Facultad de Ingeniería

Carrera de Ingeniería de Sistemas e Informática.

TRABAJO DE INVESTIGACIÓN:

Diseño de un sistema para predecir la deserción de los alumnos mediante Machine learning en la Universidad Tecnológica del Perú

Autores

Perez Bedia, Carlos Andres - 1220317

Rojas Segovia, Luis Enrique - 1627749

PARA OPTAR POR EL GRADO ACADÉMICO DE BACHILLER EN
INGENIERÍA DE SISTEMAS E INFORMÁTICA

Asesor

MSc. Mamani Ticona, Wilfredo

Lima, Perú

Agosto 2020

DEDICATORIA

Este trabajo va para mi madre, que fue mi gran apoyo y empuje a lo largo de mi crecimiento profesional y personal, así como también a toda mi familia.

Pérez Bedia, Carlos Andres.

Este trabajo va dedicado a mis padres por su ayuda constante que me brindan para continuar con cada paso de nuestra vida académica y personal.

Rojas Segovia, Luis enrique

AGRADECIMIENTO

Agradecimiento especial a nuestras familias, por el constante apoyo para poder llegar hasta a la fase final de mi carrera profesional, además por su ayuda en momentos duros de mi vida profesional, ayudándome siempre a ser constante para lograr nuestros objetivos.

Pérez Bedia, Carlos Andres.

A la UTP, por brindarnos las herramientas de tanta utilidad para nuestro aprendizaje constante, así como darnos un ambiente grato para la convivencia universitaria sin dejar de lado los objetivos de un universitario.

A nuestros maestros, que con sus enseñanzas nos guiaron hasta estas etapas de nuestras carreras y sobre todo nos prepararon para enfrentar una vida profesional con muchos retos y dar lo mejor de nosotros para lograr nuestras metas.

Rojas Segovia, Luis enrique

RESUMEN

Este trabajo de investigación permitió predecir la deserción de alumnos de la sede central de la UTP ubicado en la ciudad de Lima, teniendo como base de aprendizaje los alumnos que desertaron en semestres anteriores e implementando la tecnología de machine learning, que proporcionaron los patrones de comportamiento de los alumnos desertores y mediante las herramientas de machine learning se pudo obtener de manera anticipada los alumnos con potencial de deserción y de esta manera la Universidad Tecnológica del Perú pueda intervenir y evitar el abandono de estudios.

El aporte de esta investigación permitió identificar patrones de comportamiento que son de gran importancia para la Universidad Tecnológica del Perú, debido a que, al tener identificado a los estudiantes con intención de desertar de sus estudios, les permitirá plantear estrategias que permitan disuadir al estudiante y orientar sobre su situación académica.

La metodología utilizada fue la de Crisp dm y el algoritmo Support Machine Vector (SVM) permitió predecir por medio de comportamiento patrón de los alumnos predecir su decisión de desertar.

Palabras clave: Machine learning, aprendizaje automático, deserción universitaria, métricas de machine learning, algoritmos de predicción.

ABSTRACT

The contribution of this research will be of great importance for the UTP, because, having identified the students with the intention of dropping out of their studies, it will allow them to propose strategies that will dissuade the student and provide guidance on their academic situation.

The methodology used was that of Crisp dm and the Support Machine Vector (SVM) algorithm to predict through standard behavior of the students to predict their decision to drop out.

Keywords: Machine learning, University desertion, machine learning metrics, prediction algorithm.

ÍNDICE

INDICE DE FIGURAS

INDICE DE LAS TABLAS

INDICE DE ECUACIONES

INTRODUCCIÓN

En la actualidad machine learning o aprendizaje automático cuenta con muchas aplicaciones a nivel empresarial, medico, transporte, entre otros campos importantes como la educación, que tiene como objetivo brindar a los algoritmos la capacidad de aprender, partiendo de información previamente encontrada para analizarla y detectar patrones de comportamiento. Para finales del año 2018 el banco mundial realizó un estudio que demostró que el 30% de los alumnos peruanos de universidades abandonan sus estudios, del que la Universidad Tecnológica del Perú no es ajena, generando así un círculo de desempleo y pobreza que acarrea más problemas sociales como consecuencia.

Es por ello que en el primer capítulo de este trabajo investigación se plantea el problema de la deserción y las posibles soluciones, identificando objetivos e hipótesis delimitando la investigación. Así mismo en el capítulo 2, se revisa sobre los antecedentes de la deserción del alumnado de distintas entidades educativas del mundo. Para tener un enfoque más específico de las causas y de esa manera utilizar las herramientas de machine learning para predecir dicha decisión.

En el capítulo 3 se definirá la metodología de acuerdo al marco teórico y el porcentaje de uso que tengan en investigaciones relativamente actuales. Para en el capítulo IV elaborar el desarrollo con cada fase de la metodología y mostrar los resultados en el capítulo V.

CAPITULO I: PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción de la problemática

El problema de la deserción universitaria es a nivel mundial, sin embargo tiene más impacto negativo en sociedades más vulnerables en la estructura socioeconómica, que frenan las oportunidades de crecimiento y generan pobreza y otros problemas sociales, por ello es importante abordar este tema a fin de contrarrestar las consecuencias negativas que presentan en la sociedad.

Un estudio del Banco Mundial, a fines del año 2018, reveló que el 30% de los universitarios peruanos abandonan su carrera por diferentes motivos, asimismo el porcentaje en otros países de américa latina llega a 42% de los alumnos que desertan de sus estudios, evidenciándose así un gran problema social que trae consecuencias negativas a la sociedad. De igual manera este problema está presente en las universidades nacionales y privadas del Perú, como ocurre en la UTP donde existe un porcentaje de alumnos que eligen desertar de sus estudios, por diferentes factores, como económicos, académicos, vocacionales e incluso problemas de adaptación.

En este trabajo de investigación se buscará encontrar, a partir de los datos históricos de estudiantes que desertaron, un patrón de comportamiento que permita identificar a los alumnos con intenciones de desertar de sus estudios y de esta manera, las autoridades universitarias planteen estrategias que permitan orientar al alumno para continuar con sus estudios y brindarles las facilidades posibles para que se mantenga el vínculo con el

centro de estudios y pueda contar con las oportunidades de desarrollo profesional que los centros de educación superior brindan.

1.2 Formulación del problema

1.2.1 Problema general

¿Cómo predecir la deserción estudiantil en la UTP mediante aprendizaje automático?

1.2.2 Problemas Específicos

- ¿Qué motivos alteran la permanencia de los alumnos en la UTP?
- ¿Cómo un sistema basado en machine learning puede predecir la deserción de estudiantes en la UTP?
- ¿Cómo garantizar la fiabilidad del sistema de predicción de deserción de los alumnos?

1.3 Objetivos

1.3.1 Objetivo general

Diseñar un sistema de predicción de desertores de la universidad tecnológica del Perú, mediante machine learning.

1.3.2 Objetivos específicos

- Precisar los factores que lleva a la deserción del alumnado en la UTP.
- Diseñar un sistema que predice la deserción en la UTP basado en aprendizaje automático.

- Determinar la confiabilidad del sistema de predicción, mediante las métricas de machine learning.

1.4 Hipótesis

1.4.1 Hipótesis general

El sistema de predicción basado en machine learning permite reconocer a alumnos que pueden desertar de sus estudios en la UTP.

1.4.2 Hipótesis especificas

- El algoritmo utilizado puede identificar los motivos y patrones de los estudiantes que deciden dejar los estudios.
- La implementación de la técnica Support vector machine podrá influir significativamente en la reducción de desertores.
- El diseño del sistema de predicción de deserción estudiantil tiene un grado de certeza superior al 80%.

1.5 Justificación de la investigación

1.5.1 Teórica

En esta investigación se podrá identificar los posibles casos de deserción estudiantil, y de esta manera contribuir a conocer la importancia del uso de un sistema de predicción capaz de brindar información real y oportuna a los responsables.

1.5.2 Práctica

Existe un gran porcentaje de alumnos que deciden retirarse de su centro de estudios, el cual podrá apoyar a la educación en base a las respuestas de estudiantes que han desertado, así poder tener una mayor cobertura e información para disminuir la tasa de deserción mediante las estrategias elaboradas por la alta directiva.

1.5.3 Metodológica

Para el logro de las metas de este trabajo, se empleó distintas técnicas de machine learning que permitan conocer los factores con mayor probabilidad de deserción estudiantil.

1.6 Delimitación del estudio

1.6.1 Espacial

Este estudio se desarrollará en la UTP, en la sede centro ubicado en Lima.

1.6.2 Temporal

Los datos considerados para realizar el trabajo de investigación se encuentran entre las fechas de enero del 2017 a diciembre del 2019.

1.6.3 Conceptual

- Predicción: es la posibilidad de anticipar algo que va a suceder.
- Deserción Estudiantil: es la problemática en donde un alumno deja los estudios por diferentes motivos.
- Machine learning o Aprendizaje automático: Permite brindar a las maquinas la habilidad de poder aprender e identificar patrones automáticamente.

CAPITULO II: MARCO TEÓRICO

2.1 Antecedentes de la investigación

Actualmente la deserción de estudiantes de universidades es un problema latente en el Perú y Latinoamérica, a fines del año 2018 el banco mundial que es una asociación global que busca disminuir la pobreza, identificó que el 30% de estudiantes de las universidades peruanas desertan de sus estudios, siendo este un gran problema social, generando así un círculo de desempleo y pobreza que acarrea más problemas sociales como consecuencia. Por tanto, el presente trabajo pretende identificar ¿Cuál es la técnica de machine learning que mejor se adapta a la predicción con patrones de comportamiento?

Himmel (2012) encuentra 5 categorías principales que son factor psicológico, factor económico, factor sociológico, factor organizacional y factor de interacciones. Esto significa que al existir varios factores que llevan al abandono de estudios y para evitar esto es necesario implementar una herramienta que permita anticiparse al hecho y evitar así el abandono de estudios.

De esta manera, Villamarín en el año 2017 realizó un estudio que ayuda reconociendo de manera anticipada a los estudiantes que presenten problemas y puedan retirarse de los estudios, usando herramientas de machine learning. De esta manera permitirá a través de consultas o sentencias obtener una colección de datos para así poder clasificarlos y ordenarlos con la finalidad de determinar con que variables trabajar para luego ejecutar los mapas auto organizados con el software Matlab.

Una vez se realizó todo ese proceso pudo ejecutar y analizar los resultados obtenidos por los mapas para poder descubrir comportamientos que podrían coadyuvar a entender el fenómeno de deserción y así proponer acciones que puedan minimizar el riesgo que existe.

Por otra parte, se encuentra el caso de estudio realizado por Forero, Piñeros y Rodríguez (2019) que utiliza Azure Machine Learning Studio que permite crear e implementar soluciones de análisis predictivos, así como la manipulación de datos y los tipos de resultados, el cual se utilizó después de realizar un filtro de datos necesarios de la universidad objeto de su estudio.

Luego de la clasificación de algoritmos que han podido pasar por la evaluación de la herramienta de validación cruzada que proporciona este software y que puedan cumplir con la base del objetivo practico. Entre las conclusiones se puede decir que con los resultados obtenidos la deserción se debe a razones académicas o sociales del estudiante, confirmando que con las herramientas de machine learning se puede predecir si un estudiante desertará o no, siendo los algoritmos de aprendizaje de clasificación binaria capaces de replicarse en varias áreas con diferentes objetivos.

Precisamente García (2019), pretende poder estudiar el patrón de comportamiento de los alumnos, usando los datos que serán extraídos de la base de datos Oracle y así tener limpio los datos obtenidos para poder exportarlos en Excel, con este archivo obtenido se podrá pasar al entrenamiento del modelo mediante el lenguaje de programación Python y sus bibliotecas como Scikit-Learn, Pandas, Matplotlib y Numpy con el algoritmo XGBoost y por último el modelo podrá mostrar los resultados obtenidos. De esta manera

se recomienda entrenar al algoritmo directamente con la base de datos y no con un archivo ya que se podría ahorrar tiempo en este proceso.

Asimismo, el autor Candia en el año (2018) quien investigó sobre la predicción de rendimiento académico a partir de los propios datos del estudiante usando algoritmos de machine learning, además utiliza la metodología CRISP-DM que ayuda a encontrar un comportamiento predictivo. Esto muestra que esta metodología tiene grandes ventajas sobre los modelos predictivos debido a que cuentan con 6 importantes para este tipo de investigación.

Por ello Candia, recomienda el uso del algoritmo de árbol de decisión ya que es el más simple de comprender y de implementar y el óptimo para un modelo de predicción den machine learning.

En el año 2012 Fischer, hace referencia al método SEMMA que significa en ingles Sample, explore, Modify, Model, Access) que es usado para manejar mucha cantidad de datos con el fin de encontrar coincidencias que permita definir patrones, Además realiza una comparación entre dicha metodología y la de CRISP-DM indicando que esta última tiene muchos aspectos positivos en cuanto a la predicción de deserción por sus fases ya establecidas.

Lo expresado en líneas arriba muestra la clara preferencia que existe por la metodología CRISP-DM por algunos autores que investigan sobre la predicción del riesgo de deserción de estudiantes universitarios. Sin embargo, existen otras principales metodologías que los autores de esta investigación abordaran.

Por otro lado, En mayo de 2016, los autores Ordoñez Rosa Anela & Pastor Maria, en sus estudios sobre la deserción muestra que utilizan el algoritmo SVM y enfatizan que dicho algoritmo está más enfocado en aplicaciones que tengan que ver con datos reales y en una gran cantidad ya que al brindarle al sistema SVM la información requerida y clasificada este permite discriminar los datos y clasificarlos de acuerdo a los resultados obtenidos. Además, el algoritmo SVM es actualmente considerado como referente en el ámbito estadístico y machine learning.

Se evidencia que los autores, tiene preferencia por la metodología CRISP-DM, ya que cuenta con fases de tratamiento de la data para clasificarla y que sea óptima para el algoritmo de elección. Así mismo Miranda (2019), en su investigación hace uso de los algoritmos de SVM por dichos algoritmos buscan obtener la estructura optima del modelo como sus valores óptimos de sus parámetros para una medición más confiable.

