TESINA DE GRADO

LICENCIATURA EN SISTEMAS DE INFORMACIÓN

Universidad Autónoma de Entre Ríos
Facultad de Ciencia y Tecnología

Año 2025

“Minería de datos aplicada para la identificación de patrones de deserción universitaria”

Autores: A.S. Arrejin, Martin Raúl
A.S. Richard, Rodrigo Adrián

Director de Tesina: Bioing. Hadad, Alejandro
Co Director de Tesina: Esp. Lic. Suiva, Tamara
Colaborador: Mg. Lic. Díaz, Santiago Raúl

Revisores: Ing. Vera, Monica María Cristina
Lic. Cabrera Julio Ignacio
Lic. Pamela Bonadeo

Índice de contenido

Índice de imágenes

Índice de tablas

Agradecimientos

En primer lugar, deseamos expresar nuestro reconocimiento a la Facultad de Ciencia y Tecnología de la Universidad Autónoma de Entre Ríos (UADER), cuyos recursos académicos, humanos y materiales han sido fundamentales para la realización de este trabajo.

Asimismo, agradecemos profundamente el acompañamiento del director, Bioing. Alejandro Hadad; de la codirectora, Esp. Lic. Tamara Suiva; y del colaborador, Mg.Lic. Santiago Raúl Díaz, cuyas valiosas sugerencias, compromiso y guía constante enriquecieron significativamente el desarrollo de esta tesina.

Finalmente, un agradecimiento especial a nuestras familias y amistades, cuyo aliento, comprensión y apoyo incondicional fueron esenciales a lo largo de todo este proceso. Gracias por estar presentes en cada etapa, brindándonos la fuerza necesaria para superar los desafíos que implicó este camino académico.

A.S. Arrejin, Martín

A.S. Richard, Rodrigo Adrián

Resumen

La minería de datos se ha convertido en una herramienta esencial para el análisis de grandes volúmenes de información, permitiendo identificar patrones y relaciones significativas en diversos ámbitos. En el contexto educativo, su aplicación ofrece nuevas oportunidades para abordar problemas como la deserción en la universidad.

El objetivo es identificar patrones asociados al abandono de los estudios y desarrollar modelos predictivos que permitan detectar a los estudiantes con mayor riesgo de deserción. Este estudio se centra en los alumnos de las carreras Licenciatura en Sistemas de Información y Análisis de Sistemas pertenecientes a la Facultad de Ciencia y Tecnología de la UADER. El mismo se desarrolla a partir del análisis de datos académicos y socioeconómicos aplicando técnicas de minería de datos.

A partir de estos hallazgos, se brindará a la institución una herramienta valiosa para implementar estrategias de permanencia e intervenciones que permitan la continuidad académica.

Palabras claves: minería de datos, deserción universitaria, patrones, predicción, educación superior.

Introducción

La minería de datos ha evolucionado como una disciplina fundamental en el análisis de información a gran escala, proporcionando herramientas para descubrir patrones ocultos y generar modelos predictivos con aplicaciones en diversos sectores. Aunque el término *data mining*[1], se popularizó en la década de 1990, sus fundamentos provienen de la estadística, la inteligencia artificial (IA) y el *machine learning* (ML). Los avances tecnológicos en procesamiento y almacenamiento han permitido su expansión, facilitando el análisis automatizado de datos y mejorando la toma de decisiones.

En la actualidad, los procesos de minería de datos se encuentran estrechamente vinculados con la ciencia de datos (*data science)* y el *big data*. Según IBM (2021), la primera se concibe como una disciplina que combina las matemáticas y la estadística, la programación especializada, el análisis avanzado, la IA y el ML con conocimientos específicos del dominio. Por su parte, el *big data* hace referencia a conjuntos de datos masivos y complejos que los sistemas tradicionales de gestión no pueden manejar, y cuyo análisis adecuado permite extraer información valiosa para la toma de decisiones.Las definiciones de los principales conceptos utilizados en esta sección: patrones, modelos, minería de datos, inteligencia artificial, aprendizaje automático, programación especializada, análisis avanzado, ciencia de datos y *big data* se presentan en el (Anexo A).

En el ámbito educativo, la deserción universitaria representa un desafío persistente en América Latina y el Caribe, con tasas alarmantes que afectan tanto a los estudiantes como a las instituciones. Según el Banco Mundial (2017), aproximadamente el 50 % de los jóvenes que ingresan a la universidad no culmina sus estudios, lo que refleja una problemática multifactorial.

Este trabajo se enfocará en la deserción en las carreras de Licenciatura en sistemas de información y Análisis de sistemas de la Facultad de Ciencia y Tecnología de la Universidad Autónoma de Entre Ríos (UADER) en el periodo comprendido entre 2013 y 2023. Ambas comparten una base curricular común, lo que genera similitudes en los desafíos académicos que enfrentan los estudiantes, aunque también existen diferencias en los factores que influyen en la permanencia o abandono.

La aplicación de técnicas de minería de datos permitirá desarrollar modelos predictivos para identificar a los estudiantes con mayor riesgo de deserción. Esto podría brindar a la institución una herramienta clave para diseñar estrategias de retención más efectivas y aplicar intervenciones tempranas que favorezcan la continuidad académica.

[1] Data mining hace referencia a la minería de datos.

Objetivos

Objetivos generales

- Aplicar técnicas de minería de datos para identificar y analizar los patrones y factores que influyen en la deserción universitaria en la Facultad de Ciencia y Tecnología de la UADER.

Objetivos específicos

- Investigar las distintas técnicas de minería de datos aplicadas a la educación.
- Analizar la problemática de la deserción universitaria en la facultad.
- Recopilar y procesar datos de los registros académicos de la facultad.
- Aplicar técnicas de minería de datos para identificar patrones y factores asociados a la deserción.
- Interpretar los resultados y proponer estrategias para reducir la deserción universitaria.

Capítulo 1: Minería de datos

Se presenta el concepto de minería de datos, su utilidad en el análisis de grandes volúmenes de información y su aplicación en distintos contextos. Se explican las etapas del proceso de minería de datos, desde la selección hasta la evaluación de los resultados, así como las principales técnicas utilizadas para descubrir patrones y realizar predicciones.

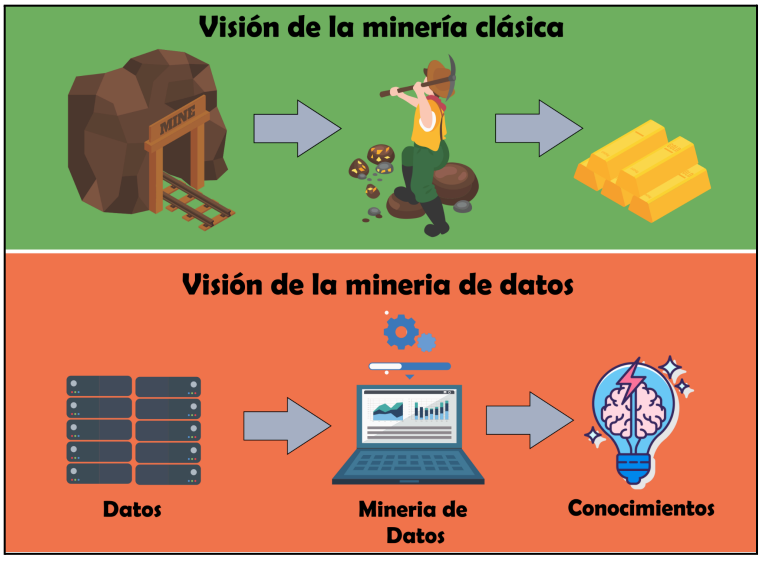
1.1. Definición y fundamentos de la minería de datos

Según Lara Torralbo (2014), fuera del ámbito informático, el término "minería" hace referencia a la actividad de extraer minerales de la corteza terrestre, pudiendo obtenerse materiales preciosos como el oro, la plata, cobre, el litio, entre otros. Por su parte, el término "dato" se define formalmente como el valor de una variable. Dicho término está estrechamente relacionado con la informática.

La combinación de ambos términos da lugar a un concepto que ha ganado relevancia en el campo de la informática: "minería de datos". Esta disciplina se enfoca en el análisis de grandes volúmenes de datos con el propósito de extraer conocimiento[2] útil de ellos. Al igual que en la minería tradicional, donde se busca obtener minerales de la corteza terrestre o del mar, la minería de datos tiene como objetivo descubrir patrones y conocimiento valioso a partir de los datos. En la Figura 1 se puede observar la comparación entre ambos tipos de minería.

Figura 1

De la minería tradicional a la minería de datos



Nota. Adaptado y actualizado de Visión de la minería de datos, de Minería de datos (p. 12), de J. A. Lara Torralbo, 2014, Universidad a Distancia de Madrid (UDIMA).

[2] Se refiere a los patrones, relaciones y tendencias ocultas que se descubren en grandes volúmenes de datos mediante técnicas analíticas y algorítmicas.

Perez Marquez (2005) define la minería de datos, como “un conjunto de técnicas encaminadas al descubrimiento de la información contenida en grandes conjuntos de datos” (p. 4). Para luego “analizar comportamientos, patrones, tendencias, asociaciones y otras características del conocimiento inmerso en los datos” (p. 4).

1.2. Proceso de extracción de conocimiento

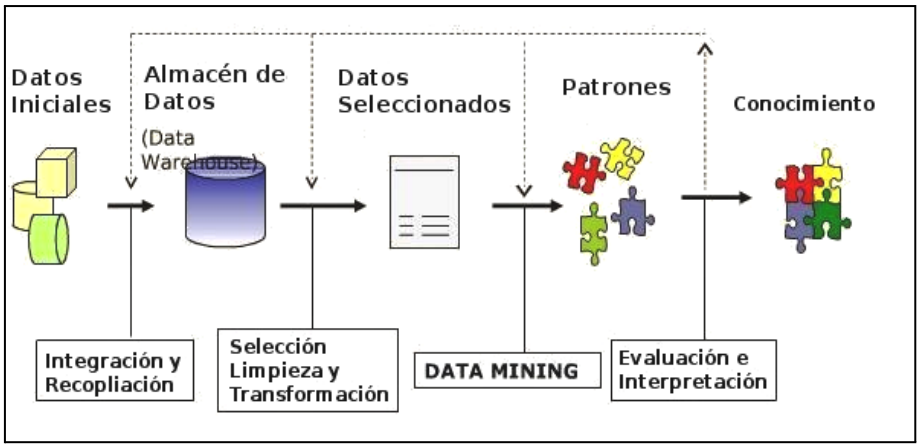
Según Lara Torralbo (2014), la minería de datos constituye una etapa esencial dentro de un proceso más amplio conocido como *Knowledge Discovery from Databases* (KDD), término utilizado para referirse a la extracción automatizada de conocimiento a partir de grandes volúmenes de datos (p. 39).

