,7,8,200,4,1,0,0,3,3,1,1,0
4,4,21,1,3,2,2,3,100,0,0,0,0,0,3,2,0,2,3,3,2,3,0
4,4,20,1,1,1,0,4,500,100,1,3,8,120,4,2,0,0,4,3,4,1
,0
4,4,20,1,1,1,3,4,1500,100,0,0,0,0,4,2,0,4,3,4,1,1,
1
4,4,22,1,1,2,1,4,500,10,0,0,0,0,4,2,0,4,3,4,1,1,0
1,4,21,1,2,2,1,3,0,0,0,0,0,0,4,1,0,0,3,3,2,1,0
1,4,20,0,1,1,0,4,3500,0,0,0,0,0,3,1,0,0,3,3,2,1,0
1,4,24,0,2,1,0,4,300,0,1,5,5,450,3,1,0,1,1,1,1,2,0
1,4,20,0,2,1,0,4,0,0,1,6,4,500,3,1,0,4,4,4,2,3,0
1,4,21,0,1,1,0,4,1000,0,0,0,0,0,3,1,0,3,4,4,3,3,0
1,4,21,0,1,1,0,3,120,0,0,0,0,0,3,1,0,4,4,3,3,3,0
1,4,25,1,2,2,1,4,100,0,1,3,3,200,3,1,0,0,1,1,1,1,1
1,4,31,0,2,1,5,3,500,0,1,5,4,400,3,1,2,3,3,2,1,1,0
1,4,23,0,2,1,2,1,100,0,1,5,8,200,3,1,0,0,1,1,1,1,0
1,4,22,0,1,1,1,3,50,0,1,7,4,200,3,1,0,0,1,1,1,1,1
1,4,30,0,1,1,2,4,1500,50,1,7,4,400,1,1,0,1,3,3,1,1
,1
1,4,28,0,2,2,0,3,0,0,1,5,5,350,4,1,0,0,1,2,2,2,0
1,4,22,0,1,2,4,1,300,500,1,5,4,300,4,1,0,0,1,3,1,1
,0
1,4,20,0,2,1,2,4,800,0,0,0,0,0,4,1,0,0,1,1,1,1,1
1,4,21,0,2,1,2,3,50,100,1,6,8,900,3,0,0,2,4,3,1,3,
1
1,4,25,0,2,2,1,4,500,100,1,7,6,400,3,1,2,1,3,3,1,3
,1
1,4,25,1,2,1,2,4,0,0,1,5,6,400,4,0,0,0,4,4,1,3,0
1,4,27,0,2,2,0,4,0,0,1,6,8,900,3,1,2,3,4,3,1,3,1
1,4,23,1,1,1,2,4,40,0,0,0,0,0,4,1,0,0,1,1,1,1,1
1,4,21,0,1,1,0,4,700,0,0,0,0,0,4,1,0,0,3,1,1,1,0
1,4,34,1,1,2,0,4,0,0,0,0,0,0,4,1,0,1,3,3,1,1,1