De las investigaciones mostradas se ha podido reafirmar que machine learning cada vez está siendo más importante no solo para el ambiente estudiantil sino también para los distintos campos profesionales, ya que cuenta con varias herramientas capaces predecir comportamientos.

La fuente que más se adaptan a la línea de esta investigación es la brindada por Candia, quien recomienda utilizar la metodología CRISP-DM por sus grandes ventajas en comparación con otras metodologías.

De igual manera la investigación que aporta mucho a este documento es el realizado por García quien recomienda el uso de lenguajes de programación como Python y sus bibliotecas como Scikit-Learn, Pandas, Matplotlib y Numpy, para realizar el machine

learning debido a su performance que se alinea con lo que se busca en este trabajo de investigación.

Asimismo, de las fuentes consultadas, se desprende que la metodología que más es usada en predicción es la metodología CRISP-DM ya que cuenta con 6 fases importantes como mostró en el estado del arte, que permiten tener un amplio conocimiento del área que se desea investigar y con el cual los autores de esta investigación encontraran apoyo alineado con el fin de este documento.

Por tanto, este documento tiene por objetivo diseñar un sistema que utilice machine learning con la metodología CRISP-DM que permita predecir los casos de riesgo y mitigar con el problema social que atraviesan muchas estudiantes en la actualidad y que significa un grave problema para la sociedad.

En cuanto a los algoritmos utilizados por los autores para predecir la deserción, van desde los algoritmos de árbol de decisión hasta los algoritmos Support Machine Vector, este último tiene una mayor aceptación por los autores de tesis en cuanto a predicción y machine learning se trataba, debido a sus características de clasificación y análisis de regresión.

2.2 Bases Teóricas

2.2.1 Deserción estudiantil

La deserción estudiantil es una problemática genera diversos impactos negativos, no solo para el alumno y el centro de estudios, sino también para el ámbito social, ya que

aporta a aumentar el desempleo y mantener los índices de pobreza (Sánchez, Barboza y Castilla, 2017).

2.2.1.1 Enfoques

Himmel (2002), indica que existen 5 enfoques que explican la deserción estudiantil:

- **Enfoque psicológico**

  Está relacionado a temas psicológicos del alumno que podría desencadenar que no esté adaptado al entorno educativo y llevarlo a la deserción.

- **Enfoque sociológico**

  Se enfatizan en la influencia de los factores externos y la relación del mismo con los demás estudiantes.

- **Enfoque económico**

  Se enfocan en el costo y beneficio que los alumnos podrán abordar durante el periodo de estudios.

- **Enfoque organizacional**

  Se refiere a las distintas características de una institución educativa así como sus servicios brindados.

- **Enfoque de interacciones**

  Se toma en consideración la integración social con otros estudiantes y docentes, así como actividades extracurriculares.

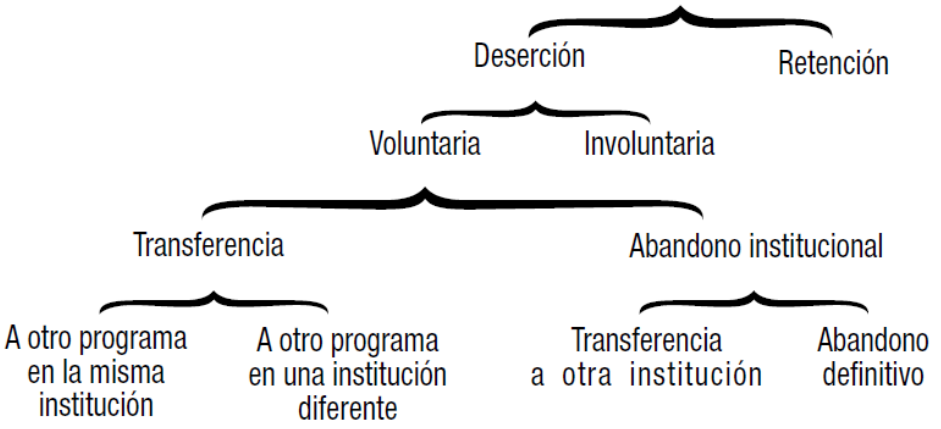


Figura 1. Niveles de los tipos de deserción.
Fuente: (Himmel, 2002). Modelos de análisis de la deserción estudiantil en la educación superior. (p. 95)

2.2.1.2 Clasificación

En cuanto a tiempo, existen 3 tipos de deserción los cuales pueden estar diferenciados por ser en tiempo precoz, temprano y tardío.

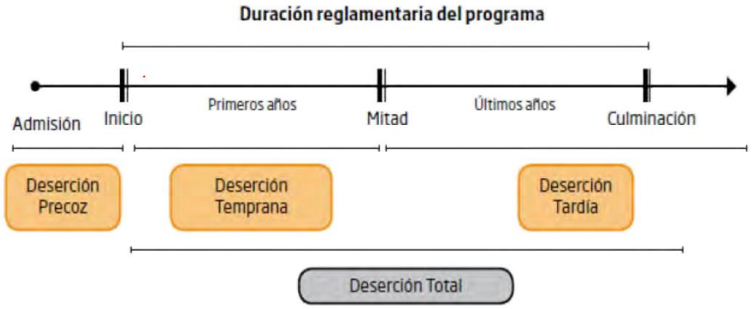


Figura 2. Deserción respecto al tiempo
Fuente: (Villamarín, 2017). Análisis de la deserción estudiantil en la FCECEP utilizando machine learning específicamente mapas auto organizados de kohonen. (p. 46)

En cuanto a espacio, la deserción puede ser cuando un estudiante decide cambiar de carrera se le llama deserción interna, cuando decide cambiar de centro de estudios es la deserción institucional y por último la deserción total.

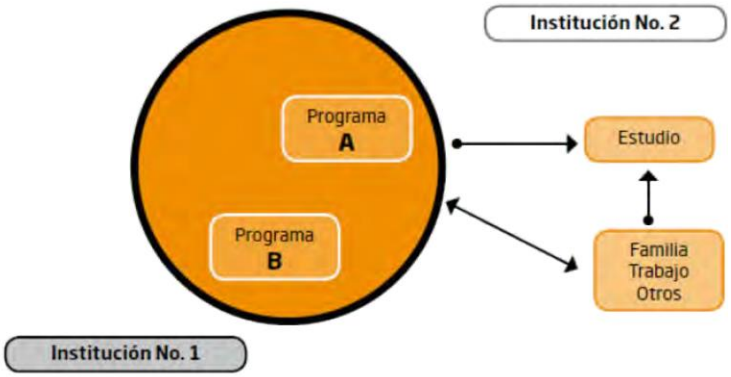


Figura 3. Deserción respecto al espacio
Fuente: (Villamarín, 2017). Análisis de la deserción estudiantil en la FCECEP utilizando machine learning específicamente mapas auto organizados de kohonen. (p. 47)

2.1.1.3 Indicadores y Metodologías

De acuerdo con la investigación de Fischer (2012), se ha podido tener en cuenta los siguientes kpi´s y metodologías para calcular la deserción, primero los indicadores de deserción semestral, que se puede evidenciar cuando un alumno no culmina su ciclo. El de tasa ponderada de deserción por nivel que se calcula de la tasa de deserción con el total de alumnos matriculados.

2.1.1.4 Modelos explicativos

- **Modelo de spady – Teoría del suicidio**

  Durante 1970, Spady elabora estudios relacionados a la deserción basada en los principios de suicidios dictaminados de Durkheim en 1951, donde establece que suicidarse no es solo por factores individuales, ya que existen hechos sociales que generan la ruptura del individuo con la sociedad. Spady indica que la deserción es una consecuencia de un individuo que no ha podido integrarse con su entorno

educativo, además, indica que las características del entorno familiar producen fuertes efectos en el estudiante, el cual pueden modificar el rendimiento académico como la adaptación del alumno.

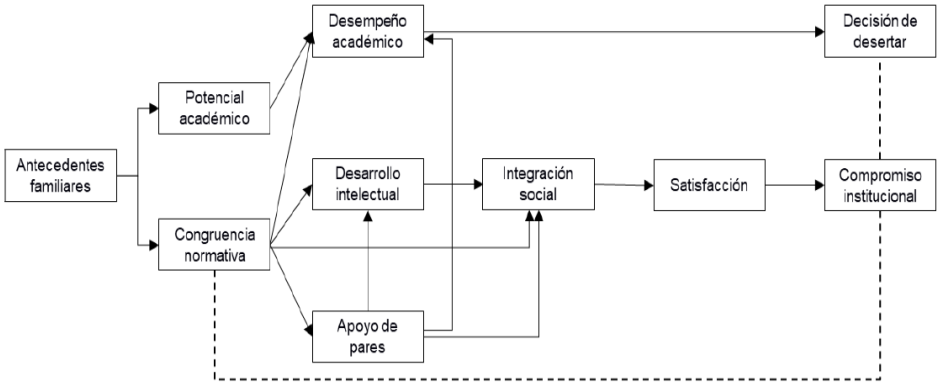


Figura 4. Modelo de Spady
Fuente: (Jurado, 2019). Diseño de un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior. (p. 15)

En la figura 4, Spady muestra un modelo en donde se puede percibir que el entorno familiar tiene un impacto en las decisiones del alumno en cuanto a su vida académica. Las dos variables que se muestra pueden intervenir en el crecimiento profesional y académico del estudiante. Además estos factores pueden afectar en la integración del estudiante a su entorno de educación. Es por ello que se puede concluir que las variables mostradas afectan en las decisiones del estudiante sobre la continuidad de sus estudios.

- **Modelo de tinto – Teoría del intercambio.**

  Tinto menciona que los alumnos continuarían con el programa estudiantil si los beneficios que reciben son superiores a la dedicación y esfuerzo.

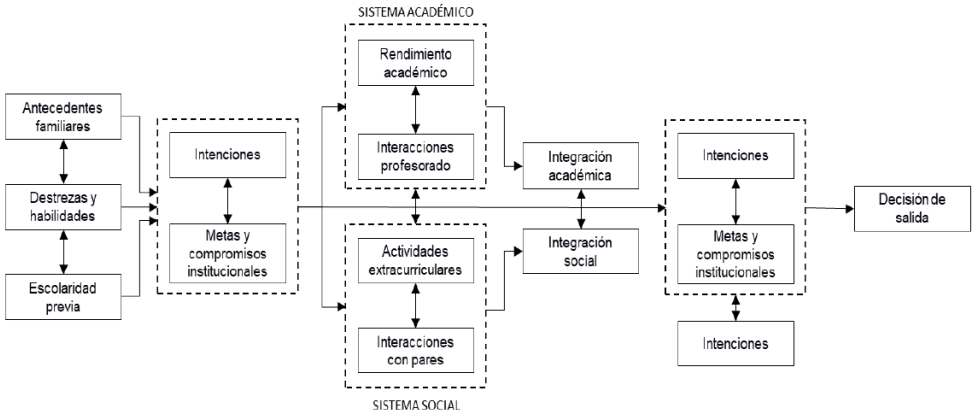


Figura 5. Modelo de Tinto**.**
Fuente: (Jurado, 2019). Diseño de un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior. (p. 16)

De acuerdo con la figura 5, Tinto consideró tres variables posibles como que tienen un gran impacto en los alumnos en sus decisiones de renunciar a sus estudios.

- **Modelo de Bean – Productividad del ambiente laboral**

  En 1985 realiza otra publicación donde amplía su trabajo, en la cual toma en cuenta estudiantes diferenciados como resultado.

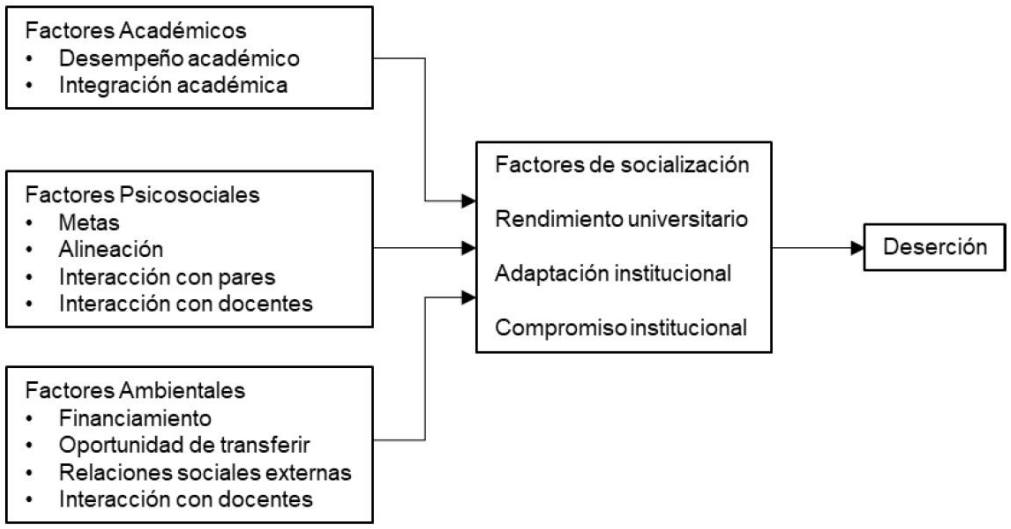


Figura 6. Modelo de Bean
Fuente: (Jurado, 2019). Diseño de un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior. (p. 17)

Según Miranda (2019), la deserción genera costos altísimos a los estudiantes, ya que algunos vienen de familias de pocos recursos y en algunas circunstancias recurren a préstamos de entidades bancarias para solventar los gastos de los estudios. De este modo el costo no impacta en los alumnos solamente sino también al sistema educativo, el cual:

- Genera un congelamiento a la institución del financiamiento.
- Se obtiene la perdida de una vacante.
- Obstruye en la evolución de la educación en el país.

2.2.2 Machine learning

Está definido como un algoritmo que tiene la finalidad de predecir sucesos futuros que son desconocidos para el sistema (Aguilar y Vázquez, 2016).

Según Forero, Piñeros y Rodríguez (2018) es una forma de inteligencia artificial donde se accede e interpreta grandes volúmenes de datos, se entrena el sistema y se predice nueva información a través de algoritmos de aprendizaje, las tareas dictadas al algoritmo y los tipos de modelos

Villamarín (2017) indica que si se desea que una maquina pueda realizar actividades que solo puede realizar el humano, se requiere un algoritmo que pueda incluir el aprendizaje del mismo, tal como es la inteligencia artificial y poder lograr que una maquina pueda distinguir patrones para la toma de decisiones.