Este proceso se observa en la Figura 2, el proceso se compone de una secuencia iterativa de fases. La primera fase es la recopilación de los datos, la siguiente fase es la preparación, que es fundamental, ya que de ella depende en gran medida el éxito en la obtención de conocimiento. Tras la preparación, los datos pasan a la fase de minería de datos, en la que se aplican una serie de técnicas para obtener modelos (Lara Torralbo, 2014, p. 40–41).

Finalmente, los modelos generados en la fase de minería de datos deben ser validados e interpretados para extraer el conocimiento deseado (Lara Torralbo, 2014, p. 41).

Figura 2

Secuencia de pasos del KDD



Nota. Adaptado y actualizado de Proceso KDD, de Aplicaciones SIG en proyectos de ordenación de montes (p. 13), de I. Fernández Morís, 2015, Universidad de Oviedo

1.2.1. Fase de preparación de los datos

La primera etapa del KDD consiste en recopilar los datos y homogeneizarlos e integrarlos, produciendo lo que se conoce como *dataset*[3] o conjunto de datos. Esta es una base de datos con un formato unificado y, además, con una estructura más adecuada para el análisis de los datos. En este contexto, la integración de datos se refiere al proceso de combinar y armonizar datos de múltiples fuentes en un formato unificado y coherente que pueda utilizarse para diversos fines analíticos, operativos y de toma de decisiones (IBM, s.f.).

Dentro de esta fase de preparación, se deben llevar a cabo tres acciones fundamentales: selección, limpieza y transformación de datos. Estas etapas garantizan la calidad y coherencia de la información antes de su procesamiento y análisis.

1.2.1.1. Selección

Según Hernández Orallo et al. (2004) "la selección de atributos relevantes es uno de los preprocesamientos más importantes, ya que es crucial que los atributos utilizados sean relevantes para la tarea de minería de datos." (p. 23).

Además, el autor agrega que lo ideal sería utilizar todas las variables y que la herramienta de minería de datos pruebe todas ellas para poder así elegir las mejores variables predictoras, en la práctica este enfoque es ineficiente. La razón principal es que el tiempo de procesamiento necesario para generar un modelo aumenta en proporción directa con la cantidad de variables incluidas (p. 23).

Por esa razón, resulta necesario la utilización de técnicas de selección para priorizar las variables más importantes.

Según Lara Torralbo (2014), “las técnicas de selección tienen como objetivo filtrar aquellos datos que no son relevantes para el análisis posterior“ (p. 44-45).

- Filtrado de atributos: algunos atributos de los datos pueden no ser de interés para el análisis. Por ejemplo, si se dispone de información sobre los clientes de una compañía y uno de los atributos es el DNI, es probable que dicho atributo debe ser descartado, ya que no aporta valor al análisis de datos.
- Filtrado de registros: en ciertos casos, es necesario eliminar algunos registros almacenados. Por ejemplo, si una empresa solo está interesada en analizar clientes cuya edad supere un determinado umbral, los registros que no cumplan con este criterio podrían ser excluidos.

El filtrado de registros es importante porque en ocasiones se dispone de un número excesivo de ellos. En estos casos, trabajar con un subconjunto más pequeño de

[3] Es una colección organizada de datos que puede incluir números, texto, imágenes, estructurados en filas y columnas.

registros, denominado muestra, permite realizar un análisis igualmente efectivo pero mucho más eficiente desde el punto de vista computacional.

Para seleccionar una muestra representativa del conjunto completo de datos se aplican diversas técnicas de muestreo, entre las más relevantes se encuentran:

- muestreo aleatorio simple: todos los elementos del conjunto de datos tienen la misma probabilidad de ser seleccionados en la muestra.
- muestreo aleatorio estratificado: su objetivo es garantizar que todos los grupos o estratos dentro del conjunto de datos estén representados de manera equilibrada en la muestra.
- muestreo por conglomerados: se seleccionan únicamente registros pertenecientes a un grupo específico. Por ejemplo, una empresa podría analizar solo a los clientes que consumen más de 200 euros mensuales en sus productos.

1.2.1.2. Limpieza

En esta etapa, según lo expuesto por Lara Torralbo (2014), se busca eliminar los datos irrelevantes o innecesarios recolectados en la fase de preparación de los datos, dado que estos pueden afectar negativamente el proceso de minería de datos. Asimismo, deben abordarse otros problemas que comprometen la calidad de los datos, como la presencia de valores atípicos u *outliers.*

Lara Torralbo (2014) destaca que estos valores anómalos pueden originarse por errores o tratarse de datos válidos que difieren significativamente del resto. Algunos algoritmos de minería de datos los ignoran, mientras que otros los descartan por considerarlos ruidos o excepciones. Sin embargo, en algoritmos más sensibles, los valores atípicos pueden alterar considerablemente los resultados. Por ejemplo, según Alaminos-Fernández (2022), “los bosques aleatorios son generalmente robustos ante ruido y *outliers*, ya que promedian las predicciones de múltiples árboles de decisión” (p. 16). Por otro lado, el autor menciona que “algoritmos como máquinas de soporte vectorial (SVM) y vecinos más cercanos (k-NN) pueden verse afectados por la presencia de ruido y valores atípicos, al depender de márgenes y distancias que los extremos distorsionan” (p. 16).

No obstante, su eliminación no siempre es recomendable, ya que en determinados contextos pueden contener información valiosa.

1.2.1.3. Transformación

La transformación de datos es crucial en el análisis de datos por varias razones. Optimiza el almacenamiento y procesamiento de los datos, mejora la eficiencia en

consultas y almacenamiento. Además, es importante para la aplicación de modelos de *machine learning*, entendido como una rama de la inteligencia artificial que utiliza algoritmos y modelos estadísticos para que los sistemas aprendan automáticamente a partir de los datos, sin ser programados de manera explícita (IBM, s.f.), dado que muchos de estos algoritmos requieren datos numéricos y normalizados para funcionar correctamente.

Según Lara Torralbo (2014), existen múltiples técnicas de transformación de datos. Por su especial relevancia, se listan las más aplicadas:

- numerización: consiste en transformar un atributo cualitativo en un atributo cualitativo[4] equivalente. Por ejemplo, un atributo de tipo booleano se puede numerizar asignando el valor Falso como 0 y Cierto como 1.
- discretización: consiste en transformar un atributo cuantitativo en un atributo cualitativo ordinal. Por ejemplo, la altura de una persona en centímetros se puede discretizar en los siguientes intervalos:

  alto ($> 180\ cm$), medio (entre $150\ cm$ a $180\ cm$) y bajo ($< 150\ cm$).
- creación de características: consiste en generar un nuevo atributo a partir de otros ya existentes. Por ejemplo:

$$Sueldo\_Bruto\_Anual = Sueldo\_Bruto\_Mensual \cdot Numero\_Pagos$$

  Donde:

  - ➔ $Sueldo\_Bruto\_Anual$: ingreso total anual de una persona antes de los descuentos por ley o impuestos.
  - ➔ $Sueldo\_Bruto\_Mensual$: salario bruto es la cantidad total que se acuerda con el empleador.
  - ➔ $Numero\_Pagos$: Es la cantidad de pagos que se reciben al año.
- normalización**:** consiste en transformar el rango de valores de un atributo. La normalización lineal uniforme es una de las más comunes y convierte los valores de un atributo a una escala en el intervalo $[0, 1]$ utilizando la siguiente fórmula:

$$Valor\_Normalizado = \frac{Valor\_Inicial - Valor\_Mnimo}{Valor\_Maximo - Valor\_Minimo}$$

  Donde:

  - ➔ $Valor\_Normalizado$: el valor resultante después de aplicar la normalización (estará entre 0 y 1)
  - ➔ $Valor\_Inicial$: el valor original que se desea normalizar.

[4] El atributo cualitativo es una característica que se describe sin utilizar números.

➔ $Valor\_Minimo$: el valor mínimo del conjunto de datos.

➔ $Valor\_Maximo$: el valor máximo del conjunto de datos.

- reducción de dimensionalidad: estas técnicas buscan disminuir el número de atributos sobre los cuales se realiza el análisis posterior. Una de las más conocidas es el *principal components analysis* (PCA), que proyecta los atributos originales en un espacio de menor dimensionalidad, de manera que los nuevos atributos retienen la mayor parte de la información relevante, eliminando redundancias y dependencias. Esto permite un análisis más eficiente en términos de costo computacional.

1.2.2. Fase de minería de datos

Hernández Orallo et al. (2004) menciona que la fase de minería de datos "es la más característica del KDD y, por esta razón, muchas veces se utiliza esta fase para nombrar todo el proceso. El objetivo de esta fase es producir nuevo conocimiento que pueda utilizar el usuario" (p. 25).

Lara Torralbo (2014) agrega "las tareas de *data mining* se suelen clasificar dependiendo del tipo de modelo que son capaces de generar" (p. 48). En este sentido, el autor distingue dos tipos principales de tareas:

- tareas predictivas: se orientan a estimar valores que aún no se conocen de uno o varios atributos para nuevos registros, apoyándose en modelos aprendidos sobre la vista minable.
- tareas descriptivas: procuran construir modelos que caractericen y expliquen los datos disponibles, su estructura y relaciones. Sin intentar prever valores futuros.

En el caso de las tareas predictivas, se incluyen la clasificación y la regresión, mientras que el agrupamiento o *clustering*, las reglas de asociación, las reglas de asociación secuenciales y las correlaciones se consideran tareas descriptivas.

A continuación, se detallan estas tareas:

- clasificación: se entiende como la construcción de un modelo capaz de asignar a cada nuevo caso una de las clases predefinidas; por ejemplo, decidir si un cliente devolverá o no un préstamo a partir de un atributo como Devuelve_Préstamo con valores {Si, No}.
- regresión: se concibe como la estimación de un valor numérico cuantitativo[5] de interés. Por ejemplo, la nota final de un estudiante, utilizando información disponible sobre el caso, a diferencia de la clasificación que predice categorías.