3,4,20,0,1,1,1,3,0,1500,0,0,0,0,4,1,1,4,4,3,1,3,0
3,4,21,1,1,1,5,1,500,0,1,5,4,700,1,1,0,2,1,1,1,1,1
3,4,21,0,1,1,2,4,800,0,1,5,5,600,3,1,0,3,3,2,1,1,0
3,4,25,0,2,1,1,1,0,0,1,3,2,100,2,1,0,3,1,1,4,4,0
3,4,20,0,1,1,4,4,1800,0,1,1,8,350,3,2,0,1,2,3,1,1,
0
3,4,23,1,2,2,2,4,800,0,1,2,8,500,1,1,0,0,4,1,1,1,1
3,4,21,1,1,1,1,1,300,0,0,0,0,0,3,2,0,3,1,3,1,1,0
3,4,23,1,2,2,4,3,1200,0,0,0,0,0,4,1,2,1,3,4,3,1,0
3,4,20,1,1,1,5,4,1000,0,0,0,0,0,2,2,2,4,3,3,2,3,0
3,4,24,1,3,2,1,4,100,0,1,7,7,200,1,1,0,4,3,3,1,1,0
3,4,23,0,1,1,1,1,2000,0,0,0,0,0,1,1,0,0,1,1,1,1,1
3,4,22,0,1,1,4,4,100,0,0,0,0,0,4,1,0,0,3,4,3,2,0
3,4,24,1,3,1,2,4,3500,100,1,6,4,1100,4,3,0,2,3,3,
1,3,0
3,4,25,1,1,2,3,4,1000,500,0,0,0,0,1,1,0,4,1,3,4,1,
0
3,4,24,1,1,1,4,4,50,0,1,4,5,200,3,1,1,2,3,1,1,3,0
3,4,23,1,3,1,2,4,1600,930,1,5,4,550,4,1,1,2,1,3,1,
1,0
3,4,22,1,1,1,3,3,200,0,0,0,0,0,3,1,0,1,3,3,2,3,0
3,4,20,0,3,1,1,4,1000,0,1,4,2,150,1,1,0,0,3,3,1,1,
0
3,4,21,0,3,2,3,3,300,200,1,7,8,300,3,1,0,1,1,1,1,1
,1
3,4,21,0,3,1,1,1,300,0,1,2,7,300,3,1,0,1,1,1,1,1,1
3,4,26,0,2,1,4,3,800,0,1,6,5,500,1,1,0,1,1,3,1,1,0
3,4,23,0,1,1,3,3,500,0,0,0,0,0,1,1,0,2,3,2,2,3,0
3,4,23,0,3,2,0,3,200,0,1,5,5,450,2,2,0,2,1,1,1,1,0
3,4,23,0,2,2,1,1,200,0,1,4,6,465,1,1,0,2,1,1,1,1,1
3,4,22,0,1,1,0,3,300,0,0,0,0,0,3,2,0,0,1,2,1,1,0
3,4,20,1,3,2,5,3,1000,700,1,3,3,500,3,1,2,1,3,3,2,
3,0
3,4,25,1,3,2,2,4,400,500,1,4,7,800,1,1,0,2,1,1,1,1
,0
2,5,23,1,2,1,0,1,1200,300,1,5,4,400,1,1,0,1,1,1,1,
1,0
2,5,22,1,1,1,5,4,300,0,0,0,0,0,4,1,1,1,3,3,2,3,0
2,5,26,1,2,1,1,3,1000,0,1,7,5,20,4,1,0,0,1,1,1,1,0
2,5,22,1,3,1,1,4,0,0,0,0,0,0,3,1,0,2,3,3,2,2,0
2,5,21,1,3,1,3,2,1000,0,1,7,6,400,1,1,0,1,3,2,1,1,
1
2,5,21,0,1,1,4,4,600,0,0,0,0,0,3,1,0,4,3,3,3,3,0
2,5,23,0,2,2,2,3,350,0,0,0,0,0,3,1,0,1,2,1,1,1,0
2,5,22,0,1,2,2,4,250,0,1,2,8,400,3,1,0,1,1,3,1,1,0
2,5,24,0,2,2,4,4,350,0,1,7,6,250,4,1,0,0,3,1,1,1,0
2,5,26,0,1,1,3,4,1500,150,0,0,0,0,4,2,0,0,3,3,1,3,
0
2,5,25,1,1,1,5,4,300,0,1,4,2,100,1,1,2,4,1,4,1,1,0
2,5,30,0,3,1,1,4,0,51,0,0,0,0,3,1,0,0,1,1,1,1,1
2,5,21,0,3,1,2,1,50,0,0,0,0,0,3,1,0,0,1,1,1,1,0
2,5,21,0,2,1,1,4,100,0,1,4,4,1000,4,1,1,0,3,4,4,4,
0
2,5,19,0,2,1,5,2,100,0,0,0,0,0,1,1,1,1,3,3,3,1,0
2,5,21,0,1,1,2,2,200,0,1,5,5,300,3,1,1,0,1,2,2,1,0
2,5,21,1,2,1,5,3,150,0,0,0,0,0,3,1,0,0,3,3,3,3,0
2,5,20,0,1,2,5,3,1200,0,0,0,0,0,3,1,0,0,1,1,3,1,0
2,5,22,0,1,1,1,3,300,0,0,0,0,0,4,1,4,3,2,3,3,1,0
2,5,23,1,2,2,1,4,900,100,1,3,3,800,4,1,0,0,3,3,1,1
,0
2,5,18,1,1,1,4,4,200,0,1,4,6,100,4,2,0,0,3,4,3,1,0
2,5,21,0,1,1,4,3,500,0,0,0,0,0,4,2,0,0,3,3,1,1,0
2,5,28,0,2,2,2,1,100,0,0,0,0,0,3,1,0,0,1,2,1,1,0
2,5,28,1,3,1,2,2,3500,0,0,0,0,0,3,1,0,0,3,3,1,1,0
2,5,26,0,2,2,0,1,0,0,1,5,6,300,1,1,0,0,4,4,3,3,0
1,5,21,0,1,2,2,3,300,0,1,2,8,60,3,1,0,0,3,3,3,4,0
1,5,22,1,3,1,3,1,200,0,1,4,3,100,1,1,3,4,1,1,1,1,0