Así mismo, Candia (2019) define a machine learning como una manera en que las computadoras pueden aprender por si solas pero con ayuda de los humanos. Y a su vez

este pueda obtener patrones que permitan construir modelos y predecir comportamientos anticipadamente.

De acuerdo con las definiciones mostradas por los autores se puede concluir que machine learning ofrece distintas formas de facilitar acciones predictivas y que se puede utilizar en los casos de deserción estudiantil, ya que por medio de los algoritmos se podrá detectar patrones que ayudaran a determinar si un estudiante llegara a abandonar o no sus estudios. El aprendizaje automático tiene distintas categorías como muestra la figura 7.

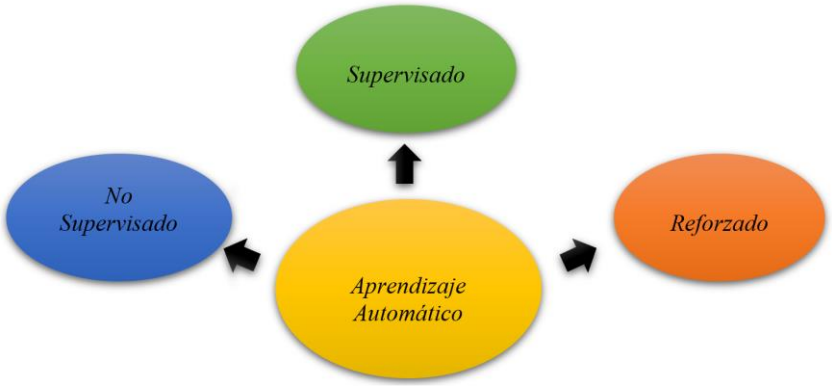


Figura 7. Categorías de Machine Learning
Fuente: (García, 2019). Implementación de un modelo computacional basado en reglas de clasificación supervisadas para la predicción de la deserción estudiantil en la Universidad Peruana Unión Filial Juliaca. (p. 17)

2.2.2.1 Aprendizaje supervisado

Es donde se entra a los algoritmos otorgándoles de preguntas, características y respuestas, denominadas etiquetas con la finalidad de que pueda combinarlas y hacer predicciones (Candia, 2019).

Tiene como objetivo generar modelos predictivos, en las que define el tipo de entrada y salida, mediante el modelo de análisis de datos como la etiquetación y clasificación (Mamani, 2019).

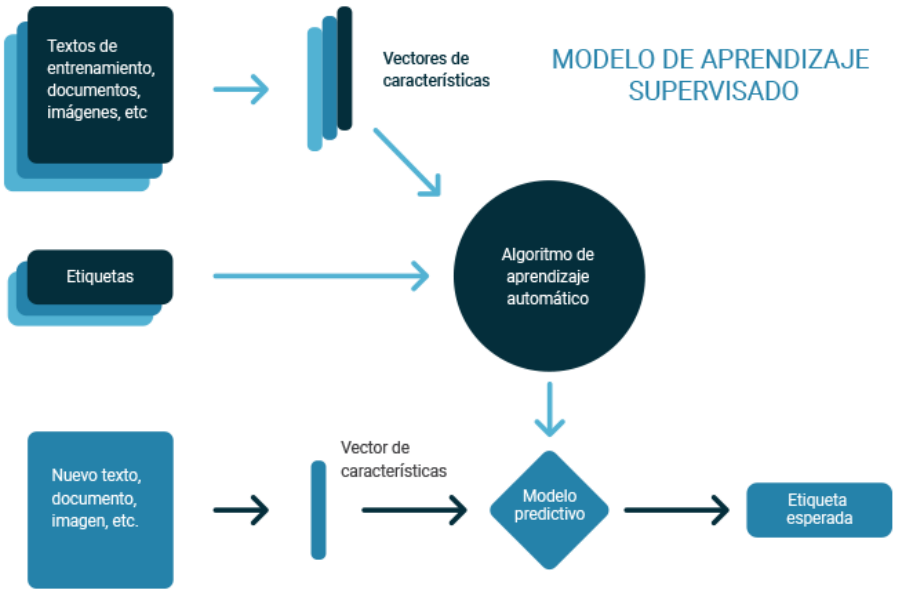


Figura 8. Flujo de Aprendizaje Supervisado
Fuente: (García, 2019). Implementación de un modelo computacional basado en reglas de clasificación supervisadas para la predicción de la deserción estudiantil en la Universidad Peruana Unión Filial Juliaca. (p. 23)

Aprendizaje supervisado tiene como algoritmos:

- Redes neuronales
- Máquinas de soporte vectorial
- Arboles de decisión

Según Candia (2019), existen dos tipos de aprendizaje supervisado:

- **Regresión**: estos algoritmos son utilizados para el aprendizaje automático sobre todo en estadística, por que indica la tendencia de un grupo con datos continuos.

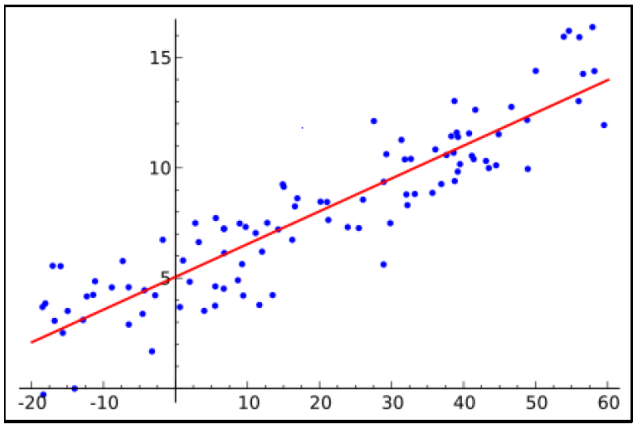


Figura 9. Aprendizaje automático: Regresión
Fuente: (Candia, 2019). Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático. (p. 20)

- **Clasificación**: estos algoritmos buscan o encuentra patrones que les permitirán clasificar objetos .

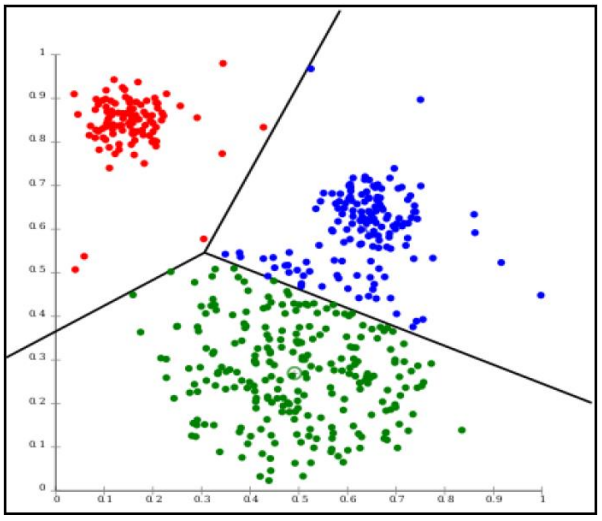


Figura 10. Aprendizaje automático: Clasificación
Fuente: (Candia, 2019). Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático. (p. 21)

2.2.2.2 Aprendizaje no supervisado

Requiere que se le brinde datos de entrada para que pueda almacenar y clasificar los patrones de comportamiento y de esta manera poder emitir un resultado.

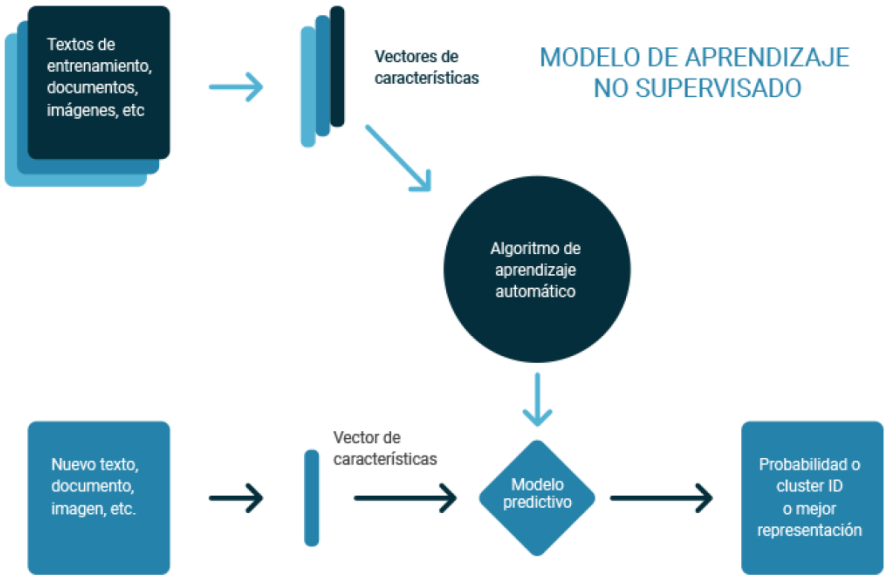


Figura 10. Flujo de Aprendizaje no supervisado
Fuente: (García, 2019). Implementación de un modelo computacional basado en reglas de clasificación supervisadas para la predicción de la deserción estudiantil en la Universidad Peruana Unión Filial Juliaca. (p. 24)

Entre los algoritmos de lenguajes no supervisados, se tienen:

- Mapas auto organizados
- K-medias

2.2.2.3 Aprendizaje reforzado

El aprendizaje reforzado emite un entorno de control para los especialistas en programación y las máquinas para identificar la conducta perfecta dentro de una situación única. Es donde un operador puede elegir la mejor actividad según la condición actual de los resultados obtenidos (García, 2019). El objetivo del aprendizaje reforzado es verificar si tiene la capacidad de adaptarse al entorno.

Figura 11. Flujo de Aprendizaje reforzado.
Fuente: (García, 2019). Implementación de un modelo computacional basado en reglas de clasificación supervisadas para la predicción de la deserción estudiantil en la Universidad Peruana Unión Filial Juliaca. (p. 25)

2.2.2.4 Árbol de decisión

El árbol de decisión se considera como un algoritmo de aprendizaje en la que se utiliza para clasificar y la cual necesita distintos variables de entrada. Esta técnica divida los datos en 2 o más grupos homogéneos diferenciándose en los datos de entrada (Candia, 2019)

Los árboles de decisión son una estructura jerárquica compuesta por un grupo de puntos finales, en la cual cada punto establece una regla o condición con capacidad de devolver verdadero o falso según los valores y atributos que se deseen analizar, de este modo la decisión final puede ser determinada e interpretada. (Chaves y Pertuz, 2016)

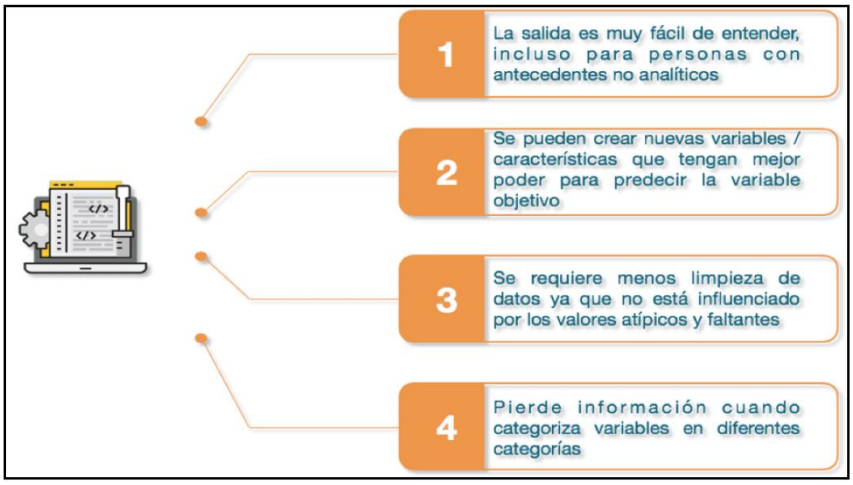


Figura 12. Características de árboles de decisión
Fuente: (Candia, 2019). Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático. (p. 31)

2.2.2.5 Redes neuronales

Las redes neuronales es un algoritmo que tiene un funcionamiento que ha sido basado en el comportamiento del cerebro humano, consta de nodos en el que se transmite las señales de entrada con el objeto de obtener información en su salida. (Carpio, 2016).

Además, según Jurado (2019), el desarrollo de un modelo artificial se guía de un patrón de optimización que no es lineal, compuestos por nodos denominados neuronas, los cuales recepcionan información de diferentes puntos y brindan datos con relevancia.

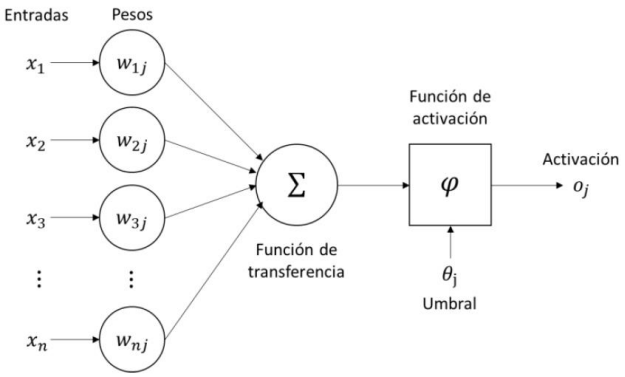


Figura 13. Estructura de una red neuronal.
Fuente: (Jurado, 2019). Diseño de un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior. (p. 29)

2.2.2.6 Máquinas de soporte vectorial (SVM)

Los SVM, son un conjunto de algoritmos que se pueden utilizar para la clasificación y análisis de aprendizaje supervisado. Muchos autores utilizan este tipo de algoritmos para separar grupos por medio de una previa clasificación de datos y además estos, pueden soportar datos altamente dimensionales.

Las máquinas de soporte vectorial aprenden de un conjunto de entrenamientos, luego se aplica a una serie de datos nuevos y busca clasificar de forma óptima mediante la generación de hiperplanos separadores, el cual tiene dos objetivos, minimizar el porcentaje de error del conjunto de entrenamiento y maximizar el margen de separación (Serrano, 2017).