[5] Son aquellas variables cuyas alternativas representan cantidades y, por ende, se expresan con números.

Por su parte, las principales tareas descriptivas de *data mining* son las siguientes:

- *clustering*: se concibe como el proceso de segmentar una población diversa en conjuntos internos uniformes (*clusters*), de modo que los elementos dentro de cada grupo compartan gran similitud entre sí. En la práctica, por ejemplo, una empresa puede agrupar a su clientela según variables como edad, sexo, cantidad de hijos e ingresos para identificar tipologías.
- asociación: se entiende como la identificación de reglas que revelan relaciones frecuentes entre atributos dentro de un conjunto de datos; en retail, por ejemplo, permite descubrir qué productos suelen adquirirse conjuntamente para apoyar decisiones comerciales.
- detección de atípicos: se concibe como la tarea de identificar casos cuyo comportamiento difiere marcadamente del patrón general del conjunto; por ejemplo, detectar operaciones potencialmente fraudulentas con tarjeta cuando aparecen compras inusuales en horario, monto o frecuencia.

1.2.3. Construcción del modelo

Según Hernández Orallo et al. (2004), la construcción del modelo permite evidenciar con mayor claridad el carácter iterativo del proceso de KDD, ya que resulta necesario explorar diversas alternativas hasta identificar aquella que resulte más efectiva para la resolución del problema. De este modo, una vez obtenido un modelo y analizados sus resultados, podría ser conveniente desarrollar otro utilizando la misma técnica con distintos parámetros o, en su defecto, aplicar enfoques y herramientas diferentes. En la búsqueda de un modelo adecuado, puede ser necesario retroceder a fases previas para realizar modificaciones en los datos utilizados o incluso ajustar la definición del problema.

Hernández Orallo et al. (2004) afirma que el desarrollo de modelos predictivos[6] exige una definición precisa de las etapas de entrenamiento y validación con el fin de garantizar predicciones robustas y precisas. El enfoque fundamental consiste en entrenar el modelo con una parte de los datos (*training dataset*) y posteriormente validarlo para evaluar su desempeño.

1.2.4. Fase de evaluación e interpretación

Hernández Orallo et al. (2004) señala que no todos los modelos generados en la fase previa cumplirán con las características esperadas para la adquisición de conocimiento. En particular, es esencial que estos modelos sean precisos, comprensibles e interesantes.

[6] Los modelos predictivos son una técnica estadística comúnmente utilizada para predecir comportamientos y resultados probables en el futuro.

Además destaca que, aunque esto depende de la tarea específica de *data mining*, en general, para evaluar un modelo se realiza mediante un enfoque que consiste en reservar un pequeño subconjunto de los datos (conjunto de prueba o *test*) que se utilizará para validar el modelo construido con el resto de los datos (conjunto de entrenamiento o *training*). Este enfoque se conoce como validación simple.

1.3. Aplicaciones de la minería de datos

Hernández Orallo et al. (2004) afirma que la minería de datos ha dejado de aplicarse exclusivamente en sectores como la distribución y la publicidad, y actualmente se utiliza cada vez más en diversas actividades cotidianas debido a su capacidad para optimizar procesos y mejorar resultados (p. 16).

En la Tabla 1 se sintetizan los principales sectores que se benefician del uso de minería de datos, junto con su objetivo típico y ejemplos de aplicación.

Tabla 1

Sectores que se benefician del uso de minería de datos

Aplicaciones financieras y banca
● Obtención de patrones de uso fraudulento de tarjetas de crédito. ● Determinación del gasto en tarjeta de crédito por grupos. ● Cálculo de correlaciones entre indicadores financieros.
Educación
● Selección o captación de estudiantes. ● Detección de abandonos y de fracaso. ● Estimación del tiempo de estancia en la institución.
Telecomunicaciones
● Establecimiento de patrones de llamadas. ● Modelos de carga en redes. ● Detección de fraude.
Análisis de mercado
● Análisis del carrito de compra (compras conjuntas, secuenciales, ventas cruzadas, etc.) ● Evaluación de campañas publicitarias. ● Análisis de la fidelidad de los clientes. Reducción de fuga.

Medicina

- Identificación de patologías. Diagnóstico de enfermedades.
- Detección de pacientes con riesgo de sufrir una patología concreta.
- Gestión hospitalaria y asistencial. Predicciones temporales para el mejor uso de recursos, consultas, salas y habitaciones.

Otras áreas

- Humanos: selección de empleados.
- Web: análisis del comportamiento de los usuarios, detección de fraude en el comercio electrónico, análisis de los logs de un servidor web.
- Turismo: determinar las características socioeconómicas de los turistas en un determinado destino o paquete turístico, identificar patrones de reservas.
- Tráfico: modelos de tráfico a partir de fuentes diversas: cámaras, GPS.

Capítulo 2: Situación de la educación universitaria y deserción estudiantil

En este capítulo se presenta un panorama general de la educación universitaria en Argentina y se analiza el fenómeno de la deserción estudiantil. Se define el concepto de deserción estudiantil, se identifican las variables que influyen en el abandono y se revisan las principales teorías explicativas. Además, se reflexiona sobre las implicancias sociales y educativas de esta problemática.

2.1. Situación de la educación universitaria pública-privada en Argentina

Sanseau et al. (2023) menciona que desde el año 2000, el acceso a la educación superior en Argentina, al igual que en la región, ha mostrado una marcada tendencia de crecimiento. Este aumento se debe principalmente a políticas que promueven el acceso igualitario a la educación y a la obligatoriedad de la escuela media. La expansión de la demanda es impulsada en gran medida por jóvenes provenientes de sectores sociales previamente excluidos, quienes buscan obtener credenciales que favorezcan su inserción en el mercado laboral.

Además afirma que este proceso de aumento de la demanda fue acompañado y reforzado por un crecimiento en la oferta educativa, con la creación de 50 instituciones universitarias durante el período 2000-2023.

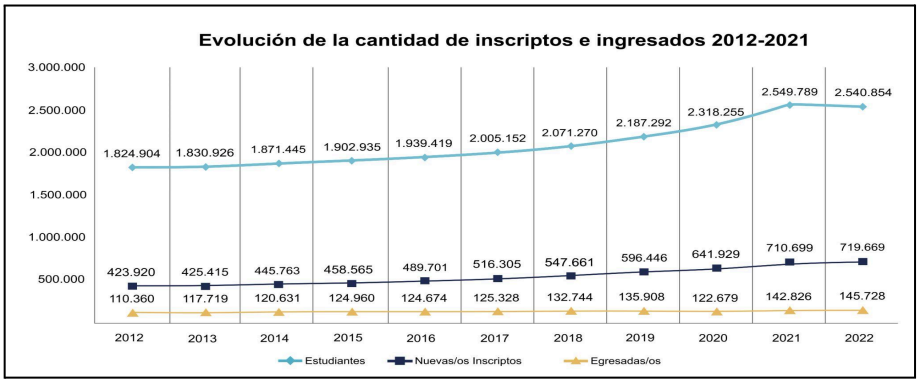
Según datos de la Secretaría de políticas universitarias (SPU):

La matrícula en educación universitaria de pregrado y grado pasó de *1.553.700* estudiantes en 2005 a *1.902.935* en 2015. Para el año 2020, la matrícula ascendió a *2.318.255* estudiantes , lo que representa un crecimiento del *49,2%* entre 2005 y 2020 (Sanseau et al., 2023, p. 6).

Los datos mencionados anteriormente se observan en la Figura 3.

Figura 3

Evolución de la cantidad de estudiantes 2012-2022



Nota. Tomado de Síntesis de información estadística universitaria 2022-2023 (p. 26), Ministerio de Educación de la Nación Argentina, Secretaría de Políticas Universitarias (2023).

Según Sanseau et al. (2023) estos datos son alentadores en términos del alcance de la educación universitaria en Argentina; sin embargo, existen limitaciones significativas respecto a la continuidad en las trayectorias educativas de los estudiantes. Una de las hipótesis con mayor respaldo para explicar este fenómeno es que, pese a las elevadas tasas de ingreso, las brechas en capital cultural, social y económico afectan negativamente la permanencia y la graduación, que continúan siendo bajas en comparación con América del Norte y Europa.

En Argentina, de acuerdo a información de la SPU:

> El 38,1*%* de ingresantes a carreras universitarias en 2019 se desvincularon al año siguiente, es decir durante su primer año de cursada. Respecto a la graduación, en el año 2020 se registró un total de *122.679* egresados/as, lo cual supone un aumento del *12,2%* desde 2011 (Sanseau et al., 2023, p. 6).

Bajo ese contexto, Sanseau et al. (2023) advierten que de este total, solo el 25,1% logró alcanzar la titulación en el tiempo teórico estimado para cada carrera.

En este marco, la cuestión del acceso y la permanencia de los estudiantes adquiere una importancia central, sobre todo en el primer año de estudios. Aunque la cantidad de inscripciones ha mostrado un crecimiento sostenido, dicho aumento no se traduce en un número proporcional de egresados, lo que pone en evidencia las dificultades que enfrentan los alumnos para sostener sus trayectorias y concluir sus carreras.

2.2. Deserción estudiantil

En este aspecto se considera que:

> La deserción estudiantil, entendida no sólo como el abandono definitivo de las aulas de clase, sino como el abandono de la formación académica, independientemente de las condiciones y modalidades de presencialidad, es una decisión personal del individuo y no obedece a un retiro académico forzoso, por el no éxito del estudiante en el rendimiento académico, como es el caso de la expulsión por bajo promedio académico o a un retiro por asuntos disciplinarios. Se dice entonces que la deserción es una opción del estudiante, influenciado positiva o negativamente por circunstancias internas o externas (Páramo & Correa Maya, 1999, p. 66).

Algunos autores señalan que:

> La deserción o el abandono universitario pueden ser definidos de manera operativa como la cantidad de estudiantes que no se matriculan en una carrera del sistema de educación superior entre uno y otro período

académico. No obstante, la problemática reviste una complejidad mayor. El abandono puede ser definitivo o bien temporal.