1,5,23,0,1,1,4,4,1500,300,0,0,0,0,2,1,0,4,3,3,2,1,
0
1,5,24,0,2,1,1,1,1000,0,1,6,5,200,1,1,2,1,1,1,1,1,
1
1,5,22,0,1,2,1,4,300,100,1,3,6,300,3,1,4,4,4,3,3,3
,0
1,5,21,0,1,2,2,3,100,50,0,0,0,0,1,1,4,4,2,1,1,1,0
1,5,28,0,1,1,1,4,930,0,1,5,4,100,4,1,1,0,4,3,3,4,0
1,5,21,0,1,1,4,3,100,0,1,6,4,200,1,1,0,2,1,1,1,1,0
1,5,20,0,1,1,4,1,600,80,1,7,6,750,1,1,0,3,1,4,4,1,
0
1,5,32,1,2,1,5,3,300,0,1,7,5,300,2,1,0,0,2,1,1,3,0
1,5,27,1,2,2,0,1,100,0,1,5,8,100,4,1,0,4,4,1,4,4,0
1,5,21,0,1,1,5,4,500,0,1,6,4,100,4,1,0,0,3,3,1,1,0
1,5,20,0,1,1,3,4,150,200,0,0,0,0,4,1,0,1,3,3,3,3,0
1,5,20,0,1,2,2,3,450,0,0,0,0,0,3,0,0,3,2,1,2,3,1
1,5,32,0,2,1,0,4,1000,0,1,7,4,3000,4,1,4,4,4,4,1,1
,0
1,5,28,0,1,2,0,1,1000,0,0,0,0,300,2,0,4,4,2,4,1,1,
0
1,5,22,1,1,1,1,1,300,0,0,0,0,100,4,2,1,1,1,1,3,3,0
1,5,23,1,3,1,1,4,1300,500,1,2,4,400,4,1,0,1,3,4,3,
3,0
1,5,20,1,2,1,3,1,50,0,1,2,8,100,3,0,0,0,3,2,1,1,0
1,5,23,1,3,1,5,4,5000,1000,1,5,4,250,4,1,1,4,3,3,
3,3,0
1,5,20,1,3,2,0,3,500,0,1,4,5,400,1,2,0,1,3,3,1,1,0
1,5,22,1,2,1,0,4,200,200,1,7,8,2100,3,0,0,1,2,3,1,
1,0
1,5,21,1,3,2,0,4,0,0,1,2,8,400,3,1,0,1,3,4,3,3,0
1,5,25,1,2,2,5,3,10000,0,1,2,8,1200,2,1,0,0,1,2,1,
1,1
1,5,21,1,1,2,4,1,8200,0,1,2,4,100,1,1,2,2,4,1,2,1,
0
1,1,20,0,3,2,4,1,950,0,1,7,6,700,1,0,4,0,3,1,1,1,1
1,1,23,1,2,1,5,1,500,0,1,7,5,450,1,1,4,0,1,1,1,1,1
1,1,18,1,1,1,3,3,4000,0,1,2,8,800,3,1,1,0,3,3,1,3,
1
1,1,39,0,2,2,0,1,0,0,1,6,8,2500,3,0,2,0,1,3,3,1,1
1,1,19,0,3,2,4,1,1500,200,1,5,8,750,3,0,3,0,2,2,1,
1,1
1,1,21,1,3,1,3,1,1200,500,1,6,6,600,1,0,2,0,3,3,1,
2,1
1,1,20,0,1,1,5,1,2000,0,1,4,8,500,1,1,3,0,3,1,2,1,
1
1,1,20,0,3,2,5,1,1500,0,1,5,5,600,2,0,4,0,1,2,2,3,
1
1,1,20,0,1,1,5,3,2500,0,0,0,0,0,4,2,0,0,3,3,2,3,0
1,1,34,0,2,1,2,3,2000,0,1,6,6,1200,3,0,0,0,2,3,3,3
,0
1,2,21,0,1,1,4,3,1800,500,1,2,8,400,3,1,1,0,2,1,3,
1,0
1,2,24,0,3,2,3,1,850,100,1,5,8,550,1,0,4,2,1,2,1,1
,1
1,2,20,1,3,1,3,1,1800,0,1,2,8,450,3,1,3,1,1,3,3,3,
0
1,2,21,0,1,1,2,3,3000,0,0,0,0,0,3,2,0,0,3,3,3,3,0
1,2,30,0,2,1,2,1,2800,0,1,6,5,750,2,0,0,0,3,3,3,3,
0
1,2,21,0,3,1,4,2,1500,0,1,7,6,800,3,0,4,1,3,3,3,1,
0
1,2,22,0,3,2,4,1,1000,0,1,6,6,600,3,1,2,1,2,1,1,1,
1
1,2,21,0,3,1,2,3,800,0,1,2,8,400,3,1,4,2,1,1,1,1,0
1,2,21,1,2,1,3,3,3200,0,0,0,0,0,4,2,0,1,3,3,3,3,0
1,2,24,0,3,1,2,2,1500,0,1,6,5,500,4,1,2,2,3,3,2,1,
0
1,2,20,1,2,1,5,1,800,0,1,6,7,800,1,0,2,1,2,2,2,2,0
1,2,20,1,3,1,3,1,1000,0,1,5,8,850,4,0,3,1,1,2,1,1,
0
1,2,22,0,3,2,1,2,1000,0,1,5,6,500,3,1,3,1,3,1,3,3,
1
1,3,22,1,3,1,0,1,1000,0,1,6,5,500,3,0,4,1,1,3,3,3,
0
1,3,23,1,3,1,4,2,1800,0,1,2,8,400,3,1,0,0,3,3,1,1,
0
1,3,24,1,3,1,2,3,2500,0,1,2,8,420,3,2,1,1,2,2,3,3,
1
1,3,24,0,1,1,3,4,2000,250,0,0,0,0,3,2,0,0,3,3,3,3,
0
1,3,23,1,3,1,3,3,3000,0,0,0,0,0,3,2,0,0,3,1,3,3,0
1,3,20,0,3,1,2,4,2400,0,1,2,8,600,3,1,1,0,3,3,1,1,
0
1,3,21,1,2,2,3,1,0,0,1,6,8,850,3,0,4,2,1,1,1,1,0