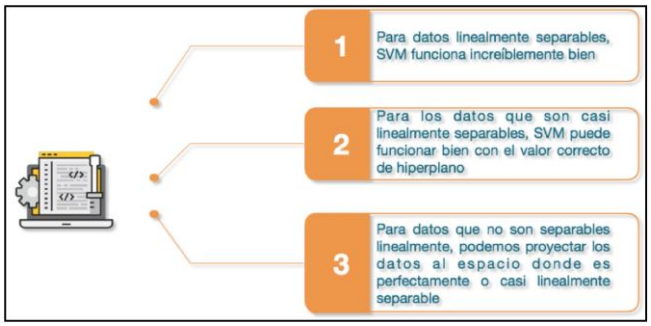


Figura 14. Características de SVM.
Fuente: (Candia, 2019). Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático. (p. 30)

En la Figura 16 se muestra en un hiperplano la separación de 2 grupos clasificados por el SVM, en el cual se clasifica por 2 grupos en el que los del lado derecho son los que no desertaron y en el lado izquierdo los que si desertaran.

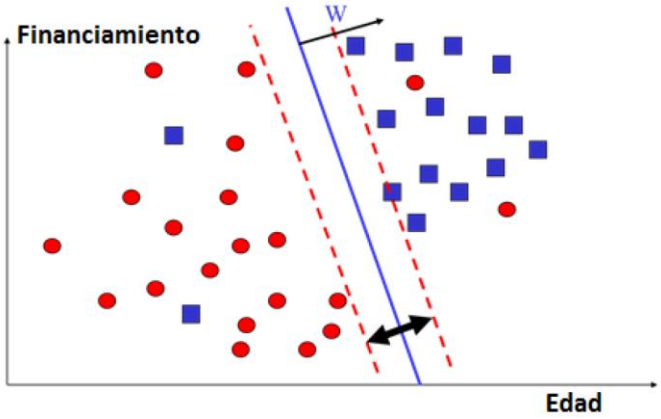


Figura 15. Representación gráfica del algoritmo SVM

Fuente: (Vásquez, 2016). Modelo predictivo para estimar la deserción de estudiantes en una institución de educación superior Universidad privada Antenor Orrego. (p. 34)

Tabla 1. Técnicas de análisis de datos ML.

Nombre	PREDICTIVO		DESCRIPTIVO		
	Clasificación	Regresión	Agrupamiento	Reglas de asociación	Correlaciones/ Factorizaciones
Redes neuronales	•	•	•		
Árbol de decisión	•	•			
Redes de Kohonen			•		
Regresión logística	•			•	
Kmeans			•		
Apriori				•	
Naive Bayes	•				
Vecinos más próximos	•	•	•		
Análisis factorial y de componentes principales					•
Twostep, Cobweb			•		
Algoritmos genéticos y evolutivos	•	•	•	•	•
Maquinas de vectores de soporte	•	•	•		
CN2 rules(cobertura)	•			•	
Análisis discriminante multivariante	•				

Fuente: (Villamarín, 2017)

2.2.3 Minería de datos

Se le considera un grupo de técnicas que tiene como fin el descubrimiento de datos contenidos en grandes grupos de información, en donde se analizan comportamientos, asociaciones, patrones y otras cualidades que se encuentran en los datos (Candia, 2019).

Figura 16. Aplicaciones de Minería de datos
Fuente: (Miranda, 2019). Diseño de un proceso de alertas tempranas para disminuir las deserciones de los estudiantes de primer año en una institución de educación superior. (p. 14)

2.2.4 Metodología SEMMA

SEMMA es la metodología que cuenta con un proceso que clasifica, examina y elabora una gran cantidad de datos con la finalidad de descubrir patrones desconocidos (Fischer, 2012).

Según Chaves y Pertuz (2016), existen 6 fases de SEMMA.

Figura 17. Fases de la metodología SEMMA**.**
Fuente: (Chaves y Pertuz, 2016). Aplicativo web para el análisis, selección y admisión de aspirantes al programa de ingeniería de sistemas de la universidad Cundinamarca en la extensión facatativa utilizando modelos predictivos de minería de datos. (p. 44)

- **Muestreo**

  Es donde se toma muestra de los datos disponibles, el cual debe ser lo suficientemente grande para que contenga información relevante y lo suficiente pequeña para que pueda procesar rápidamente (Chaves y Pertuz, 2016).

- **Exploración**

  En esta etapa es donde se explora los datos en búsqueda de relaciones desconocidos, teniendo como objetivo familiarizarse con los datos y formular nuevas hipótesis (Chaves y Pertuz, 2016).

- **Modificación**

  Es donde se realiza la preparación de datos, limpieza de valores anómalos, tratamiento a datos faltantes y se crea o modifica las variables a trabajar (Chaves y Pertuz, 2016).

- **Modelado**

  Es la creación del modelo a trabajar cuya labor es permitir la predicción de las variables respuesta a partir de las variables explicativas (Chaves y Pertuz, 2016).

- **Evaluación**

  Es donde se evalúa la utilidad y exactitud de los modelos obtenidos.

SEMMA propone generar nuevas hipótesis llevando a repetir el proceso iterativamente.

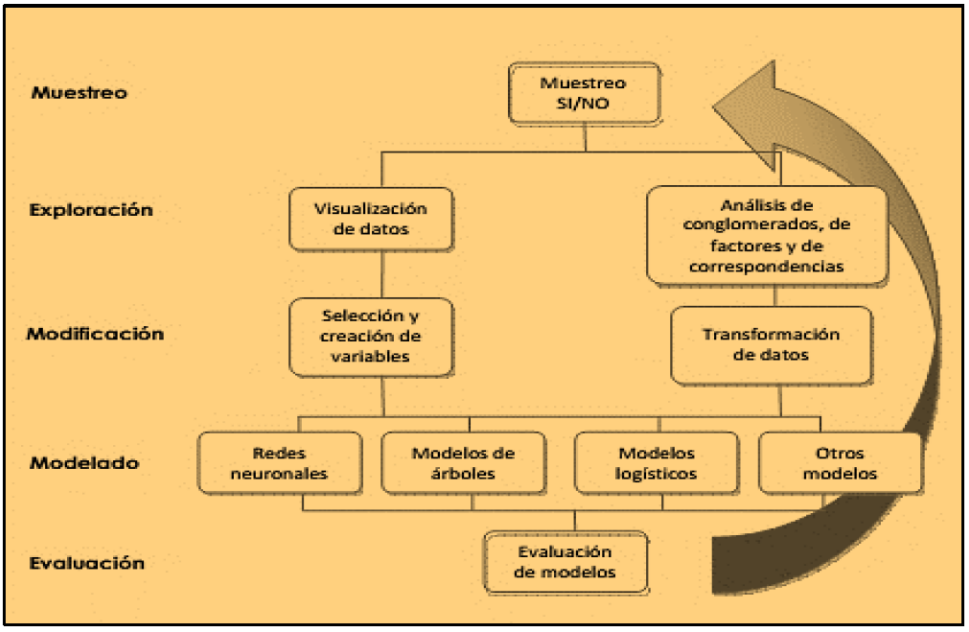


Figura 18. Iteración de las fases metodología SEMMA.
Fuente: (Chaves y Pertuz, 2016). Aplicativo web para el análisis, selección y admisión de aspirantes al programa de ingeniería de sistemas de la universidad Cundinamarca en la extensión facatativa utilizando modelos predictivos de minería de datos. (p. 44)

2.2.5 Metodología KDD

KDD, Se representa por un conjunto de procesos que tiene como objetivo poder recuperar datos, preparar información e interpretar los resultados obtenidos (Jurado, 2019).

Asimismo, esta metodología cuenta con 5 fases:

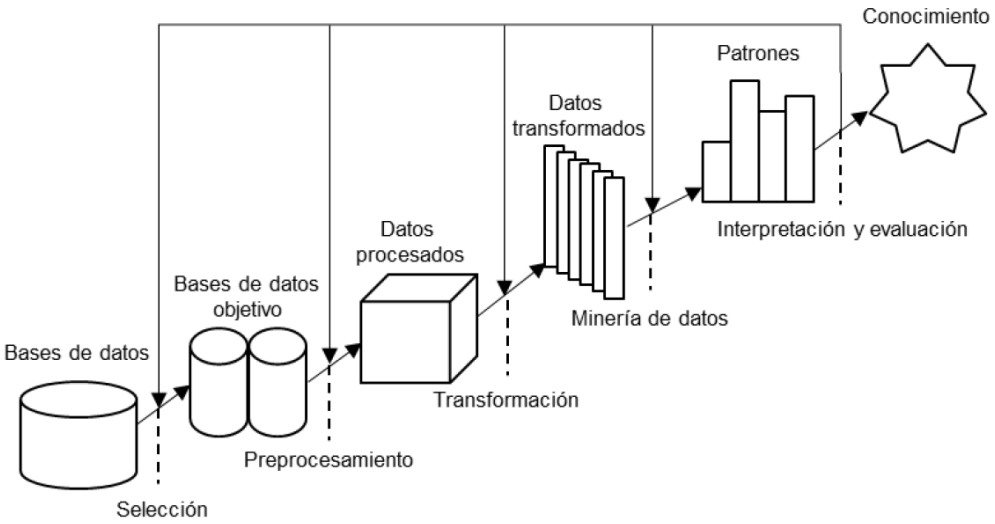


Figura 19. Fases de la metodología KDD

Fuente: (Jurado, 2019). Diseño de un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior. (p. 19)

- **Selección**

  Es la etapa donde se extrae conocimiento, donde se define la fuente de datos que se analizara dependiendo del objetivo del estudio (Chaves y Pertuz, 2016).

- **Pre procesamiento**

  Es la clasificación de los datos y limpieza para su próxima transformación.

- **Transformación**

  Es donde se requiere cambiar los datos procesados a un formato apropiado (Jurado, 2019).

- **Modelización**

  En esta etapa se elige la técnica y algoritmo a utilizar, para poder obtener los patrones dependiendo del conocimiento que se requiere (Chaves y Pertuz, 2016).

- **Interpretación y evaluación**

  Es donde se establece las medidas cuantitativas, para validar si el modelo cumple con los requisitos e identificar los patrones con mayor relevancia (Jurado, 2019).

2.2.6 Metodología CRISP-DM

Fue presentada por las SPSS company, Daimer Chrysler y NCR en el año 1999, como una metodología abierta, construida en base a experiencia de sus creadores, es decir mediante un enfoque práctico. Dentro de ella contienen 6 fases, con sus respectivas tareas y resultados:

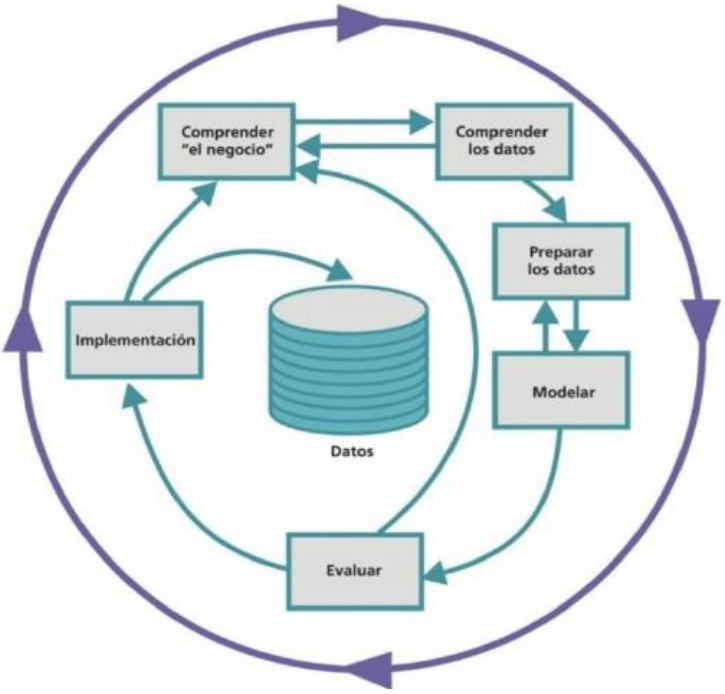


Figura 20. Fases de CRISP-DM.
Fuente: (Miranda, 2019). Diseño de un proceso de alertas tempranas para disminuir las deserciones de los estudiantes de primer año en una institución de educación superior. (p. 11)

- **Fase 1: Comprensión de negocio:**

  Tiene como objetivo comprender el negocio desde su organización hasta las áreas involucradas (De la Cruz, 2016).

- **Fase 2 : Comprensión de datos:**

  Durante la comprensión se analiza la data inicial del estudiante que servirá como datos de aprendizaje para el modelo (Carpio, 2016).

- **Fase 3: Preparación de datos:**

  Aquí, se discriminan los atributos que no son de relevancia para los resultados que se desea obtener (Chaves y Pertuz, 2016).

- **Fase 4 : Fase de modelado:**

  Se decide el modelo a utilizar para que pueda ser evaluado con los datos de ingreso (Fischer, 2012).

- **Fase 5 : Evaluaciones:**

  Es donde se evalúa el modelo, teniendo en cuenta si ha cumplido con los criterios del éxito del problema (Galán, 2015).

- **Fase 6: Despliegue:**

  En esta parte se activa siempre haya modelos que puedan ser implementados, en donde se realiza la documentación, elaboración de plan de funcionamiento y preparación de presentaciones y resultados (Miranda, 2019).

2.2.7 Métricas de rendimiento

Una vez obtenido los resultados con ayuda del algoritmo es necesario averiguar cuan efectivo es el modelo utilizando las principales métricas de rendimiento a fin de evitar el sobreajuste de datos y por lo tanto una mala predicción de nuestro modelo, para ello se empleara la matriz de confusión. (Miranda, 2019).

2.2.7.1 Matriz de confusión

Es una herramienta de métrica para la efectividad de un algoritmo. Para poder emplear esta herramienta es necesario separar nuestra data de entrenamiento 80% train y 20% test esto quiere decir que se podrá comparar el resultado obtenido a partir de estos 2 para obtener una predicción acertada que beneficie al fin de esta investigación.

Tabla 2. Matriz de Confusión

		CLASE	
		+	-
Predicción de Clases	+	PT	PF
	-	NF	NT

Fuente: (Miranda, 2019)

PT: Núm. de positivos definidos como positivos correctos.