Por otra parte, no siempre la ausencia de rematriculación significa el abandono del sistema universitario, en ocasiones corresponde a cambios de carrera, cambios de instituciones o breves discontinuidades sin pérdida de la regularidad (Sanseau et al., 2023, p. 8).

En la literatura reciente sobre educación superior en Argentina se observan distintas formas de operacionalizar la deserción y el abandono. Sanseau et al. (2023, p. 8) proponen medir el abandono a partir de la ausencia de rematriculación entre períodos académicos consecutivos y advierten que esta situación puede corresponder tanto a abandonos definitivos como a cambios de carrera o de institución. En un estudio de caso sobre la carrera de Bioingeniería de la Universidad Nacional de San Juan, Seminara (2020, p. 90) define al alumno desertor como todo estudiante que no se inscribió en el sistema durante dos años consecutivos y distingue, además, la categoría de *alumno crónico* en función de la cantidad de materias aprobadas por año respecto de la trayectoria ideal prevista en el plan de estudios. Por su parte, Calloni (2025, p. 10) operacionaliza el "riesgo de abandono" en la Facultad de Ciencias Exactas y Naturales de la UBA a partir de un umbral mínimo de actividad académica por semestre, calculado a partir de inscripciones, aprobaciones y finales registrados en sistema.

Estos trabajos evidencian, de manera concreta, que la deserción y el abandono universitario suelen definirse mediante criterios cuantitativos basados en umbrales de reinscripción, actividad académica o rendimiento, ajustados a las particularidades de cada institución y carrera. En la presente tesina se retoman y adoptan estas perspectivas operativas.

2.3. Variables asociadas a la deserción estudiantil

Según Páramo y Correa Maya (1999), "de las variables asociadas con la deserción, algunas tienen mayor o menor grado de significancia e incidencia en la misma" (p. 2). Además, agregan que "podrían citarse como variables, entre otras: Misión-Visión de la institución educativa, ambientes educativos, modelos pedagógicos, cultura universitaria, perfil ocupacional y profesional" (p. 3).

Existen varias teorías sobre la deserción de los estudiantes a lo largo de los años. A continuación se mencionan dos de ellas. Según el modelo de Tinto[7] (1975, 1987), la decisión de permanecer o abandonar los estudios depende del grado de integración del estudiante en la institución, tanto en lo académico como en lo social. Este proceso se ve

[7] Hace referencia a un enfoque teórico que analiza el abandono o la permanencia en la universidad desde una perspectiva organizacional.

afectado desde el ingreso por tres conjuntos de factores previos: rasgos personales y familiares, habilidades y capacidades, y experiencias preuniversitarias. Dichos factores influyen en las metas y el compromiso con la universidad y permiten anticipar experiencias favorables o desfavorables al inicio de la trayectoria académica (Fonseca & García, 2016, p. 31). Por su parte, Fonseca y García, el modelo de Bean y Metzner (1985) intenta explicar el abandono universitario de estudiantes no tradicionales, refiriéndose a aquellos que poseen características diferenciadas del alumnado tradicional. Por ejemplo: son mayores de edad, no residen en campus universitarios y cursan estudios a tiempo parcial. Considerando esta tipología, estos autores desarrollan un modelo explicativo basado en cuatro dimensiones que inciden en la decisión de abandonar los estudios (p. 31-32):

- rendimiento académico: distingue a quienes permanecen de quienes abandonan. El promedio académico es un indicador importante.
- variables académicas y psicológicas: se consideran a los hábitos de estudio, asesoramiento académico recibido, ausentismo, compromiso, satisfacción, estrés académico, y certeza respecto a los objetivos educativos.
- antecedentes preuniversitarios: abarca aspectos como el desempeño en la educación secundaria, el estado de inscripción, la edad, la etnia, el género, etc.
- variables ambientales: son variables externas a la universidad. Abarcan la utilidad percibida de los estudios, las responsabilidades familiares, la cantidad de horas de trabajo, los estímulos externos, y la posibilidad de transferirse a otra institución.

En síntesis, Bean y Metzner (1985) concluyen que “si las variables académicas y ambientales presentan resultados propicios, los estudiantes optan por quedarse en la institución; no obstante,si las variables ambientales son adversas, tienden a abandonar la universidad, incluso si las variables académicas son ventajosas” (p. 33).

Sin embargo, para Páramo y Correa Maya (1999) la deserción estudiantil está asociada a ciertas variables, que influyen con diferente intensidad. Algunas son:

- ambientes educativos universitarios.
- ambientes familiares.
- proceso educativo y acompañamiento al estudiante en su formación.
- edad. La mayoría de los estudiantes universitarios son muy jóvenes.
- adaptación social del estudiante desertor con sus pares u homólogos.
- bajos niveles de comprensión unidos a la falta de interés y apatía por programas curriculares.
- modelos pedagógicos universitarios diferentes a los modelos de bachillerato, que imprime un alto nivel de exigencia.

- programas micro-curriculares universitarios rígidos con respecto a los de su formación secundaria, de alta intensidad temática, dispuestos en corto tiempo.
- evaluaciones extenuantes y avasalladoras. Las evaluaciones y trabajos universitarios tienen mucho mayor nivel de complejidad que las de secundaria.
- cursos no asociados ni aplicables con su ejercicio profesional.
- factores económicos que impiden la continuidad del desertor en la Universidad. Cantidad de oferentes.
- orientación profesional.

2.4. Teorías sobre la deserción estudiantil

Según Cabrera et al. (2006), “a la hora de valorar el abandono académico en la enseñanza superior se han propuesto diferentes modelos y teorías explicativas[8], que nosotros agrupamos en cuatro grandes enfoques: modelo de adaptación, modelo estructural, modelo economicista y modelo psicopedagógico” (p. 182).

En relación con los enfoques explicativos, Cabrera et al. (2006) exponen los modelos previamente mencionados:

- modelo de adaptación: sostiene que la deserción ocurre cuando el estudiantado no alcanza una integración académica y social suficiente en la vida universitaria. La sensación de no pertenencia favorece el aislamiento y promueve la reasignación del tiempo y los recursos hacia actividades percibidas como más beneficiosas o menos costosas.
- modelo estructural: interpreta la deserción como resultado de tensiones y desigualdades propias de los subsistemas político, económico y social. Desde una perspectiva crítica, se reconoce que la universidad puede reproducir dichas condiciones, influyendo en la decisión de abandono de ciertos grupos.
- modelo economicista: concibe la deserción como una decisión racional en la que se comparan costos y beneficios de continuar los estudios frente a usos alternativos del tiempo y los recursos. Si las alternativas prometen mayores rendimientos futuros que la permanencia, se opta por el retiro. Se fundamenta en la teoría del capital humano.
- modelo psicopedagógico: integra aportes de los modelos de adaptación y estructural, incorporando variables psicológicas y educativas. Plantea que factores como las estrategias de aprendizaje, la postergación de recompensas, la calidad del vínculo docente–estudiante, la perseverancia ante obstáculos y la claridad de metas inciden de manera determinante en la permanencia o el abandono.

[8] Las teorías explicativas son marcos que buscan entender las causas y efectos de un fenómeno.

2.5. La deserción como problema educativo y social

Los autores Páramo y Correa Maya (1999) consideran que:

> A menudo los educadores y administradores de la educación se refieren a la deserción como la enfermedad más aguda del sistema y tratan de diseñar correctivos para hacerla rebajar al mínimo, ésto ocurre especialmente en los niveles educativos superiores universitarios (p.70).

Páramo y Correa Maya (1999) sostienen que las reformas educativas implementadas en el país para enfrentar la problemática de la deserción lograron resultados positivos en un inicio; sin embargo, estas medidas no se mantuvieron en el tiempo y finalmente fueron abandonadas.

De acuerdo con Delors (1996), la prioridad de los sistemas educativos debería ser reducir la vulnerabilidad social de niños y jóvenes de contextos desfavorables para romper el círculo de pobreza y exclusión. Para ello, propone identificar tempranamente las desventajas a menudo ligadas al entorno familiar y aplicar políticas de discriminación positiva para quienes enfrentan mayores dificultades. En esa línea, Páramo y Correa Maya (1999) subrayan la necesidad de definir indicadores de deserción aplicables en el ámbito universitario. Además destacan la necesidad de instaurar métodos pedagógicos especiales. Proponen, en este sentido, la flexibilización de los programas académicos profesionales, la reorientación de los perfiles y la modernización tanto de la administración como de la infraestructura física cuando sea necesario.

Finalmente, consideran que la deserción representa, por excelencia, “un problema del sistema educativo, íntimamente relacionado con sus entornos, contornos y dintornos, tales como los ambientes educativos, las situaciones familiares y las exigencias ambientales y culturales” (Páramo y Correa Maya, 1999, p. 71) que inciden directamente sobre quien abandona sus estudios

Capítulo 3: Diseño Metodológico y Tratamiento de Datos

En este capítulo se presenta el enfoque metodológico adoptado para el desarrollo del trabajo. Se describe el tipo de investigación y el diseño elegido, así como las características de la población objeto de estudio.

3.1. Enfoque y diseño de la investigación

Según Hernández Sampieri (2014), “la investigación es un conjunto de procesos sistemáticos, críticos y empíricos que se aplican al estudio de un fenómeno o problema” (p. 4). En el presente estudio se adopta un enfoque cuantitativo, por cuanto se ajusta a la naturaleza del problema: parte de un planteamiento definido y, mediante el análisis de datos, busca proporcionar una respuesta al problema formulado; se trabaja con datos numéricos y procedimientos estadísticos (Hernández Sampieri, 2014, p. 4–6).

La investigación se encuadra en un diseño no experimental, el cual, según Hernández Sampieri (2014), “podría definirse como la investigación que se realiza sin manipular deliberadamente variables. Es decir, se trata de estudios en los que no se hacen variar en forma intencional las variables independientes para ver su efecto sobre otras variables” (p. 152). En este caso, se analizan datos provistos por la facultad, sin manipulación de las variables, utilizando únicamente la información existente.

Por último, el estudio se enmarca en un diseño longitudinal, dado que, como indica Hernández Sampieri (2014), “los diseños longitudinales recolectan datos en diferentes momentos o periodos para hacer inferencias respecto al cambio, sus determinantes y consecuencias. Tales puntos o periodos generalmente se especifican de antemano” (p. 154). Esto concuerda con el presente trabajo, que analiza observaciones a lo largo de 10 años.