1,3,20,0,1,1,1,3,2000,0,1,1,8,400,3,2,0,0,3,3,3,3,
0
1,3,21,0,3,1,5,2,1000,0,1,3,8,250,3,1,2,0,3,1,1,1,
0
1,3,22,0,3,2,5,4,1500,0,0,0,0,0,0,3,1,0,0,3,3,3,3,0
1,3,22,0,2,2,5,1,0,0,1,6,8,1000,3,0,3,2,1,1,1,1,0
1,3,23,1,3,1,4,4,2000,0,1,3,8,600,3,1,1,1,3,1,1,1,
1
1,3,21,1,3,1,5,1,1000,0,1,5,4,500,1,0,4,2,1,1,1,1,
0
1,3,20,0,1,1,3,3,2500,0,0,0,0,0,0,3,2,0,0,4,4,4,4,0
1,3,21,0,1,1,2,4,2500,0,0,0,0,0,0,4,2,0,0,4,4,4,4,1
1,3,20,1,2,2,5,1,800,0,1,6,6,1200,3,0,4,1,3,3,1,1,
0
1,3,22,0,2,2,2,1,1500,0,1,6,6,900,1,0,4,2,1,1,1,1,
0
1,3,20,1,1,1,3,4,2500,0,1,2,8,400,4,2,0,0,3,3,3,3,
1
1,3,21,0,2,1,2,1,1000,0,1,6,6,500,1,0,4,2,1,1,1,1,
0
1,3,23,0,1,1,5,3,1500,0,1,2,8,600,3,2,0,0,3,3,3,3,
0
1,3,20,0,1,2,3,3,1200,0,1,5,5,600,3,1,2,0,3,3,3,3,
0
1,4,23,0,1,1,3,4,2500,0,0,0,0,0,0,3,1,0,0,3,3,4,3,0
1,4,22,0,2,2,5,1,0,0,1,6,8,950,1,0,3,0,1,1,1,1,0
1,4,25,0,1,2,2,1,1800,0,1,2,8,600,3,1,4,2,3,3,3,3,
0
1,4,23,1,1,1,3,1,2000,200,1,5,8,600,3,1,0,0,3,3,3,
3,0
1,4,24,0,1,1,1,4,2000,0,1,1,8,240,3,2,0,0,3,3,3,3,
1
1,4,23,0,1,1,3,4,3000,0,1,2,8,220,4,2,0,0,4,4,4,4,
0
1,4,24,0,1,1,2,1,1200,0,1,5,6,700,1,0,2,1,1,1,1,1,
0
1,4,25,1,1,2,2,1,2000,0,1,6,8,1000,3,0,0,0,3,3,3,3
,0
1,4,20,0,1,1,2,3,2500,0,0,6,4,0,3,1,1,0,3,3,3,3,0
1,4,20,1,1,1,4,3,3500,0,0,7,5,0,4,1,0,0,4,4,4,4,1
1,4,21,0,2,2,4,1,1000,0,1,7,6,800,1,1,1,0,1,1,1,1,
0
1,4,21,1,1,1,2,3,2500,0,0,5,5,0,3,1,0,0,1,1,1,3,0
1,4,20,1,1,1,4,4,3000,0,0,5,4,0,4,1,0,0,3,3,3,3,0
1,4,20,1,2,1,4,1,1000,0,1,6,6,600,1,0,2,1,1,1,1,1,
0
1,4,24,0,1,1,3,3,1200,0,1,6,5,650,3,0,0,2,3,3,3,3,
0
1,4,23,1,1,1,2,3,2500,0,0,5,6,0,3,1,2,0,3,3,3,3,0
1,4,24,0,2,2,2,1,1000,0,1,6,8,1000,1,0,2,0,3,1,1,1
,0
1,4,22,1,1,1,4,3,2500,0,0,7,6,0,3,1,0,0,3,3,4,3,0
1,4,21,0,1,1,3,3,2500,0,0,5,4,0,3,1,0,0,4,3,3,3,0
1,4,22,0,1,1,3,3,2200,250,0,6,6,0,3,1,0,1,3,3,3,3,
0
1,4,21,0,1,1,3,1,1800,0,0,2,4,0,1,1,1,0,3,3,3,3,1
1,4,20,0,1,1,5,4,2200,0,1,5,5,750,3,1,1,0,3,3,3,3,
0
1,4,21,0,1,1,3,1,1200,0,1,6,6,750,3,1,2,0,3,3,3,3,
0
1,4,19,0,1,1,2,4,2500,0,0,0,0,0,4,2,2,2,4,4,4,4,0
1,4,21,0,1,2,2,1,1200,0,1,2,8,200,3,1,0,0,3,3,3,3,
0
1,4,21,0,2,2,2,1,1000,0,1,5,8,600,1,0,4,1,4,1,1,1,
0
1,4,21,0,1,1,3,3,2400,0,1,2,8,220,3,2,0,0,3,2,2,2,
0
1,4,22,0,1,1,1,4,2800,0,0,0,0,0,0,2,2,0,0,3,3,3,3,0
1,4,20,1,1,1,3,3,2000,0,1,5,5,600,4,1,0,0,3,3,3,1,
0
1,4,21,1,1,1,4,1,1500,0,1,6,8,850,1,0,4,2,1,1,1,1,
1
1,4,22,0,1,1,3,3,1800,0,1,6,6,500,3,1,0,0,3,3,3,3,
0
1,4,20,1,2,2,4,1,1000,0,1,6,8,650,1,0,4,2,1,1,1,1,
1
1,5,22,0,1,1,4,3,1200,0,1,2,8,280,3,2,0,0,3,3,3,3,