NT: Núm. de negativos definidos como negativos correctos.

NF: Núm. de positivos definidos como negativos incorrectos.

PF: Núm. de negativos definidos como positivos incorrectos.

2.2.7.2 Métricas de performance

2.2.7.2.1 Precisión

Se le llama precisión a la proporción que existe en los positivos reales, es decir que es la proporción al total de predicciones que si fueron acertadas correctamente. (Miranda, 2019).

$$\pi = \frac{pt}{pt + pf} \qquad 1$$

En la cual:

PT es el Núm. Verdadero +

PF es el Núm. Falso -

2.2.7.2.2 Sensibilidad

La sensibilidad son valores que pueden identificar la capacidad de limitar los casos positivos de los casos negativos. O también llamado con la habilidad tener sensibilidad con los sucesos positivos. (Miranda, 2019).

$$\rho = \frac{pt}{pt + nf} \qquad 2$$

Donde:

pt = Núm. de verdaderos +

nf = Núm. de falsos -

2.2.7.2.3 Puntaje micro promedio (F1)

Para poder definir el puntaje micro promedio es necesario tener los datos de la precisión y de la sensibilidad definida anteriormente. En este cálculo se busca obtener una puntuación única que represente ambas variables a fin de obtener un ponderado que indicará en porcentaje. (Miranda, 2019).

$$\boldsymbol{F(\text{micro} - \text{averaged}) = \frac{2\pi}{\pi + \rho}} \qquad 3$$

Donde:

π = Precisión

ρ = Sensibilidad

Cabe mencionar que el puntaje micro promedio, es considerado el promedio de las categorías.

2.3 CONTEXTO DE LA INVESTIGACIÓN

2.3.1 Deserción estudiantil en universidades peruanas

Como definen muchos autores la deserción estudiantil es un problema mundial que es necesario enfrentar y en muchos de estas investigaciones dan énfasis a la tecnología como herramienta para minimizar las deserciones estudiantiles, previamente identificando los patrones que son síntomas de un desertor y de esta manera con la estrategia de los centros educativos, intervenir y persuadir al estudiante.

Los autores de esta investigación abordan predecir la deserción con una investigación cualitativa utilizando las herramientas de machine learning para lograr disminuir la tasa de deserción. Actualmente se tiene una estadística que aproximadamente el 30% de alumnos deja sus estudios en el Perú.

2.3.2 Machine learning como herramienta para la predicción

El aprendizaje automático es bastante usado para predecir situaciones con patrones de comportamiento, en muchas investigaciones existen el uso de diversas tecnologías para predecir como la minería de datos, que en este caso los autores de esta investigación la usaran como fuente de información.

2.3.3 Universidad Tecnológica del Perú (UTP)

Este trabajo va dirigido a la UTP sede centro, de la ciudad de Lima que inicio con sus operaciones desde el año 1997. Para el análisis y el desarrollo del diseño de machine

learning, los autores se centraron en el campus de Lima centro en donde realizaran la toma de información y análisis de los alumnos que desertaron desde 2017 hasta el 2019 obteniendo los datos base para entrenar al modelo de machine learning.

CAPITULO III: METODOLOGIA DE LA INVESTIGACIÓN

3.1 DISEÑO DE LA INVESTIGACIÓN

3.1.1 Tipo de investigación.

Este trabajo hará uso del algoritmo SVM, Para ello se necesitará utilizar la metodología cuantitativa ya que permitirá revisar la información recuperada a partir de los algoritmos y de esta manera cuantificar la información, con el objetivo de resolver las preguntas de investigación de este proyecto.

3.1.2 Instrumentos y técnicas de investigación.

Se requiere obtener una data histórica de los alumnos que desertaron, para ello es necesario realizar un análisis documental que recopile la siguiente información:

- Obtención de la data de alumnos desertores que cuenta la Universidad Tecnológica del Perú
- Los atributos considerados de los estudiantes que dejaron sus estudios.

3.1.3 Población y muestra

3.1.3.1 Población

Se consideró a los alumnos que desertaron dejando registro del motivo del abandono de sus carreras.

3.1.3.2 Muestra

Se consideró que la muestra se debe realizar con alumnos que desertaron y dejaron información histórica de las razones de sus retiros, desde el año 2017 al 2019. Además se está tomando como muestra, a los alumnos con las siguientes características:

- El total de alumnos de la sede central: 55% de alumnos son varones, y el 45% son mujeres.
- Cualquier nivel socioeconómico.
- Cualquier edad.
- Sede de lima centro de la UTP.

3.1.3.3 Trabajo de campo

Se recopilo información de los alumnos en el que se evidencia que tienen en cuenta muchos atributos al momento de almacenar las razones posibles de deserción del alumno, para ello cabe mencionar que existen distintos factores de deserción. Además, en el trabajo de campo se realizó un levantamiento de información en conjunto con la gerencia de retención de la UTP.

A continuación se mostrarán las entidades y los atributos que se toma en cuenta para almacenar en los registros históricos la información de los alumnos que desertaron durante el año 2017 hasta el 2019

Tabla 3. Tabla de entidades y atributos de histórico de alumnos desertores

ENTIDAD	ATRIBUTOS	OBSERVACIÓN
Alumno	Id_Alumno Nombres Apellidos DNI Fecha_nacimiento Año Ingreso Sexo Estado Civil Sector Económico Becas Lugar de nacimiento	La tabla alumno almacena los datos de los estudiantes.
Carrera	IdCarrera Descripción	Esta entidad está relacionada a la tabla alumnos en el que especifica la carrera que está cursando el alumno
Notas	Id_Notas Notas Promedio Ponderado	La tabla notas tiene relación con la tabla ciclos, en el que se almacena las notas y promedio del alumno
Cursos	Id_Cursos Descripción	Esta tabla almacena los cursos llevados por el alumno durante su carrera
Ciclo	Id_Ciclo Descripción	Tabla ciclo que está relacionada al alumno en el que indica su ciclo actual y los restantes
Empleo	Id_Empleo Empresa Hora ingreso Hora Salida Dirección	Esta tabla muestra si el alumno se encuentra laborando actualmente y sus horarios como referencia.
Modalidad	Id_Modalidad Descripcion	Esta tabla, está relacionada a la carrera y muestra la modalidad ya sea pregrado o para gente que trabaja
País	Id_Pais Descripcion	Esta tabla muestra los datos del país origen del estudiante
Ciudad	Id_Ciudad Descripcion	Esta tabla muestra los datos de la ciudad donde reside el alumno
Distrito	Id_Distrito Descripcion	Esta tabla muestra el distrito donde reside el alumno

Apoderado	Id_Apoderado Nombre Apellidos Dni Parentesco	Esta tabla muestra los datos del apoderado del alumno de la Universidad Tecnológica del Perú.
Asistencia	Id_Asistencia Fecha y hora	Esta tabla, nuestra el registro de las asistencias a clases del alumno durante sus ciclos
Estudios previos	Id_Estudios_Previos Nombre_centro_estudios Observacion Ciclo	Esta tabla, indica los estudios previos que realizó el estudiante, ya sea en otra universidad o instituto.
Deudas	Id_Deudas Descripcion Id_Pagos Record Fecha	Esta tabla muestra las deudas acumuladas del alumno durante su vida académica
Pagos	Id_Pagos Matricula Pension Fecha	Esta tabla almacena los pagos realizados por el alumno y las fechas de cada aporte.
Deserción	Id_Desercion Id_alumno Motivo Fecha_desercion	Esta tabla, almacena los datos de los alumnos que desertaron así como el motivo que ellos indican del abandono de sus estudios.
Estrategias_Deserción	Id_Estrategias_Desercion Descripcion_estrategia	Esta tabla muestra las posibles estrategias que se podría emplear por la gerencia de retención.

En el cuadro anterior se puede notar, que existe relación entre las entidades Deserción y “estrategias_desercion” en el que llevan el registro de los motivos del alumno y las posibles estrategias que se pueden. Asimismo se tiene los datos completos de sus cursos, notas, pagos, si cuenta con empleo o no, además del lugar donde reside. Dichos datos son relevantes para la investigación debido a que pueden brindar los datos históricos y mediante machine learning y el algoritmo de Support Vector Machine para predecir si un

estudiante que actualmente este cursando alguna carrera, este con posibles intenciones de abandonar sus estudios.

3.2 METODOLOGIA DE LA IMPLEMENTACIÓN

Para diseñar un sistema que permita predecir la deserción de estudiantes mediante machine learning, existen muchas metodologías que muchos autores recomiendan, para la elección de la metodología es necesario tener referencias y por ello en la siguiente figura se puede observar una comparación de las metodologías más usadas.

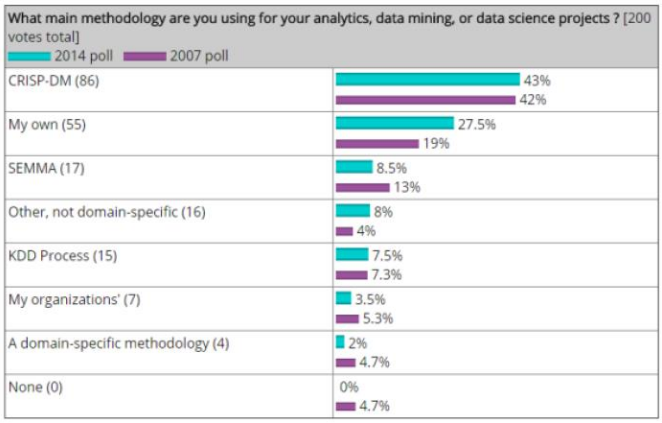


Figura 21. *Metodologías más usadas*
Fuente: (García, 2019). Implementación de un modelo computacional basado en reglas de clasificación supervisadas para la predicción de la deserción estudiantil en la Universidad Peruana Unión Filial Juliaca. (p.35)

Como muestra la figura 28, la metodología CRISP-DM y sus 6 pasos permiten desde la comprensión del negocio hasta la explotación y evaluación del modelo es la metodología CRISP-DM, y con fundamento en la base teórica mostrada en este trabajo de investigación, esta metodología es también usada para predecir comportamientos mediante machine learning.

En la siguiente tabla se podrá observar las diferencias que tiene las metodologías con mayor uso en exportación de datos.

Tabla 4. Tabla de comparación entre metodologías de minería de datos.

CRITERIOS/METODOLOGIAS	CRISP-DM	SEMMA	KDD
Metodología Estructurada	SI.	SI.	SI.
Metodología Independiente	SI.	NO.	SI.
Ampliamente Usada	SI.	NO.	NO.
Mejora la calidad de resultados en proyectos de Data Mining.	SI.	SI.	SI.
Herramientas y técnicas independientes.	SI.	SI.	SI.
Finalidad diversa (Ej. Ampliamente estable en la resolución de problemas. variados).	SI.	SI.	SI.
Fácil de Implementar	SI.	SI.	SI.

Fuente: (García, 2019)

Así mismo a continuación se muestra una tabla con criterios y puntuación brindada para seleccionar que metodología tiene mejor utilización para esta investigación.

Tabla 2. Tabla de selección de metodología por criterios

	Amplia información en internet	Amplia documentacion	Grupos en internet	Uso de empresas	Proyectos de investigación	TOTAL
Proceso KDD	3	2	4	2	4	15
SEMMA	2	4	4	3	3	16
CRISP-DM	4	5	5	4	5	23
KDD	2	3	2	3	4	14
TOTAL	11	14	15	12	16	68

Como se puede observar la metodología CRISP DM es bastante demandada en cuanto a predicción y machine learning esto debido a sus 6 fases con las que se puede entender el negocio, comprender los datos, preparar los datos, modelar los datos, evaluar los datos e implementarlos.

Es por ello que los autores de esta investigación han visto conveniente la utilización de la metodología CRISP DM para desarrollar una investigación basado en predicción.

3.2.1 FASES DE LA METODOLOGÍA

3.2.1.1 Comprensión de negocio.

Según la metodología CRISP-DM es importante comprender el negocio, es decir involucrarse con las áreas de la UTP para poder analizarla alineado con el objetivo de este proyecto.

Para lograr esto se tiene que tener en cuenta las siguientes tareas iniciales a realizar.

Tarea 1:

- Identifique las áreas claves que pueden almacenar información del alumnado de la sede central de la UTP.

Tarea 2:

- Describa el problema en forma general.

Tarea 3:

- Describa las soluciones que la universidad emplea actualmente para evitar la deserción.

3.2.1.2 Comprensión de datos

Es necesario obtener la data general del alumno para analizar los datos obtenida y modelarlo de acuerdo a los requerimientos del proyecto. Por lo que se accederá a los

registros y base de datos de la UTP a fin de obtener la cantidad mayor posible de datos de alumnos que se retiraron durante el año 2017 al 2019.

Para lo cual se tiene que tener en cuenta las siguientes tareas:

- De los datos obtenidos, se debe realizar una copia del mismo discriminando los datos que no son relevantes tales como, Nombre de los padres, Lugar de nacimiento, Teléfono.
- Considerar los datos de los alumnos que desertaron en el complejo central de la UTP, así como limitarlos desde los años 2017 al 2019.

3.2.1.3 Preparación de los datos

Se debe seleccionar la muestra obtenida del área encargada de gestionar la información de los alumnos que desertaron ya a partir de ella clasificar los datos, eliminando valores que puedan tener datos vacíos así como también agregar atributos que permita dicha clasificación sin modificar los datos reales de los alumnos desertores.

3.2.1.3.1 Selección de los datos

Una de las sub fases de la preparación de datos es la selección para ello se tiene que tener en cuenta las siguientes actividades. En el cual debería identificar que datos son importantes para el modelo, es decir que debe hacerse una discriminación de los datos relevantes en esta investigación.

Además, se tendrá que realizar la selección de elementos en la cual se considerar a los alumnos que desertaron entre los años 2017 al 2019 solo para la sede central, para luego

poder seleccionar los atributos que serán de vital importancia para el modelo ya que serán los datos de entrada para ejecutar el aprendizaje automático.

3.2.1.3.2 Limpieza de los datos

Para tener una data de calidad para el análisis del sistema de machine learning es necesario realizar un filtro de datos.