3.2. Población

La población de la investigación serán los estudiantes que se inscribieron en el periodo comprendido entre 2013 y 2023 a las carreras de Análisis de sistemas y Licenciatura de sistemas de información de la Facultad de Ciencia y Tecnología de la Universidad Autónoma de Entre Ríos, sita en Oro Verde.

3.3. Muestreo

La muestra estará conformada por la totalidad de estudiantes que cursaron durante el período antes mencionado. Los datos serán extraídos del sistema de gestión académica SIU Guaraní, que contiene la información actualizada y detallada de los estudiantes matriculados, sus inscripciones y avances en las distintas asignaturas. Esta

base de datos permite obtener un panorama completo y confiable de la población objeto de estudio, facilitando un análisis cuantitativo riguroso y representativo.

Es importante destacar que, para preservar la privacidad y confidencialidad de los estudiantes, no se utilizaron datos sensibles ni información personal identificable. De acuerdo con la normativa argentina sobre protección de datos personales, se entiende por datos personales toda información relativa a una persona física o jurídica que permite identificarla, mientras que los datos sensibles abarcan, entre otros aspectos, datos sobre origen racial o étnico, opiniones políticas, convicciones religiosas o filosóficas, afiliación sindical, salud o vida sexual (Argentina.gob.ar, s.f.). En este estudio solo se emplearon variables estrictamente necesarias para el análisis, garantizando el cumplimiento de la Ley 25.326 y demás disposiciones vigentes sobre protección de datos.

3.4. Herramientas de software utilizadas

En cuanto a las herramientas de software, se trabajó en Google Colab, una plataforma gratuita de Google que permite escribir y ejecutar código Python directamente desde el navegador, sin necesidad de instalación ni configuración adicional.

El lenguaje de programación utilizado fue Python, junto con librerías que facilitaron el tratamiento de los datos y la construcción de los modelos, entre ellas:

- Pandas: manipulación y análisis de datos en forma de tablas.
- NumPy: operaciones numéricas y manejo de arreglos.
- Scikit-learn: implementación de algoritmos de aprendizaje automático.

Estas herramientas se emplearon para las tareas de limpieza y preparación de los datos, el entrenamiento de los modelos de clasificación y la evaluación de su desempeño.

3.5. Recolección de datos

Los datos fueron obtenidos a partir de los registros académicos de los estudiantes, los cuales pueden incluir calificaciones, participación en actividades extracurriculares y resultados en pruebas estandarizadas. La fuente exclusiva de esta información fue la base de datos académica proporcionada por la Secretaría Académica de la Facultad de Ciencia y Tecnología.

Se dispuso inicialmente de un conjunto amplio de variables, detalladas en el listado que se presenta en las Tablas 2, 3 y 4. No obstante, a lo largo del desarrollo del estudio se llevó a cabo un proceso de depuración y análisis que se presenta en el capítulo 5. Cabe destacar que Hernández Orallo et al. (2004) señalan que, en minería de datos, no es eficiente utilizar todas las variables disponibles. “Obviamente, esta forma de

trabajar no funciona bien, entre otras cosas porque el tiempo requerido para construir un modelo crece con el número de variables" (p.23).

En la misma línea, Lara Torralbo (2014) indica que las técnicas de selección tienen como objetivo filtrar datos que no son relevantes para el análisis posterior.

Tabla 2

Variables académicas

Variables académicas	
Variable	Descripción
Carrera	Nombre de la carrera universitaria en la que está inscripto.
Título	Título que se otorga al finalizar la carrera correspondiente.
Plan_nombre	Nombre o código del plan de estudios vigente.
Año_ingreso	Año en que el alumno se inscribió por primera vez a la carrera.
Egresado	Información si tiene algún título previo de pregrado o grado.
Porcentaje_avance	Porcentaje de materias aprobadas en relación al total del plan de estudios.
Cant_aprobadas	Total de materias aprobadas hasta la fecha del análisis.
Promedio_con_aplazo	Promedio general de notas, incluyendo las notas de los aplazos
Promedio_sin_aplazo	Promedio general de notas considerando solo las materias aprobadas.
Cnt_materias	Total de materias obligatorias previstas en el plan de estudios.
Año_reinsc	Año de la última reinscripción administrativa realizada.
Cant_equiv	Número de materias acreditadas por equivalencia provenientes de otras instituciones o carreras.

Nota. La tabla presenta las principales variables académicas consideradas en el análisis, junto con una breve descripción de su significado y alcance en el contexto del estudio.

Tabla 3

Variables sociodemográficas

Variables sociodemográficas	
Variable	Descripción
Sexo	Género declarado (por ejemplo: M o F).
Cant_hijos	Número de hijos (por ejemplo: no tiene, uno, dos, más de dos).
Situacion_padre	Situación actual del padre del alumno (por ej.: vive, fallecido, ausente).
Situacion_madre	Situación actual de la madre del alumno (por ej.: vive, fallecida, ausente).
Tipo_vivienda	Tipo de vivienda en la que reside el alumno (por ej.: propia, alquilada, prestada).
Loc_per_lect	Localidad o lugar de residencia del alumno durante el período lectivo.
Loc_origen	Localidad de origen del alumno antes de comenzar sus estudios universitarios.

Nota. La tabla muestra las variables sociodemográficas consideradas en el estudio.permiten contextualizar el perfil del estudiante, aportando información sobre su entorno familiar, económico y residencial.

Tabla 4

Variables económicas y laborales

Variables económicas y laborales	
Variable	Descripción
Costeo_familiar	Indica si los estudios son financiados por su familia.
Costeo_social	Indica si recibe algún tipo de ayuda social para costear sus estudios.
Costeo_trabajo	Indica si se financia sus estudios con ingresos provenientes de su trabajo.
Costeo_beca	si recibe una beca para financiar sus estudios
Trabajo_hace	Indica si actualmente se encuentra trabajando.
Trabajo_existe	Indica si ha tenido alguna experiencia laboral previa.
Trabajo_ocupacion	Tipo o rubro del trabajo que desempeña.
Horas_sem	Cantidad de horas semanales que el alumno dedica a su trabajo.
Trabajo_carrera	Indica si el trabajo actual del está relacionado con su carrera universitaria.

Nota. Las variables económicas y laborales aportan información clave sobre las condiciones de sostenimiento y dedicación del estudiante.

3.6. Metodología del proceso de minería de datos (CRISP-DM)

Según IBM (2021), *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*) es un estándar ampliamente aceptado que sirve como guía para el desarrollo de proyectos de minería de datos. Como metodología, describe las fases típicas de un proyecto, las tareas asociadas a cada fase y las relaciones entre dichas tareas. Además agrega que “el modelo de *CRISP-DM* es flexible y se puede personalizar fácilmente”.

Las fases del modelo *CRISP-DM* se describen del siguiente modo, siguiendo la descripción de Hernández Orallo et al. (2004, p. 601–602):

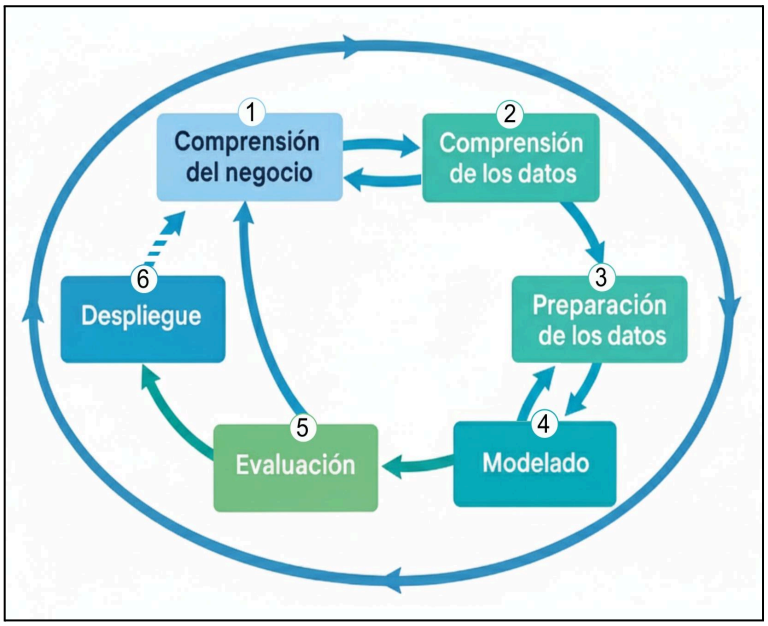
1. comprensión del negocio: se centra en recopilar los datos disponibles, conocer sus características, detectar problemas de calidad y explorar de forma preliminar su potencial para responder a los objetivos de negocio.
2. comprensión de los datos: se trata de recopilar y familiarizarse con los datos, identificar los problemas de calidad de datos y ver las primeras potencialidades o subconjuntos de datos que puede ser interesante analizar (para ello es muy importante haber establecido los objetivos de negocio en la fase anterior).
3. preparación de los datos: el objetivo es construir el conjunto de datos que se utilizará para el modelado, mediante selección, limpieza, construcción, integración y formateo de los datos.

4. modelado: en esta fase se aplican las técnicas de minería de datos sobre el conjunto preparado, ajustando parámetros y obteniendo uno o varios modelos candidatos.
5. evaluación: aquí se analiza si los modelos obtenidos no solo son válidos estadísticamente, sino también útiles para responder a los requerimientos planteados en la fase de negocio y decidir los pasos a seguir.
6. despliegue: consiste en poner en práctica los resultados del modelado; esto implica integrar los modelos en los procesos de decisión, elaborar informes y documentar la experiencia.

En la Figura 4 se observa las fases del modelo antes descritas.

Figura 4

Fases del modelo de referencia CRISP-DM



Nota. Adaptado de *Introducción a la minería de datos* (p. 601), por J. Hernández Orallo, M. J. Ramírez Quintana y C. Ferri, 2004, Pearson Educación.

3.7. Limpieza de datos

Amazon Web Services (s.f.) considera a la limpieza de datos como una etapa crucial, ya que los datos sin procesar pueden contener numerosos errores, que pueden afectar a la precisión de los modelos de *machine learning* y dar lugar a predicciones incorrectas. Según Lara Torralbo (2014), las tareas de limpieza de datos intentan resolver dos problemas frecuentes: los valores atípicos y los valores faltantes.