0
1,5,21,1,1,1,2,3,1800,150,0,0,0,0,4,1,0,0,3,3,3,3,
0
1,5,21,0,1,2,3,3,2000,0,0,0,0,0,3,2,0,0,3,3,4,4,0
1,5,21,0,1,1,0,4,1800,0,1,1,8,200,2,1,2,1,4,4,4,4,
0
1,5,22,0,1,1,1,1,1600,0,1,2,8,320,2,1,4,3,1,1,1,1,
1
1,5,22,0,1,1,3,3,1400,0,1,6,4,500,3,1,2,2,3,3,3,3,
0
1,6,23,0,1,1,2,3,1500,0,1,2,5,250,3,1,1,2,3,3,3,3,
1
1,6,24,1,2,1,2,3,1200,0,1,5,5,550,3,1,0,0,3,3,3,3,
0
1,6,23,0,1,1,3,4,1800,0,1,2,8,380,4,1,1,1,4,4,4,4,
0
1,6,24,0,1,1,2,4,2500,0,0,0,0,0,3,2,1,3,3,3,3,3,0
1,6,21,0,1,1,2,3,2400,0,0,0,0,0,3,2,0,2,3,3,1,3,0
1,6,20,0,1,1,3,3,2800,0,0,2,5,0,3,1,0,2,3,3,2,2,0
1,6,20,0,1,1,1,3,1500,0,0,3,5,0,4,1,0,2,4,4,4,4,0
1,6,22,0,1,2,1,1,1200,100,1,2,6,200,2,0,2,3,4,3,3,
2,0
1,6,23,1,1,2,2,2,1500,0,1,2,5,350,3,1,1,2,1,3,2,2,
0
1,6,20,0,1,1,3,3,1800,0,0,0,0,0,3,2,0,0,3,3,3,3,0
1,6,21,0,1,1,2,3,2400,0,0,0,0,0,3,2,2,2,3,3,3,3,0
1,6,21,0,1,1,4,3,1800,0,1,2,5,150,3,2,0,2,3,3,3,2,
0
1,6,24,0,1,1,3,3,1000,250,1,5,4,400,3,1,1,0,3,3,3,
3,0
1,6,23,1,1,2,2,3,1200,1000,1,6,5,0,3,1,1,2,3,3,2,2
,0
1,6,24,0,1,1,2,3,1200,0,1,2,8,280,3,1,0,1,3,3,1,1,
0
1,6,25,0,1,1,2,4,2000,0,0,6,4,0,3,1,1,3,3,3,4,3,0
1,6,22,1,1,1,3,3,1900,0,1,2,8,200,3,1,1,3,3,3,3,3,
0
1,6,23,1,1,1,1,3,1600,0,1,2,8,250,3,1,1,0,3,3,3,3,
0
1,6,21,0,1,2,1,3,1200,4,1,1,8,280,4,1,0,4,3,3,3,2,
0
1,6,23,1,1,1,3,3,1800,0,1,1,8,270,3,2,0,0,3,4,3,3,
0
1,6,23,0,1,1,4,3,2000,0,0,0,0,0,4,2,2,2,3,3,4,4,0
1,6,23,0,1,1,2,3,1800,0,1,2,8,300,3,1,0,3,3,3,3,3,
0
1,6,23,0,1,1,3,3,1600,0,1,6,4,0,3,1,0,2,3,3,3,4,0
1,6,22,0,1,1,1,4,1500,0,1,2,8,250,3,2,1,4,3,4,2,2,
0
2,2,18,1,2,1,5,1,1000,0,1,6,6,600,1,0,4,1,1,1,1,1,
1
2,2,19,1,1,1,4,1,800,0,1,6,6,650,1,0,3,1,3,1,3,1,1
2,2,20,0,1,1,2,4,2000,0,0,0,0,0,4,2,0,0,3,3,3,3,0
2,2,18,0,1,1,3,4,1800,0,1,2,8,240,3,2,1,0,3,3,4,4,
0
2,2,19,0,1,1,4,3,1600,0,0,6,8,0,3,2,1,1,3,3,3,3,0
2,2,19,1,1,1,3,1,1000,0,1,2,8,250,1,1,2,1,3,1,1,1,
1
2,2,20,0,1,2,1,1,800,0,1,5,6,450,1,0,3,2,3,1,1,1,1
2,2,19,1,1,1,2,4,1900,0,1,1,8,200,4,2,0,0,3,3,3,3,
0
2,2,20,0,1,1,4,3,1500,0,0,0,0,0,3,2,1,1,3,3,3,3,0
2,2,19,1,2,1,5,1,750,0,1,6,6,550,1,0,2,2,1,1,1,1,1
2,2,19,1,1,1,3,3,500,0,0,5,7,0,1,0,2,0,1,1,1,2,0
2,2,19,1,2,1,3,1,1000,0,1,6,6,500,1,0,3,2,1,1,1,1,
1
2,2,19,0,1,1,2,3,650,0,1,2,8,300,1,2,1,0,3,1,1,1,0
2,2,19,0,1,1,4,1,500,0,1,5,6,400,1,0,2,1,3,1,3,3,1
2,2,18,0,1,2,3,1,1000,0,1,2,8,350,1,2,3,2,1,1,1,1,
1
2,2,19,0,1,1,5,3,1600,0,0,0,0,0,4,2,0,0,3,3,3,3,0
2,3,22,0,1,1,2,4,1200,0,1,1,8,240,3,2,1,0,3,3,3,3,
0
2,4,21,1,1,1,3,3,1500,0,1,6,6,500,3,1,0,1,3,3,3,2,
0