- **Data perdida**. Es posible que los alumnos que no han completado el cuestionario se tengan que omitir de los modelos posteriores. Lo que se puede hacer es modelar las diferencias de los motivos de deserción entre los alumnos que responden y los que no responden el cuestionario. Si estos dos conjuntos de alumnos tienen motivos de deserción similares, los cuestionarios que faltan son menos preocupantes.
- **Errores de datos.** Los errores detectados durante el proceso de exploración se pueden corregir en esta fase. La mayoría de las veces, sin embargo, la entrada correcta de datos se debe realizar desde la plataforma donde se registra a un alumno que desertó.
- **Errores de mediciones.** Al igual que los cuestionarios perdidos, se trata de un problema difícil, porque es posible que no se disponga del tiempo o recursos disponibles para recopilar la información del alumno desertor.

3.2.1.3.3 Elaboración de nuevos datos

El investigador debe realizar un proceso de exploración para comprobar que la creación de los datos se ha realizado correctamente.

3.2.1.4 Modelado

Para este caso, interesa encontrar la relación de las calificaciones que se han obtenido en las primeras evaluaciones con la deserción. Así mismo, se tomará en cuenta las notas como las variables independientes y si deserta o no como la variable dependiente.

3.2.1.4.1 Seleccionar las técnicas de modelado

Los autores de este trabajo consideraron el algoritmo Support Machine Vector (SVM) debido a la gran ventaja que tiene para la predicción de patrones de comportamiento. Además cuenta con un alto porcentaje de aciertos en comparación con muchos otros modelos.

3.2.1.4.2 Elaboración de un diseño de comprobación

Se denomina al uso del 75% de los datos como primera iteración y la segunda iteración se refiere al 25% de los datos que son restantes, después de haber sido entrenado el modelo.

3.2.1.5 Evaluación

Es donde se debe evaluar los datos obtenidos en relación con la precisión y efectividad de los resultados del modelo. Así mismo verificar si se pueden aplicar a los objetivos del proyecto.

3.2.1.6 Explotación

Es la fase en el cual se utiliza los nuevos conocimientos obtenidos para poder explicar al cliente sobre los datos obtenidos como resultado, es decir se le presenta la documentación para que pueda intervenir en la decisión del alumno anticipadamente con el fin de evitar su deserción.

- Crear una planificación para el despliegue e integración del modelo con sus sistemas. Registre los detalles técnicos como los requisitos de base de datos para los resultados del modelo.
- Para cada descubrimiento se debe crear un plan para difundir la información a los estrategas de la organización.
- Identificar los problemas de despliegue y realice un plan de contingencia.

Se debe elaborar el informe final para poder resolver los cabos sueltos de las documentaciones previas y comunicar detalladamente los resultados obtenidos.

- Descripción con detalles de la situación problemática
- Brindar el conocimiento adquirido mediante los resultados.
- Emitir la documentación completa.

CAPITULO IV: DESARROLLO DE LA SOLUCIÓN

Este capítulo tiene como objetivo implementar la solución, identificando las fases de la metodología elegida para la elaboración de este proyecto y realizándolas a partir de datos brindados por la Universidad Tecnológica del Perú, para después ser probado con las métricas del aprendizaje automático.

4.1 PROPUESTA DE SOLUCIÓN

4.1.1 Comprensión de negocio.

Para la comprensión del negocio, es conveniente realizar el entregable 1 de obtener detalles de la organización de la UTP, y analizar cada área que involucra a los alumnos de dicha universidad, con el fin de mapear los datos relevantes de los estudiantes y tener una base sólida con un margen de error menor que consumirá el algoritmo que se encargará de predecir el comportamiento y concluir si es un alumno desertor. Como parte de la actividad del entregable 1 es identificar si existe un área que se encargue de gestionar la deserción y en dicha información general muestra que el área que se encarga de retener alumnos es la GERENCIA DE RETENCIÓN, Para revisar la información general de la Universidad Tecnológica del Perú. Ver ANEXO II.

Como parte del entregable 2, se requiere describir el problema el cual trata de la deserción de estudiantes de diferentes ciclos y para ello los autores de esta investigación han visto a bien delimitar dicha investigación usando como muestra a los alumnos de la

sede central de la ciudad Lima, así como también delimitar los años de los estudiantes que desertaron desde el 2017 hasta el 2019.

Mientras que el entregable 3, actualmente la UTP gestiona y da como solución disuadir a los alumnos realizando un acercamiento con ellos cuando los alumnos empiezan a faltar a clases o a no cumplir con las pensiones, es por ello que la universidad cuenta con muchos programas como socioeconómicas, de salud mental que tienen como objetivo guiar al alumno en su vida académica a fin de contribuir en gran porcentaje a la continuidad del alumno con sus estudios superiores.

La desventaja de este tipo de solución es que es reactiva, es decir que cuando el alumno ya se encuentra en la decisión potencial de desertar, la universidad intenta persuadir y muchas veces no es efectiva.

4.1.2 Comprensión de datos.

Se pretende obtener la data de los alumnos de la UTP, que será brindada por el área de gerencia de retención.

- Verificar la información de los alumnos de la UTP. (ANEXO 3)

Los datos obtenidos en dicha recolección para una mejor comprensión es importante separarlos por categorías como muestra a continuación la siguiente imagen:

- Atributos Demográfico del estudiante:

Apellidos y nombres	Sexo	Carrera profesional
Día de nacimiento	Teléfono	Ubigeo Nacimiento
Sociedad civil	DNI	Nivel socioeconómico
Dirección de residencia	Datos de Apoderado	Dirección laboral

- Atributos con relación a la Universidad

Código Estudiante	Asistencia del alumno	Modalidad de estudios
Carrera Elegida	Becas	Matriculas
Notas	Pagos	Promedio ponderados

- Atributos de gestión de desertores

Código Estudiante	Estrategia de retención
Motivo de deserción	Estudios previos
Fecha de deserción	Deudas del alumno

Con la obtención de los atributos ya clasificados se tienen los datos comprendidos y se necesitaría preparar los datos.

4.1.3 Preparación de los datos.

Después de la clasificación de los atributos se seleccionará los atributos relevantes para la investigación los cuales se convertirán en las variables independientes y 1 variable independiente que se refiere a la variable de salida también conocido como TARGET, que es el que indicará si un alumno esta con mucho potencial de desertar o no. A continuación se detalla las variables.

Tabla 3. Tabla de variables para el modelo de predicción

Número	Código	Descripción	Valor
1	NOTA_A	Notas Alumno	Número Entero +
2	SEXO_A	Sexo Alumno	0-1
3	ECIVIL_A	Estado Civil	1-4
4	PROMP_A	Promedio Ponderados	Número Real +
5	DEUDASA_A	Record de deudas	Número entero +
6	ASISTENCIA_A	Record Asistencia	Número entero +
7	CICLO_A	Ciclo del alumno	Número entero +
8	UBI_TRAB	Ubigeo de trabajo	Número entero +

9	ING_TRAB	hora de ingres del trabajo	Número entero +
10	SAL_TRAB	hora de salida del trabajo	Número entero +
11	UBI_RES	Ubigeo de Residencia	Número entero +
12	NSOCIOECO_A	Nivel Socio-Económico	1-3
13	BECA_A	Becas Inscritas	0-1
14	MOT_DESER	Factor de deserción	1-3
VARIABLES DE SALIDA			
15	TARGET	Indica si el estudiante es desertor o no	0 – 1

Para la comprensión de los valores de las variables mostradas en el cuadro anterior, se detalla las siguientes tablas.

Tabla 4. Tabla de género del alumno

GENERO	
Variable	Valor
Masculino	1
Femenino	0

Nota: Esta tabla muestra el valor de los tipos de género.

Tabla 5. Tabla de estado civil del alumno

ESTADO CIVIL	
Variable	Valor
Casado	1
Soltero	2
Divorciado	3
Viudo	4

Nota: Esta tabla muestra el valor de los tipos de estado civil.

Tabla 6. Tabla del nivel socioeconómico

NIVEL SOCIOECONOMICO	
Variable	Valor
Alta	1
Media	2
Baja	3

Nota: Esta tabla muestra el valor de los niveles socioeconómicos.

Tabla 7. Tabla de factor de deserción.

FACTOR DE DESERCIÓN	
Variable	Valor
Factor Social	1
Factor Ambiental	2
Factor Económico	3

Nota: Esta tabla muestra el valor de los factores de deserción.

4.1.4 Modelado.

En esta fase, se hace la comparación y elección del modelo de acuerdo a muchas otras investigaciones con la especialidad de identificar patrones de comportamiento. Es por ello que los autores de esta investigación, en base a distintos trabajos de investigación con el objetivo de predecir comportamientos, se concluyó que la técnica de clasificación con mejor beneficio es el algoritmo Support Vector Machine. Los detalles de este algoritmo están en el capítulo II, en el punto 2.2.2.6, página 30.

El objetivo que tiene esta etapa es de entrenar a los clasificadores SVM con los datos de entrenamiento que son proporcionados por la Universidad desde las variables independientes ya establecidas. El modelo SVM separa los datos para clasificar en desertores o no.

4.1.5 Evaluación.

4.1.5.1 Obtención de la data.

Para la evaluación del modelo se ha tomado en cuenta solo los alumnos que estén estudiando en la UTP sede Centro.

Para este proyecto de investigación se está analizando con 1533 alumnos mostrando como desertores a 46 y a 1487 como alumnos no desertores. La obtención de estos datos se detalla en los siguientes recuadros.

Tabla 8 Cuadro de desertores y no desertores.

	Cantidad	%
Alumnos Desertores	46	3%
Alumnos No desertores	1487	98%
Total	1533	100%

De los datos obtenidos se utiliza un porcentaje del 67% del total de los datos aplicadas en el entrenamiento con un total de 1027 alumnos de ellos han sido desertores 30 alumnos.

Tabla 9. Tabla de desertores y no desertores con el total de datos de entrenamiento

	Cantidad	%
Alumnos Desertores	30	3%
Alumnos No desertores	997	98%
Total	1027	100%

De tal modo que el 33% de validación tiene un total de 506 alumnos, de los cuales el modelo clasificó como desertores a 15 alumnos como desertores y a 491 alumnos como no desertores.

Tabla 10. Tabla de ratios de desertores y no desertores con los datos de validación

	Cantidad	%
Alumnos Desertores	15	3%
Alumnos No desertores	491	98%

Total	506	100%

4.1.5.2 Pruebas

- **Entrenamiento**

En esta fase se obtuvo un 99.9% acertados. Se calculó el dato de cada modelo SVM con el método de pesos:

Porcentaje de Acierto de cada SVM = W x (1 – E) x (100%)

Donde:

- W es el peso de los modelos.
- E el error.

Además, para conseguir el % total de acierto, se suma cada uno de los porcentajes parciales de cada modelo SVM.

$$\%\ total\ de\ acierto = \sum W\ x\ (1 - E)\ x\ (100\%)$$

Como se observa, existe un porcentaje bajo en cuanto a error.

- **Validaciones**

En esta etapa se obtuvo 99.75% que se mostrará cómo se realizó el cálculo.

$$\%\ de\ acierto = \frac{DC+NDC}{DC+NDC+DI+\ NDI}\ x\ 100\ \%$$

Donde:

- DC es el número de alumnos desertores clasificados de forma correcta.

- NDC es el número de alumnos no desertores clasificados de forma correcta.
- DI es el número de alumnos desertores clasificados incorrectamente.
- NDI es el número de alumnos no desertores clasificados con error.

Tabla 11. Matriz de confusión de la etapa de validación

Clasificación	Desertores	No desertores
Alumnos Desertores	13	2
Alumnos No desertores	6	485

Para la medición de los resultados, se ha utilizado la matriz de confusión obteniendo los resultados.

$$\%\ de\ acierto = \frac{13 + 485}{13 + 485 + 2 + 6}\ x\ 100\ \%$$

$$\%\ de\ acierto = 98.41\ \%$$

La matriz de confusión de la tabla 11, indica un resultado aceptable ya que brinda la confiabilidad por el alto porcentaje de precisión de predicción de alumnos desertores.

4.1.6 Despliegue.

En esta etapa, ya habiendo mapeado los datos brindados se procede a la documentación con los reportes de los alumnos vigentes evaluados, y se pone a disposición de la UTP la información obtenida de dicho estudio, para que el área de Gerencia de retención direccione sus esfuerzos disuasorios a los alumnos que tienen más

riesgo de desertar de acuerdo al modelo de predicción brindado por machine learning y de esta manera disminuir la cantidad de deserciones de los estudiantes de la UTP

4.2 Prototipos

En la figura 22, se visualiza la interfaz principal del sistema en la que el usuario podrá ingresar colocando su respectivo usuario y contraseña de acceso.

Figura 22. Interfaz principal del sistema de predicción de deserción estudiantil.

Fuente: Propia.

Una vez el usuario haya logrado ingresar correctamente, aparecerá la pantalla de bienvenida con el dashboard principal en la que mostrará a los alumnos con probabilidad de deserción del último periodo en la figura 23.

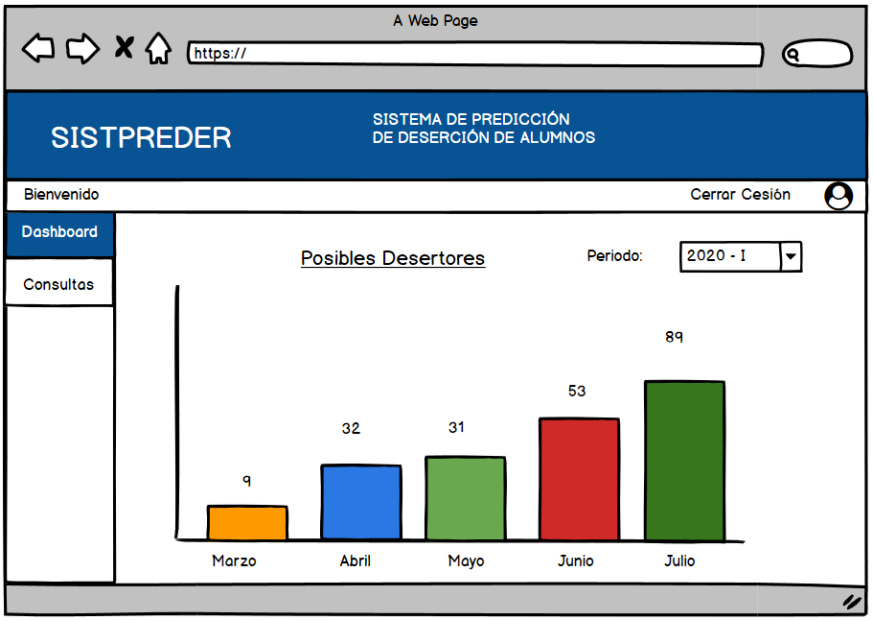


Figura 23. Interface de bienvenida con el dashboard.
Fuente: Propia

En la figura 24, mostrara la opción de consultas, en la cual se podrá observar el listado de los alumnos que el sistema ha podido detectar como desertor o no desertor.