3.7.1 Valores atípicos y valores erróneos

Los valores erróneos se refiere a completar un campo con valores incorrectos.

Como menciona Lara Torralbo (2014) también “suele ser común encontrar valores que, claramente, son erróneos”. Por ejemplo, es posible encontrar un código postal inexistente en una dirección” (p. 46).

Según Ferrero (2017), un dato atípico u *outlier* es una observación que se aleja considerablemente del resto de datos en un conjunto. Estos valores pueden tener un gran impacto en el análisis estadístico, especialmente si se utilizan medidas sensibles a su presencia, como la media aritmética.

Por otro lado, Hernández Orallo et al. (2004) menciona que “un valor erróneo y un valor anómalo no son lo mismo. Como hemos dicho, existen casos en los que los valores extremos se categorizan como anómalos estadísticamente pero son correctos, es decir, representan un dato fidedigno de la realidad” (p. 77).

3.7.2 Valores faltantes o *missing values*

Según Lara Torralbo (2014), sucede con frecuencia que, para muchos de los registros analizados, falta cierta información. Estos valores ausentes se denominan valores faltantes o *missing values*. Por ejemplo, si una persona no tiene asociado un valor para el atributo $Estado\_Civil$, puede deberse a que dicha persona no desea que se conozca dicha información.

También destaca que es importante analizar cautelosamente los valores faltantes ya que, en ocasiones, aportan información interesante (p. 45).

Para Lara Torralbo (2014), las causas más comunes de la presencia de valores faltantes se encuentran en:

- fallas en el dispositivo de lectura de datos.
- cambios en los procedimientos de recolección.
- uso de múltiples fuentes de datos, lo que puede generar inconsistencias.

Por lo tanto, resulta imprescindible analizar la causa específica del dato faltante antes de elegir la estrategia de tratamiento más adecuada.

3.7.3 Tratamientos de valores atípicos y erróneos

Lara Torralbo (2014) señala que, una vez identificados los valores erróneos o valores atípicos, las opciones más comunes para su tratamiento son las siguientes:

- pasar por alto el valor erróneo y continuar con el análisis.
- eliminar toda la columna asociada al valor erróneo.
- eliminar el registro que contiene el valor erróneo.
- reemplazar el valor erróneo por un valor correcto, el cual puede ser estimado mediante técnicas de predicción (media, mediana o moda).

3.7.4 Tratamientos de datos faltantes

El tratamiento de los valores faltantes es una etapa crítica en el preprocesamiento de datos, ya que su presencia puede comprometer la calidad de los modelos de minería de datos. Según Hernández Orallo et al. (2004), los valores faltantes (*missing values*) pueden ser reemplazados por varias razones:

- el método de minería de datos empleado puede no tratar bien los campos faltantes.
- se desea agregar datos pero los valores faltantes no permiten agregarlos correctamente (totales, medias, etc.)
- si el método es capaz de tratar campos faltantes es posible que ignore todo el ejemplo (produciendo un sesgo) o el método de sustitución de campos faltantes que no sea adecuado debido a que no conoce el contexto.

Hernández Orallo et al. (2004), advierten que un problema frecuente es que “a veces los campos faltantes no están representados como nulos” (p. 74). Por ejemplo, en campos sin formato: direcciones o teléfono como “no tiene”, códigos postales o números de tarjeta de crédito con valor –1, etc. Además, plantea que “a veces son las propias restricciones de integridad las que causan el problema.” (p. 75)

Finalmente, lista las posibles acciones sobre datos faltantes son:

- ignorar: algunos algoritmos son robustos a datos faltantes (por ejemplo árboles de decisión).
- eliminar: solución extrema.
- filtrar: sesga los datos, muchas veces las causas de un dato faltante están relacionadas con casos o tipos especiales.
- reemplazar el valor: se puede intentar reemplazar el valor manualmente (si son pocos) o automáticamente por un valor que preserve la media o la varianza para valores numéricos, o por la moda para valores nominales. Una manera más sofisticada de estimar un valor es predecirlo a partir de los otros ejemplos, esto se denomina “imputación”, utilizando cualquier técnica predictiva de aprendizaje automático (clasificación o regresión) o técnicas más específicas (por ejemplo, determinar el sexo a partir del nombre).
- segmentar: se segmentan las tuplas por los valores que tienen disponibles. Se obtienen modelos diferentes para cada segmento y luego se combinan.
- modificar la política de calidad de datos y esperar hasta que los datos faltantes estén disponibles.

Según Hernández Orallo et al. (2004), quizás la solución más frecuente sea reemplazar el valor. Esto conlleva a la pérdida de información, o a que se invente información, con los riesgos que pueda tener de que sea errónea.

También plantea que cuando son datos poco densos porque tienen niveles muy altos de valores faltantes (encima del 50 por ciento), reemplazar o filtrar no parece ser la solución.

En el contexto del tratamiento de valores faltantes, la imputación simple se presenta como una alternativa ampliamente utilizada. Consiste en aplicar un algoritmo que realiza una única estimación y emplea el valor obtenido para reemplazar el dato faltante (Codificando Bits, s.f.). Según este recurso, las tres técnicas más comunes en el ámbito del *machine learning* y la ciencia de datos son:

- imputación por la media o la mediana: consiste en calcular la media o mediana (Ver Anexo B), de los valores conocidos en la variable afectada y utilizar ese valor para reemplazar los datos faltantes. Aunque es fácil de implementar, este método tiene como desventaja que, al reemplazar múltiples datos faltantes con un mismo valor, se puede distorsionar la distribución original de la variable.
- imputación por regresión: en este caso cada dato faltante es reemplazado con el valor predicho por un modelo de regresión (Ver Anexo B) . Para esto primero se combina la información de la columna con los datos faltantes con columnas en donde los datos están completos para así ajustar un modelo de regresión. Y luego se usa este modelo para predecir los datos faltantes.
- imputación por paquete caliente: en este caso, el dato faltante es reemplazado con valores tomados de datos “cercanos” al dato faltante. Dentro de esta categoría el método más usado es el de k-vecinos más cercanos. Este algoritmo busca los k valores más cercanos (donde k es un número entero, como 2, 3, o 10 por ejemplo) y reemplaza el valor faltante con el promedio de estos vecinos.

En relación con las técnicas de imputación, una alternativa más robusta que las mencionadas anteriormente es la imputación múltiple, considerada en la actualidad una de las más utilizadas (Codificando Bits, s.f.).

En la imputación múltiple, en lugar de realizar una única estimación se generan múltiples estimaciones, que luego se combinan para producir un único valor, que será el usado para reemplazar el dato faltante correspondiente, con lo cual se puede disminuir el sesgo de la estimación.

El método de imputación múltiple más usado es el algoritmo *multiple imputation by chained equations* (MICE)[9]. En este algoritmo, de forma iterativa, se harán progresivamente cada vez mejores estimaciones de los datos faltantes.

[9] Es un algoritmo que realiza imputación de datos faltantes mediante la creación de múltiples modelos de regresión iterativos para cada variable incompleta.

Inicialmente la primera estimación no es muy buena, pues se hace con la imputación por la media que referida anteriormente, pero en los pasos restantes se aplica una regresión lineal (Apéndice B), entre pares consecutivos de columnas, y el procedimiento se repite una y otra vez hasta completar un número pre-definido de iteraciones.

El objetivo de este procedimiento es que, a medida que avanzan las iteraciones, las estimaciones se vuelvan progresivamente más precisas y converjan hacia el valor real.

Capítulo 4: Técnicas de minería de datos

En este capítulo se describen las técnicas de minería de datos que pueden ser utilizadas para el análisis y predicción de la deserción estudiantil. Cada una se presenta con sus características principales, ventajas, desventajas y aplicaciones típicas.

4.1. Aprendizaje supervisado y no supervisado

Antes de describir las técnicas, es necesario presentar dos conceptos fundamentales en el campo del aprendizaje automático, los cuales son definidos por Google Cloud (s.f.):

- aprendizaje supervisado: es una categoría del aprendizaje automático que usa conjuntos de datos etiquetados para entrenar algoritmos para predecir resultados y reconocer patrones. Los datos que se usan en el aprendizaje supervisado están etiquetados, lo que significa que contienen ejemplos de entradas (llamadas atributos) y salidas correctas (etiquetas). Los algoritmos analizan un gran conjunto de datos de estos pares de entrenamiento para inferir cuál sería un valor de resultado deseado cuando se les pide que hagan una predicción sobre datos nuevos.
- aprendizaje no supervisado: se basa en el uso de datos sin etiquetar, es un tipo de aprendizaje automático que aprende de los datos sin supervisión humana. El modelo recibe datos sin procesar, sin etiqueta, tiene que inferir sus propias reglas y estructurar la información en función de similitudes, diferencias y patrones sin instrucciones explícitas sobre cómo trabajar con cada dato.

4.2. Agrupamiento

El agrupamiento, también denominado *clustering***,** constituye una de las tareas descriptivas más representativas de la minería de datos. Su finalidad es identificar agrupaciones “naturales” dentro del conjunto de datos, sin la necesidad de contar con etiquetas previas. A diferencia de los métodos de clasificación, el agrupamiento no parte de clases predeterminadas, sino que las infiere a partir de las similitudes entre los datos. El principio básico consiste en maximizar la semejanza entre los elementos de un mismo grupo y, a su vez, minimizar entre los diferentes grupos. Esta técnica también recibe el nombre de segmentación, ya que divide los datos en subconjuntos que pueden o no ser excluyentes entre sí. (Hernández Orallo et al., 2004).

Por otro lado, según Lara Torralbo (2014), para poder establecer los diferentes grupos de objetos similares entre sí, es necesario contar con un mecanismo que permita comparar cada par de objetos. Entonces, se requiere una medida de distancia o

similaridad entre dichos objetos. Se dividen en: clúster jerárquicos y clúster no jerárquicos.