2,4,20,1,1,1,1,2,1800,0,0,0,0,0,3,2,0,0,3,3,3,3,0
2,4,21,0,1,1,4,3,2200,0,0,0,0,0,3,2,1,1,3,3,3,3,0
2,4,21,1,1,1,1,3,1600,0,1,6,5,300,3,1,1,1,4,4,3,2,
0
2,4,20,0,1,1,3,3,1700,200,1,6,6,0,3,1,1,2,3,3,3,4,
1
2,4,21,1,1,1,4,4,1200,0,1,6,6,450,4,0,1,1,3,4,3,1,
0
2,4,21,1,1,1,1,2,1500,0,1,2,8,400,3,2,0,0,2,3,4,4,
0
2,4,21,0,1,1,3,1,600,0,1,6,6,450,1,0,3,3,1,1,2,2,1
4,6,22,1,1,1,4,3,1500,0,1,2,8,400,3,1,1,2,3,1,1,1,
1
4,6,23,1,1,1,2,3,1000,0,1,5,4,800,3,1,2,2,4,3,3,2,
0
4,6,22,1,1,1,2,4,900,0,1,2,4,320,4,1,2,3,3,3,3,1,0
4,6,24,0,1,1,3,3,1200,0,0,0,0,0,3,2,0,0,4,4,1,4,0
4,6,23,0,1,1,1,4,1600,0,0,0,0,0,4,2,2,2,3,3,3,3,0