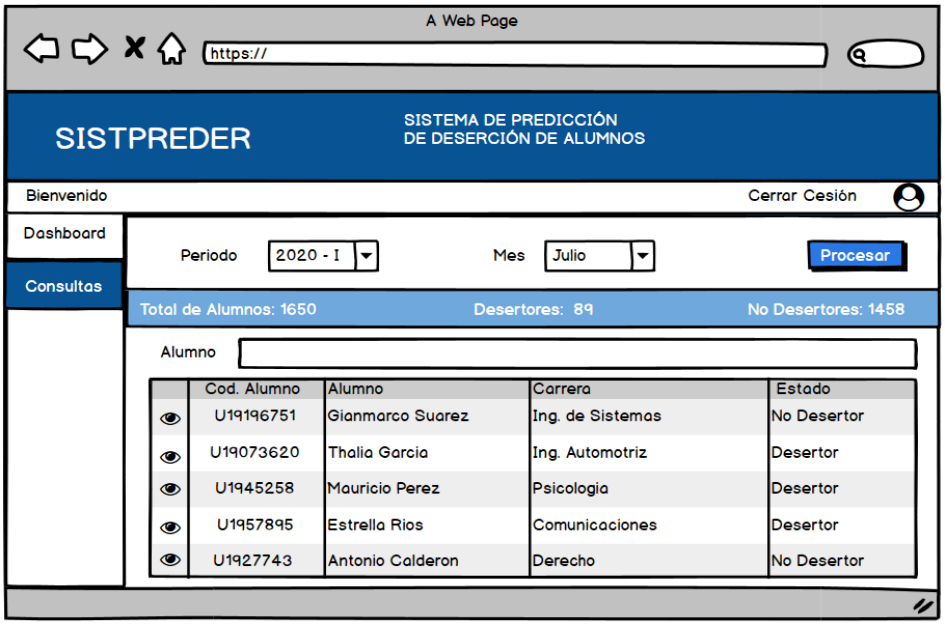


Figura 24. Interfaz de consultas del sistema

Fuente: Propia.

Para poder visualizar el detalle de un alumno se tendrá que seleccionar al estudiante que desea visualizar y te dirigirá a la pantalla donde en la que se mostrara los datos con mayor importancia del alumno seleccionado.

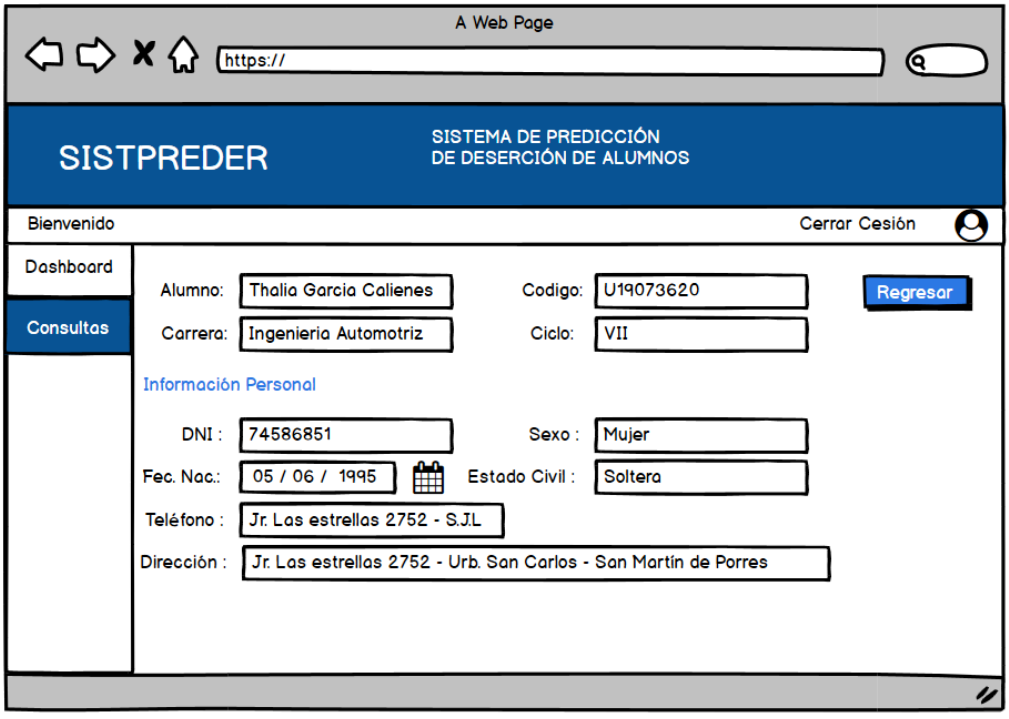


Figura 25. Interface de detalles de alumnos
Fuente: Propia.

CAPITULO V: CONCLUSION Y RECOMENDACION

5.1 Conclusión.

Durante el desarrollo de este trabajo de investigación se evaluó el problema de la deserción de estudiantes de la UTP. Por tanto se identificó por medio de la data del alumnado los que desertaron durante los años 2017 al 2019 con el fin de recolectar la información que posteriormente con el modelo Support Vector Machine (SVM) encontrar patrones de comportamiento que puedan ayudar a predecir los alumnos que son potencialmente desertores de sus estudios de educación superior. Se utilizó los hábitos y perfiles de dichos estudiantes para encontrar coincidencias con los alumnos actuales, tal como se mostró es posible detectar a los alumnos que piensan desertar mediante machine learning. Así mismo se usó la metodología CRISP-DM que permite dividir por fases el diseño del sistema de predicción.

Por ello se concluye que el algoritmo utilizado SVM identifica los factores con mayor influencia de deserción estudiantil debido a que utiliza como base de aprendizaje la información de alumnos que ya desertaron y permite predecir los patrones de comportamiento y por lo tanto el factor con mayor influencia.

Además la implementación del SVM interviene fuertemente en la predicción de los alumnos de la UTP permitiendo que los encargados del área de Gerencia de retención puedan intervenir para disuadir la decisión del alumnado.

Por último, se obtuvo un grado de certeza superior a 90% superando el grado base que se trazó en la hipótesis de esta investigación.

5.2 Recomendaciones

En el presente trabajo se empleó el algoritmo de Support Vector Machine (SVM), que es muy utilizado en documentos de investigación en cuanto a predicción de comportamiento. Además en este proyecto se utilizó la función Kernel debido a que permite que el SVM se vuelva no lineal teniendo una distribución parecida a la distribución Normal.

Para futuras investigaciones se recomienda usar la función Kernel RBF que optimiza los parámetros de entrada y de esta manera se puede obtener un mejor resultado con una población más grande.

REFERENCIAS

Aguilar, L. y Vásquez, Y. (2016). *Principal component analysis (PCA) para mejorar la performance de aprendizaje de los algoritmos support vector machine (SVM) y red neuronal multicapa (MLN)*.Tesis de título profesional. Universidad privada Antenor Orrego, Trujillo Perú. Recuperado de http://repositorio.upao.edu.pe/bitstream/upaorep/3398/1/RE.SIS_LUIS.AGUILAR_YNDRA.VASQUEZ_PRINCIPAL.COMPONENTE_DATOS.PDF

Candia, D. (2019). *Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático*. Tesis de maestría, Universidad nacional san Antonio de abad, Perú. Recuperado de http://repositorio.unsaac.edu.pe/bitstream/handle/UNSAAC/4120/253T20191024_TC.pdf?sequence=1&isAllowed=y

Carpio, J. (2016). *Modelo de predicción para la morosidad en el otorgamiento de crédito financiero aplicando metodología CRISP-DM*. Tesis para obtener el grado de ingeniero de sistemas, Universidad Andina Néstor Cáceres Velázquez, Juliaca, Perú. Recuperado de http://repositorio.uancv.edu.pe/bitstream/handle/UANCV/743/TESIS.pdf?sequence=3&isAllowed=y

Chaves, L y Pertuz C. (2016). *Aplicativo web para el análisis, selección y admisión de aspirantes al programa de ingeniería de sistemas de la universidad Cundinamarca en la extensión facatativa utilizando modelos predictivos de minería de datos*. Recuperado de http://repositorio.ucundinamarca.edu.co/bitstream/handle/20.500.12558/2102/Tesis%20de%20Grado.pdf?sequence=3&isAllowed=y

De la cruz, V. (2019). *Diseño de un modelo predictivo basado en machine learning para el control de la deserción de estudiantes den la Universidad Ricardo Palma*. Tesis para título profesional, Lima, Perú. Recuperado de http://190.12.70.20/bitstream/UNTELS/489/1/De_la_Cruz_Victor_Trabajo_Suficiencia_2019.pdf

Fischer, S. (2012), *Modelo para la automatización del proceso de determinación de riesgos de deserción en estudiantes universitarios*. Tesis de Maestría. Universidad de Chile, Santiago de Chile. Recuperado de http://repositorio.uchile.cl/bitstream/handle/2250/111188/cf-fischer_ea.pdf?sequence=1&isAllowed=y

Forero L. , Pineros, Y. & Rodríguez J. (2019). *Machine learning for the identification of students at risk of academic Universidad Federico Caldas*. Articulo Colombia. Recuperado de http://repository.udistrital.edu.co/bitstream/11349/15890/1/ForeroZeaLeidyDaniela2019.pdf

Galán, V. (2015). *Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en un entorno universitario*. Recuperado de https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

García, J. (2019). *Implementación de un modelo computacional basado en reglas de clasificación supervisadas para la predicción de la deserción estudiantil en la Universidad Peruana Unión Filial Juliaca*. Recuperado de https://repositorio.upeu.edu.pe/bitstream/handle/UPEU/1975/Jacob_Tesis_Licenciatura_2019.pdf?sequence=1&isAllowed=y

Himmel, E. (2002). *Modelos de análisis de la deserción estudiantil en la educación superior. Calidad de la Educación,* (17), 91-108. Recuperado de https://www.calidadenlaeducacion.cl/index.php/rce/article/view/409/409

Jurado, M. (2019). *Diseño de un modelo predictivo de la deserción estudiantil de postgrado en una institución de educación superior*. Recuperado de https://www.dspace.espol.edu.ec/retrieve/133935/D-CD110079.pdf

Mamani, D. (2019). *Modelo de minería de datos basado en factores asociados para la predicción de deserción estudiantil universitaria*. Recuperado de http://repositorio.unam.edu.pe/bitstream/handle/UNAM/94/T095_72389106_T.pdf?sequence=1&isAllowed=y

Miranda, F. (2019). *Diseño de un proceso de alertas tempranas para disminuir las deserciones de los estudiantes de primer año en una institución de educación superior*. Tesis de maestría, universidad de Chile, Chile. Recuperado de http://repositorio.uchile.cl/bitstream/handle/2250/172649/Dise%C3%B1o-de-un-proceso-de-alertas-tempranas-para-disminuir-las-deserciones.pdf?sequence=1

Ordoñez, R. y Pastor, M. (2016). *Sistema de predicción de clientes desertores de tarjetas de crédito para la banca peruana usando support vector machine*. Tesis de título, Universidad Nacional Mayor de San Marcos, Perú. Recuperado de http://cybertesis.unmsm.edu.pe/handle/cybertesis/4931

Russo, C. (2019), *Minería de datos aplicada a estrategias para minimizar el rezago academic y la deserción universitaria en carreras de informática de la UNNOBA*. Recuperado de http://sedici.unlp.edu.ar/bitstream/handle/10915/79958/Documento_completo.pdf-PDFA1b.pdf?sequence=1&isAllowed=y

Sánchez, G., Barboza, M., y Castilla, H. (2017). *Análisis de la deserción y los factores asociados a la permanencia estudiantil en una universidad peruana*, (69), 169-191. Recuperado de https://ciencia.lasalle.edu.co/ap/vol1/iss69/6/

Vásquez J., (2016). *Modelo predictivo para estimar la deserción de estudiantes en una institución de educación superior Universidad privada Antenor Orrego*.

Villamarín, J. (2017). *Análisis de la deserción estudiantil en la FCECEP utilizando machine learning específicamente mapas auto organizados de kohonen*. Recuperado de http://red.uao.edu.co/bitstream/10614/9618/1/Tc07288.pdf

ANEXOS

ANEXO 1

FICHA DE TAREA INVESTIGACIÓN

FACULTAD DE INGENIERIA

CARRERA DE INGENIERIA DE SISTEMAS EINFORMATICA

1. **Título del trabajo de la tarea de investigación propuesta**

Título: DISEÑO DE UN SISTEMA PARA PREDECIR LA DESERCIÓN DE LOS ALUMNOS MEDIANTE MACHINE LEARNING EN LA UNIVERSIDAD TECNOLÓGICA DEL PERÚ

1.1 Competencia de carrera (Pág. Web UTP.; en Pregrado elegir Carrera; ir a Malla de Carrera, buscar las competencias alineadas con el título)

PARA DESARROLLAR LA INVESTIGACIÓN PROPUESTA, LOS ALUMNOS DEBEN CONTAR CON LAS SIGUIENTES COMPETENCIAS:

- ✓ Desarrollo de aplicaciones móvil
- ✓ Diseño y desarrollo de juegos interactivos
- ✓ Programación de interfaces y dispositivos periféricos
- ✓ Sistemas distribuidos
- ✓ Interacción hombre – maquina
- ✓ Análisis y diseño de sistemas de información
- ✓ Desarrollo de software 1 y 2

2. **Indique el número de alumnos posibles a participar en este trabajo**. (máximo 2)

   Número de Alumnos: DOS (2)

3. **Indique si el trabajo tiene perspectivas de continuidad después que el alumno obtenga el Grado Académico para la titulación por la modalidad de tesis o no.**
   LOS PARTICIPANTES EN ESTA INVESTIGACIÓN PUEDEN CONTINUAR CON ESTE TEMA PARA OBTENER EL TÍTULO DE INGENIERO EN SISTEMAS O INGENIERO DE SOFTWARE, PARA LO CUAL SE DEBE INNOVAR DE ACUERDO CON LAS NUEVAS TENDENCIAS Y TECNOLOGÍAS DISRUPTIVAS.