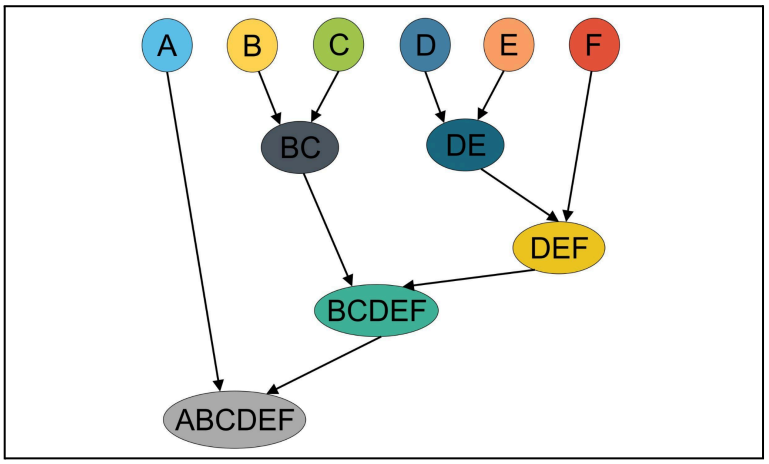
4.2.1. *Clusters* jerárquicos

AprendeIA (2020) destaca que las técnicas jerárquicas generan agrupamientos en forma de sucesión ordenada, formando una estructura en árbol denominada dendograma[10]. Entre los algoritmos más destacados se encuentran AGNES (estrategia aglomerativa) y DIANA (estrategia divisiva).

- aglomerativo: este algoritmo, que se puede observar en la Figura 5, se inicia con puntos de datos individuales y, de forma sucesiva, se fusionan los clústeres mediante el cálculo de la matriz de proximidad de todos los clústeres en el nivel actual de la jerarquía para crear una estructura similar a un árbol. Una vez que se ha creado un nivel de clústeres donde todos los clústeres tienen poca o ninguna similitud entre ellos, el algoritmo se mueve al conjunto de clústeres recién creados y repite el proceso hasta que haya un nodo raíz en la parte superior del gráfico jerárquico.

Figura 5

Clúster aglomerativo



Nota. Tomado de Algoritmo Agrupamiento Jerárquico-Teoría, por Aprende IA (2020).

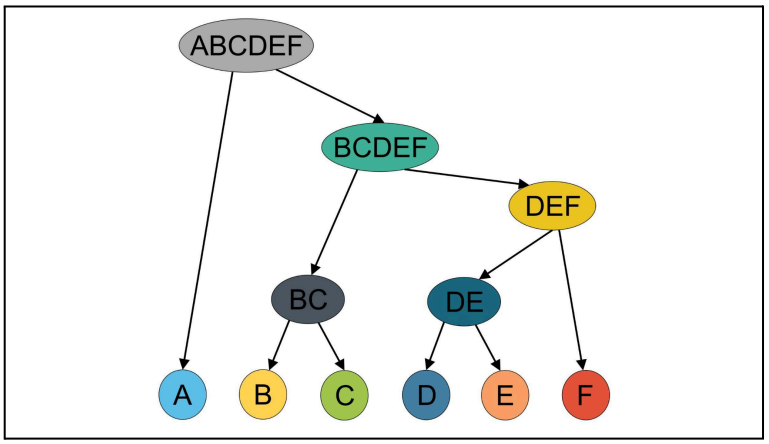
- divisivo: un enfoque de arriba hacia abajo divide sucesivamente los puntos de datos en una estructura similar a un árbol. El primer paso es dividir el conjunto de datos en clústeres utilizando un método de agrupación plana como *k-means*. Los clústeres con la mayor suma de errores cuadrados (SSE), se particionan posteriormente mediante un método de agrupamiento plano. El algoritmo se detiene ya sea cuando llega a nodos individuales o algún SSE

[10] Un dendrograma es una estructura en forma de árbol que visualiza el proceso de agrupación jerárquica.

mínimo. El particionamiento divisivo, que se puede observar en la Figura 6, permite una mayor flexibilidad tanto en términos de la estructura jerárquica del árbol como del nivel de equilibrio en los diferentes clústeres.

Figura 6

Clúster divisiva



Nota. Tomado de Algoritmo Agrupamiento Jerárquico-Teoría, por Aprende IA (2020).

4.2.2.C*lusters* no jerárquicos

Este enfoque busca dividir un conjunto de datos en grupos separados, sin superposición entre ellos. Los objetos se asignan a los clusters según su cercanía a un representante de cada grupo. La cantidad de clusters se define al inicio y, luego de varias iteraciones, se obtiene una agrupación considerada óptima. El algoritmo más conocido dentro de esta categoría es *k-means* (k-medias).

4.3. Clasificación

La clasificación es una subcategoría del aprendizaje supervisado cuyo objetivo es predecir las etiquetas de clase categórica de nuevas instancias, basadas en observaciones pasadas. Estas etiquetas de clase son discretas, valores desordenados que se pueden entender como membresías grupales de las instancias (Raschka y Mirjalili, 2019, p. 25).

4.3.1 Clasificación lineal

En la clasificación lineal, el clasificador calcula una combinación lineal de las características de entrada y define regiones de decisión separadas por hiperplanos en el espacio de características; si las clases son linealmente separables, este tipo de máquina resulta suficiente para resolver el problema de clasificación (Haykin, 2009).

Algunos de los clasificadores que utilizan funciones lineales para separar clases son los clasificadores lineales: regresión logística, máquina de vectores de soportes (kernel lineal), Naive Bayes y perceptrón.

4.3.1.1 Regresión logística

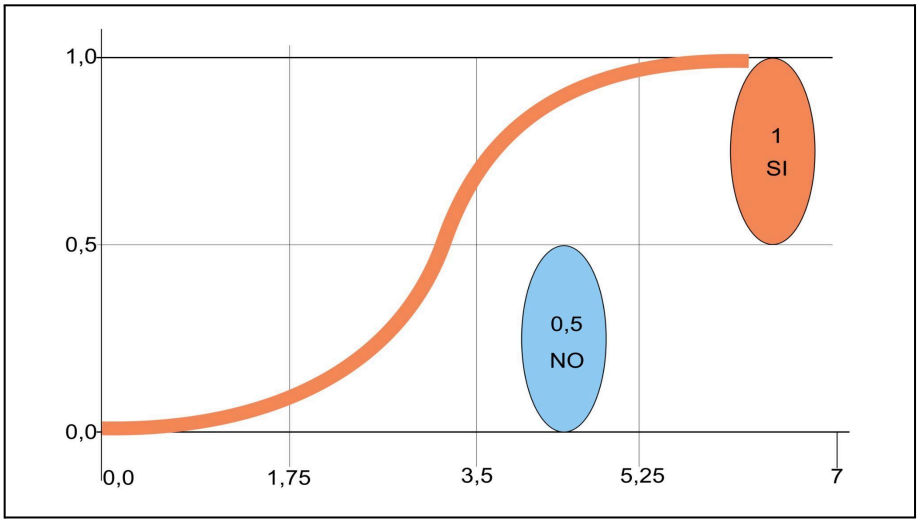
DataScientest (2025) explica que "la regresión logística es un modelo estadístico para estudiar las relaciones entre un conjunto de variables cualitativas Xi y una variable cualitativa Y. Se trata de un modelo lineal generalizado que utiliza una función logística como función de enlace".

Un modelo de regresión logística estima la probabilidad (p), de que un suceso ocurra (valor 1) o no ocurra (valor 0) ajustando sus coeficientes. Esa probabilidad siempre queda entre 0 y 1. Si el valor predicho supera un umbral definido, se considera que el evento es probable; si queda por debajo, se asume que no.

La regresión logística se basa en la función logística o sigmoide para obtener la probabilidad p (Ver Anexo B). Esta función es una curva en forma de S, esta se observa en la Figura 6, que puede tomar cualquier número de valor real entre 0 y 1 (DataScientest, 2025).

Figura 7

Regresión logística



Nota. Adaptado de Trabajo de Fin de Grado en Ingeniería Informática (p. 12), por M. Muñoz Ramos, 2025, Universidad Politécnica de Madrid.

Si la curva va a infinito positivo la predicción se convertirá en 1, y si la curva pasa el infinito negativo, la predicción se convertirá en 0. Si la salida de la función sigmoide es mayor que 0.5, se puede clasificar el resultado como 1 o SI, y si es menor que 0.5 se

puede clasificar como 0 o NO. Por su parte si el resultado es 0.75, se puede decir, en términos de probabilidad[11], hay un 75% de probabilidades de que el evento ocurra.

Como todos los análisis de regresión, la regresión logística es un análisis predictivo. Se usa para describir datos y explicar la relación entre una variable binaria dependiente y una o más variables independientes nominales, ordinales o de intervalo.

Por otro lado, Amazon Web Services (s.f.) señala que la regresión logística constituye una técnica clave dentro de la inteligencia artificial y el aprendizaje automático; además, describe a los modelos de *machine learning* como programas que pueden entrenarse para ejecutar tareas complejas de procesamiento de datos sin intervención humana directa.

Además, Amazon Web Services (s.f.) enumera diversas ventajas y desventajas asociadas al uso de este algoritmo, entre las cuales se destacan las siguientes ventajas:

- simplicidad: son matemáticamente menos complejos que otros métodos en *machine learning*.
- velocidad: pueden procesar grandes volúmenes de datos a alta velocidad porque requieren menos capacidad computacional, como memoria y potencia de procesamiento.
- flexibilidad: puede encontrar respuestas a preguntas que tienen dos o más resultados finitos. También puede usarlo para procesar datos.
- visibilidad: el análisis de regresión logística ofrece a los desarrolladores una mayor visibilidad de los procesos de software internos que otras técnicas de análisis de datos. La solución de problemas y la corrección de errores también son más fáciles porque los cálculos son menos complejos.

Por otro lado, es importante reconocer las desventajas, tales como:

- dificultad con datos no lineales: necesita técnicas adicionales (como transformación de variables o inclusión de términos polinomiales) para ajustarse correctamente a relaciones no lineales en los datos.
- sensibilidad a valores atípicos: la presencia de *outliers* puede afectar los resultados del modelo, generando estimaciones poco precisas o inestables.
- limitaciones en problemas complejos: por su simplicidad, puede no ser suficiente para capturar relaciones complejas o interacciones entre múltiples variables, lo que limita su capacidad predictiva en ciertos contextos.

[11] La probabilidad es una medida de la posibilidad de que ocurra un evento.