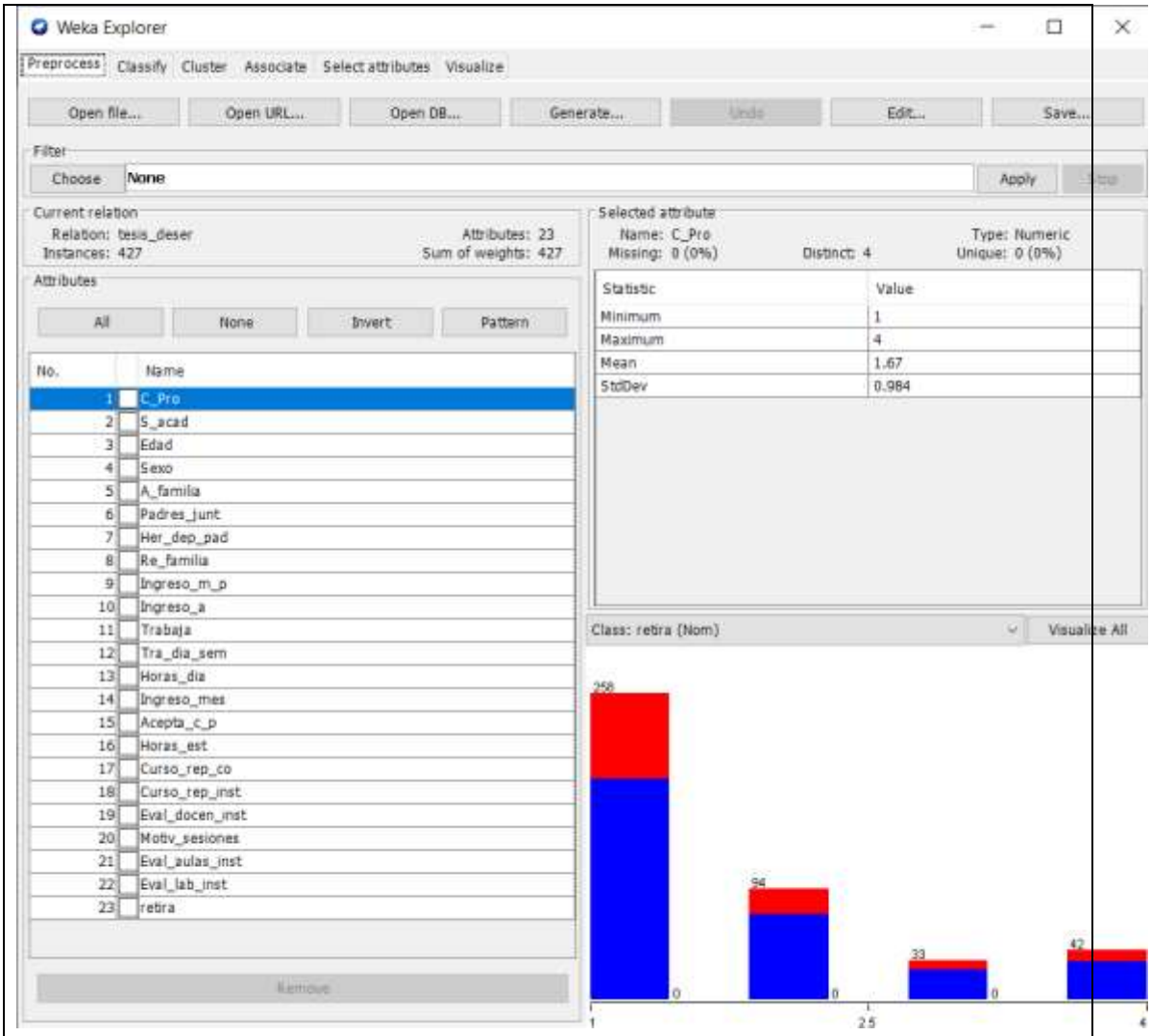
Anexo 2. Proyección del total de atributos recogidos