4. **Enuncie 3 o 5 palabras claves que le permitan al alumno realizar la búsqueda de información para el Trabajo en Revistas Indizadas en WOS, SCOPUS, EBSCO, SciELO, etc desde el comienzo del curso y otras fuentes especializadas.**
   Ejemplo:

Palabras Claves	REPOSITORIO 1	REPOSITORIO 2	REPOSITORIO 3
1. Machine learning	EBSCO	SCOPUS	SCHOLAR- GOOGLE
2. Aprendizaje automático	EBSCO	SCOPUS	SCHOLAR- GOOGLE
3. Tecnologías disruptivas	EBSCO	SCOPUS	SCHOLAR- GOOGLE

1. Entrenamiento constante	EBSCO	SCOPUS	SCHOLAR- GOOGLE
2. Lógica artificial	EBSCO	SCOPUS	SCHOLAR- GOOGLE

1. **Como futuro asesor de investigación para titulación colocar:**

   a. **Nombre** : **EFRAIN LIÑAN SALINAS**
   b. **Código Docente** : **C15136**
   c. **Correo institucional** : **C15136@UTP.EDU.PE**
   d. **Teléfono** : **975079743**

2. **Especifique si el Trabajo de investigación:**
   a. **Contribuye a un trabajo de investigación de una Maestría o un doctorado de algún profesor de la UTP,**
   b. **si está dirigido a resolver algún problema o necesidad propia de la organización,**
   c. **si forma parte de un contrato de servicio a terceros,**
   d. **corresponde a otro tipo de necesidad o causa (Explicar cuál)**

3. **Explique de forma clara y comprensible al alumno los objetivos o propósitos del trabajo de investigación.**
   EL OBJETIVO DEL PRESENTE PROYECTO IMPLEMENTAR MACHINE LEARNING PARA PREDECIR LA DESERCIÓN DE ESTUDIANTES DE LA UNIVERSIDAD TECNOLOGICA DEL PERU.

4. **Brinde al alumno una primera estructuración de las acciones específicas que debe realizar para que le permita al alumno iniciar organizadamente su trabajo.**
   EL ALUMNO DEBERÁ INVESTIGAR SOBRE ESTUDIOS SIMILARES AL PROYECTO, EL CUAL PERMITIRÁ SUSTENTAR LOS ANTECEDES Y EL MARCO TEÓRICO, PARA ELLO DEBERÁ UTILIZAR REPOSITORIOS ESPECIALIZADOS E INVESTIGACIONES EN DIVERSAS UNIVERSIDADES TANTO NACIONALES E INTERNACIONALES

5. **Incorpore todas las observaciones y recomendaciones que considere de utilidad al alumno y a los profesores del curso para poder desarrollar con éxito todas las actividades.**
   LOS DOCENTES Y LOS ALUMNOS ANTES DE INICIAR CON ESTE PROYECTO DEBERÁN INVESTIGAR SOBRE EL ESTADO DEL ARTE PERTINENTE AL TEMA PROPUESTO, ADEMÁS LOS ASESORES DEBEN ESTAR ENSEÑANDO CURSOS RELACIONADOS AL TEMA; ASIMISMO LOS ALUMNOS DEBEN HABER LLEVADO CURSOS RELACIONADOS AL DESARROLLO DE APLICATIVOS Y METODOLOGÍAS DE DESARROLLO DE SOFTWARE.
6. **Fecha y docente que propone la tarea de investigación**
   Fecha de elaboración de ficha: 25/07/2019

7. **Esta Ficha de Tarea de Investigación ha sido aprobada por:**
   (Sólo para ser llenada por la Dirección Académica)
   Nombre: _____
   Código: _____

ANEXO 2

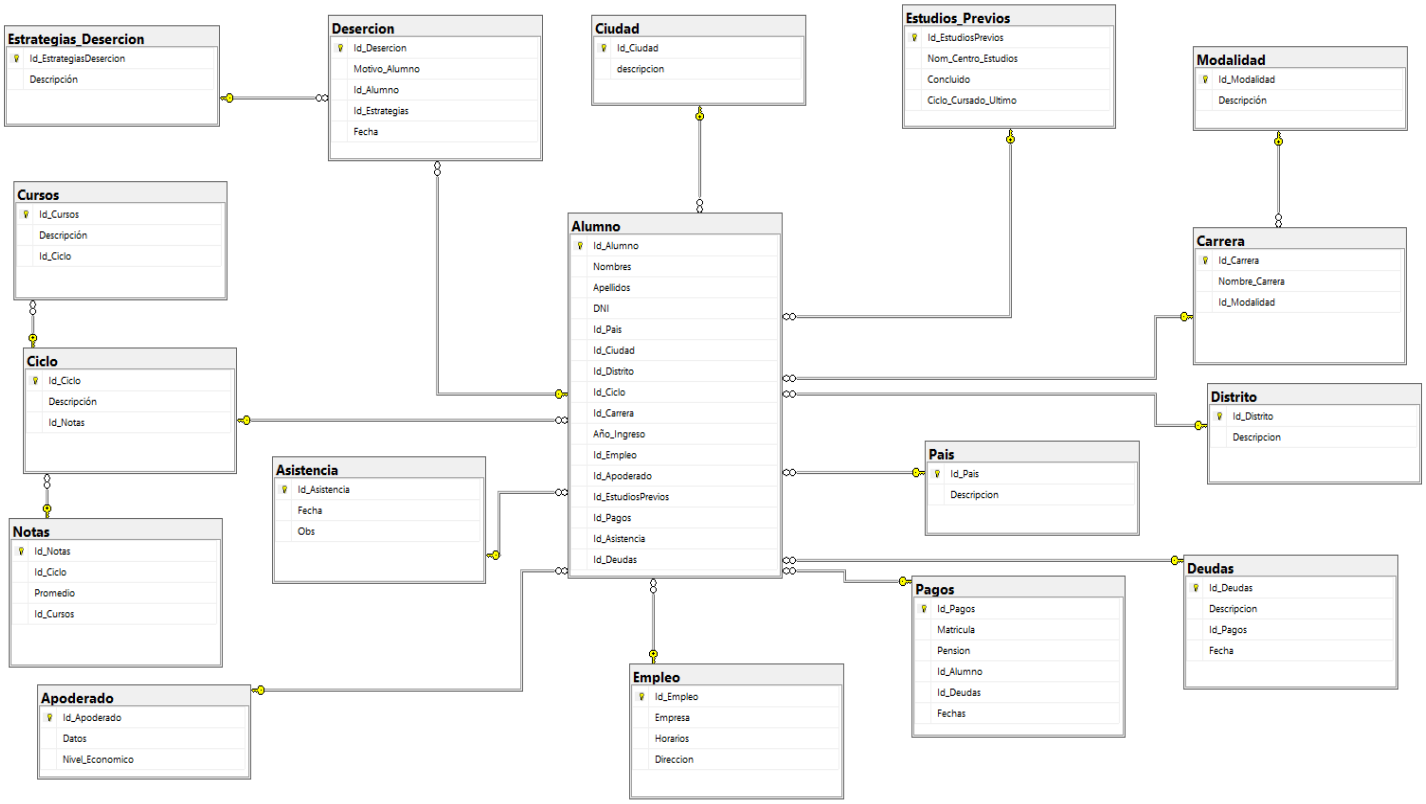
MATRIZ DE CONSISTENCIA

Problema **General**	**Objetivo** **General**	**Hipótesis** **General/ planteamiento de la propuesta**	**Variables** **de estudio**	**Instrumentos**
¿Cómo predecir la deserción estudiantil en la Universidad Tecnológica del Perú mediante el uso de Machine learning?	Diseñar un sistema de predicción de deserción de alumnos de la universidad tecnológica del Perú, mediante machine learning	El sistema de predicción basado en machine learning contribuye a identificar a los alumnos en riesgo de deserción de la Universidad Tecnológica del Perú.	Deserción estudiantil	Histórico de estudiantes desertores
Problemas específicos	**Objetivos Específicos**	**Hipótesis**		
¿Cuáles son los factores que alteran la permanencia de los estudiantes en la Universidad Tecnológica del Perú? ¿Cómo un sistema basado en machine learning puede predecir la deserción estudiantil en la Universidad Tecnológica del Perú? ¿Cómo garantizar la fiabilidad del sistema de predicción de deserción de estudiantes de la Universidad Tecnológica del Perú?	Identificar los factores de deserción de los estudiantes en la Universidad Tecnológica del Perú. Determinar la estructura óptima para la predicción de deserción estudiantil basado en SVM Realizar la prueba del modelo óptimo para la deserción de estudiantes de la Universidad Tecnológica del Perú.	El algoritmo utilizado podrá identificar los factores con mayor influencia en la deserción estudiantil. La implementación de la técnica Support vector machine podrá influir significativamente en la predicción de la deserción estudiantil. El diseño del sistema de predicción de deserción estudiantil tiene un grado de certeza superior al 80%.	Machine learning	Machine support Vector (SVM)

ANEXO 2: Link del reglamento general de la UTP.

https://www.utp.edu.pe/sites/default/files/reglamento_general_3.pdf

ANEXO 3: Base de datos de Alumnos de la UTP



ANEXO 4

Declaración de Autenticidad y No Plagio (Grado Académico de Bachiller)

Por el presente documento, yo Carlos Andres Perez Bedia, identificado/a Con DNI N° 46038187, egresado de la carrera de Ingeniería de sistemas e informática, informo que he elaborado el Trabajo de Investigación denominado "Diseño de un sistema para predecir la deserción de los alumnos mediante Machine learning en la Universidad Tecnológica del Perú".

Para optar por el Grado Académico de Bachiller en la carrera de Ingeniería de sistemas e informática, declaro que este trabajo ha sido desarrollado íntegramente por el/los autor/es que lo suscribe/n y afirmo que no existe plagio de ninguna naturaleza. Así mismo, dejo constancia de que las citas de otros autores han sido debidamente identificadas en el trabajo, por lo que no se ha asumido como propias las ideas vertidas por terceros, ya sea de fuentes encontradas en medios escritos como en Internet.

Así mismo, afirmo que soy responsable solidario de todo su contenido y asumo, como autor, las consecuencias ante cualquier falta, error u omisión de referencias en el documento. Sé que este compromiso de autenticidad y no plagio puede tener connotaciones éticas y legales. Por ello, en caso de incumplimiento de esta declaración, me someto a lo dispuesto en las normas académicas que dictamine la Universidad Tecnológica del Perú y a lo estipulado en el Reglamento de SUNEDU.

LIMA, 30 de JULIO de 2020.

……

(firma)

ANEXO 5

Declaración de Autenticidad y No Plagio (Grado Académico de Bachiller)

Por el presente documento, yo Luis Enrique Rojas Segovia, identificado/a Con DNI N° 71582685, egresado de la carrera de Ingeniería de sistemas e informática, informo que he elaborado el Trabajo de Investigación denominado "Diseño de un sistema para predecir la deserción de los alumnos mediante Machine learning en la Universidad Tecnológica del Perú".

Para optar por el Grado Académico de Bachiller en la carrera de Ingeniería de sistemas e informática, declaro que este trabajo ha sido desarrollado íntegramente por el/los autor/es que lo suscribe/n y afirmo que no existe plagio de ninguna naturaleza. Así mismo, dejo constancia de que las citas de otros autores han sido debidamente identificadas en el trabajo, por lo que no se ha asumido como propias las ideas vertidas por terceros, ya sea de fuentes encontradas en medios escritos como en Internet.

Así mismo, afirmo que soy responsable solidario de todo su contenido y asumo, como autor, las consecuencias ante cualquier falta, error u omisión de referencias en el documento. Sé que este compromiso de autenticidad y no plagio puede tener connotaciones éticas y legales. Por ello, en caso de incumplimiento de esta declaración, me someto a lo dispuesto en las normas académicas que dictamine la Universidad Tecnológica del Perú y a lo estipulado en el Reglamento de SUNEDU.

LIMA, 30 de JULIO de 2020.

....

(Firma)

ANEXO 6

Formulario de Autorización de Publicación

en el Repositorio Académico de la UTP

En calidad de autor(es) del trabajo titulado:

"Diseño de un sistema para predecir la deserción de los alumnos mediante Machine learning en la Universidad Tecnológica del Perú",

Para obtener:

[X] Grado Académico de Bachiller [ ] Título profesional

Carrera: Ingeniería de sistemas e informática

Manifiesto que nuestra obra es original y que en su producción no hemos usurpado derechos de autor o de terceros, siendo el material de nuestra exclusiva autoría. Por lo tanto, el/el autor (es) de este trabajo que a continuación nos presentamos:

Datos personales (llenar un cuadro por cada autor)

Nombres y apellidos: Carlos Andres Perez Bedia	
Código: 1220317	
Correo: 1220317@utp.edu.pe	Teléfono/ celular: 986633199

Nombres y apellidos: Luis Enrique Rojas Segovia	
Código: 1627749	
Correo: 1627749@utp.edu.pe	Teléfono/ celular: 992385745

Decidimos:

[X] Autorizar la publicación en forma inmediata.
[ ] No autorizar la publicación (especificar motivo)

A la Universidad Tecnológica del Perú para colocarlo en su Repositorio

Institucional y sea así de libre acceso/consulta.

En el caso de No autorizar su publicación, existe un periodo de embargo a los 2 años de manera automática.

Es por eso que, mediante la presente dejamos constancia de que lo que estamos entregando a la Universidad es la versión final y aprobada por el jurado.

Fecha: 30 / 07 / 2020

Carlos Andres Perez Bedia	
Nombres y apellidos	Firma
Luis Enrique Rojas Segovia	
Nombres y apellidos	Firma

ANEXO 7

RESULTADOS DE TURNITIN