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

FECHA: Octubre 2020.

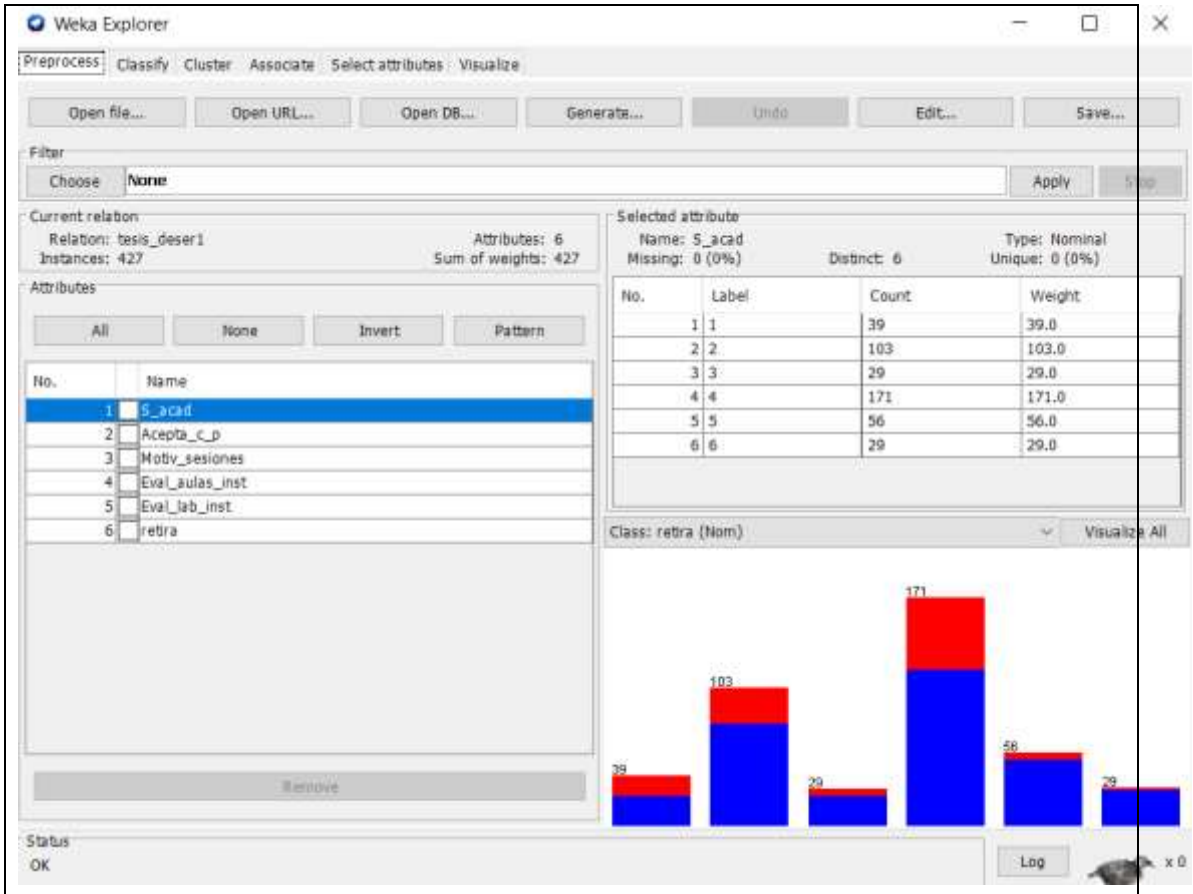
Anexo 3. Proyección de atributos codificados



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.

Este libro se terminó de publicar en la editorial

Instituto Universitario
de Innovación Ciencia y Tecnología Inudi Perú

INUDI PERÚ

INSTITUTO UNIVERSITARIO DE INNOVACIÓN
CIENCIA Y TECNOLOGÍA

ISBN: 978-612-5069-42-9

9 786125 069429

EDITADA POR
INSTITUTO
UNIVERSITARIO
DE INNOVACIÓN CIENCIA
Y TECNOLOGÍA INUDI PERÚ