4.3.1.2 Máquinas de vectores de soporte

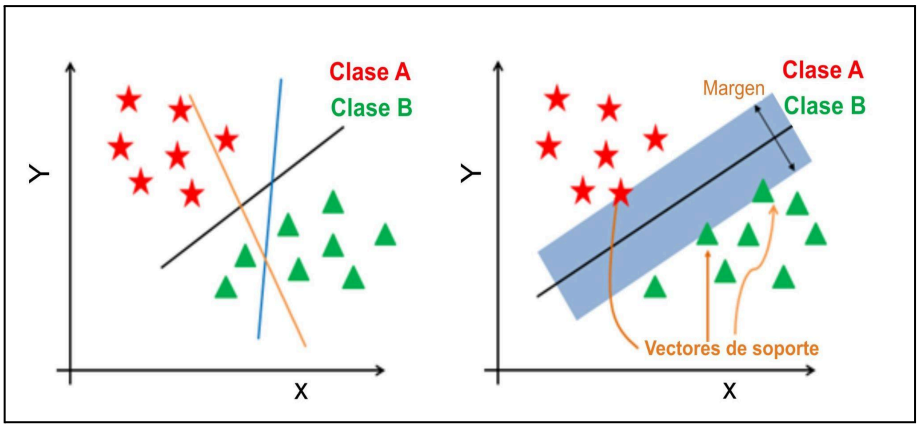
Según DataCamp (2024) menciona que “en general, las máquinas de vectores de soporte o SVM por sus siglas en inglés (*support vector machine algorithms*) se consideran un enfoque de clasificación, pero pueden emplearse en problemas de clasificación y de regresión”. Además, se indica que las SVM pueden trabajar con múltiples variables tanto continuas como categóricas y que, en un espacio de varias dimensiones, construye un hiperplano para separar las clases. De forma iterativa ajustan ese hiperplano para reducir el error y hallar el hiperplano de máximo margen, dicho de otro modo, el que mejor divide el conjunto de datos en clases (DataCamp, 2024).

Cómo funcionan las máquinas de vectores de soporte

DataCamp (2024) menciona que el objetivo principal es segregar el conjunto de datos dado de la mejor manera posible. La distancia entre los puntos más cercanos se conoce como margen. El objetivo es seleccionar un hiperplano[12] con el máximo margen posible entre vectores de soporte[13] en el conjunto de datos dado. SVM busca el hiperplano de máximo margen en los pasos siguientes.

Figura 8

Ilustración del hiperplano y los vectores de soporte en SVM



Nota. Adaptado de *Clasificación SVM con Scikit-Learn en Python*, por DataCamp (2023).

Generar hiperplanos que segregan las clases de la mejor manera. En la Figura 8, la gráfica de la izquierda muestra tres hiperplanos: negro, azul y naranja. Aquí, el azul y el naranja tienen mayor error de clasificación, pero el negro separa correctamente las dos clases. Selecciona el hiperplano con la máxima segregación de los puntos de datos más cercanos, como se muestra en la gráfica de la derecha.

[12] Es un plano de decisión que separa un conjunto de objetos que pertenecen a diferentes clases.

[13] Son los puntos de datos más próximos al hiperplano. Estos puntos definirán mejor la línea de separación calculando los márgenes.

Según DataCamp (2019), indica las ventajas y desventajas de este clasificador.

Las ventajas que se destacan son:

- buena exactitud: los clasificadores SVM suelen ofrecer una elevada precisión en las tareas de clasificación.
- uso eficiente de memoria: utilizan menos memoria, dado que en la fase de decisión emplean solo un subconjunto de los puntos de entrenamiento (vectores de soporte).
- buen desempeño con márgenes claros: funcionan adecuadamente cuando existe un margen de separación claro entre las clases.
- eficiencia en espacios de alta dimensión: resultan eficaces cuando se trabaja con un gran número de variables o características.

Por otro lado, posee desventajas, tales como:

- entrenamiento costoso en grandes volúmenes de datos: no son adecuadas para conjuntos de datos muy grandes debido a su elevado tiempo de entrenamiento.
- mal desempeño con clases solapadas: su rendimiento se ve afectado cuando las clases se superponen y no hay un margen de separación definido.
- sensibilidad al tipo de kernel: son modelos sensibles a la elección del kernel, por lo que una selección inadecuada puede deteriorar el desempeño del clasificador.

4.3.1.3. Clasificador Bayes ingenuo (*Naive Bayes*)

Según GeeksforGeeks (2025), Naive Bayes es un algoritmo de clasificación que utiliza la probabilidad para predecir la categoría a la que pertenece un punto de datos, asumiendo que todas las características no están relacionadas.

La idea principal del clasificador Naive Bayes es utilizar el teorema de Bayes para clasificar datos según las probabilidades de diferentes clases dadas sus características:

- es un clasificador probabilístico simple que cuenta con muy pocos parámetros, los cuales se utilizan para construir modelos de aprendizaje automático capaces de predecir a una velocidad mayor que otros algoritmos de clasificación.
- es un clasificador probabilístico porque asume que una característica del modelo es independiente de la existencia de otra, cada una contribuye a las predicciones sin que exista relación entre ellas.
- el algoritmo Naïve Bayes se utiliza en la filtración de spam, el análisis de sentimientos, la clasificación de artículos, entre otras.

El clasificador Bayes se denomina "ingenuo" porque asume que la presencia de una característica no afecta a las demás. El teorema de Bayes (Ver Anexo B) proporciona un método basado en principios para invertir las probabilidades condicionales.

IBM (s.f.) indica las siguientes ventajas y desventajas de este clasificador.

Las ventajas que se destacan son:

- menor complejidad: es un algoritmo sencillo con parámetros fáciles de estimar, por lo que suele ser uno de los primeros que se enseñan en cursos de ciencia de datos.
- buena escalabilidad: comparado con la regresión logística, resulta más rápido, eficiente y requiere poco almacenamiento.
- apto para datos de alta dimensión: es especialmente útil en contextos como la clasificación de documentos, donde se manejan muchas variables.

Por otro lado, posee algunas desventajas, tales como:

- frecuencia cero: es una variable categórica no aparece en el conjunto de entrenamiento, puede generar una probabilidad posterior igual a cero. Este problema se puede mitigar mediante técnicas como el suavizado de Laplace.
- hipótesis poco realista: la suposición de independencia condicional rara vez se cumple completamente en la práctica, lo que puede afectar negativamente la precisión del modelo.

4.4.2. Clasificación no lineal

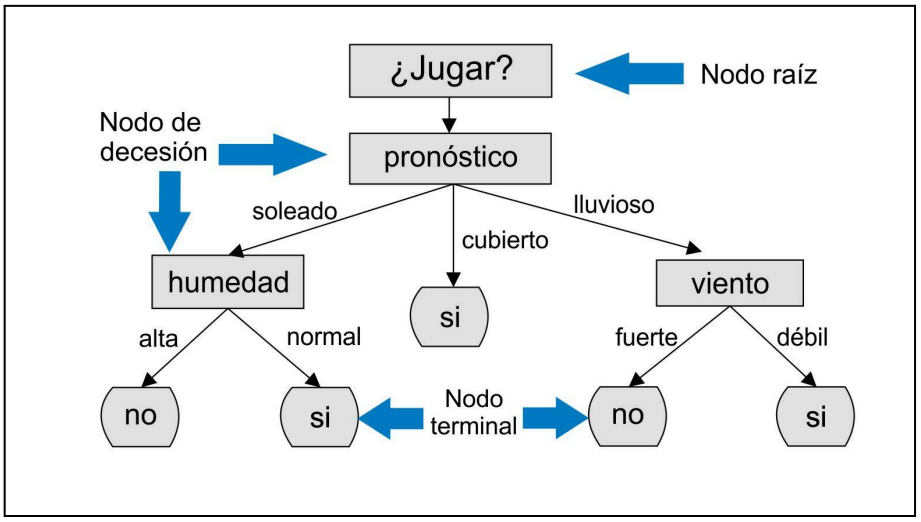
Según freeCodeCamp (2025), las tareas de clasificación no lineal presentan patrones que los modelos lineales no pueden detectar, debido a que las características mantienen entre sí relaciones no lineales.

4.4.2.1. Árbol de decisión

El árbol de decisión es un algoritmo de aprendizaje supervisado, categoría del aprendizaje automático, que utiliza conjuntos de datos etiquetados para entrenar modelos capaces de predecir resultados y reconocer patrones, tal como se describió en la sección 4.1. Se utiliza tanto para tareas de clasificación como de regresión. Se representa gráficamente en forma de árbol, como se observa en la Figura 9, donde cada nodo representa una decisión, una posible acción o un resultado. Las ramas del árbol indican las opciones disponibles y las hojas representan los resultados finales o las consecuencias de cada decisión (IBM, 2021).

Figura 9

Árbol de decisión



Nota. Adaptado de Introducción a la minería de datos (p. 30), por J. Hernández Orallo, M. J. Ramírez Quintana y C. Ferri Ramírez, 2004, Pearson Educación.

Un árbol de decisión es una técnica de modelado que se utiliza en el análisis de decisiones. Al proporcionar una estructura visual para analizar decisiones complejas, los árboles de decisión ayudan a las empresas a alcanzar sus objetivos de manera más efectiva y eficiente.

Según IBM (2021), sus ventajas son:

- fácil interpretación: fácil de entender y visualizar, esto útil para comunicar decisiones complejas a diferentes partes interesadas de una organización.
- modelado no lineal: puede modelar relaciones no lineales entre variables, por ello es adecuado para problemas con múltiples variables y relaciones complejas.
- no requiere asunciones de distribución: a diferencia de otros métodos estadísticos, no requieren asumir una distribución específica de los datos, lo que lo hace útil cuando los datos son difíciles de modelar.
- manejo de datos mixtos: pueden manejar eficazmente datos mixtos que incluyen tanto variables categóricas como numéricas, sin necesidad de preprocesamiento extenso.

Por otro lado, posee desventajas, tales como:

- sensibilidad a pequeños cambios: pequeños cambios en los datos de entrada, lo que puede llevar a diferentes árboles y decisiones finales.
- propensión al *overfitting*: existe el riesgo de que se ajuste demasiado a los datos de entrenamiento, lo que puede resultar en un rendimiento deficiente en datos nuevos y no vistos.

- dificultad con variables continuas: tiende a funcionar mejor con variables categóricas o discretas en lugar de variables continuas, lo que puede requerir discretización previa de los datos.
- inestabilidad: es inherentemente inestable, lo que significa que pequeños cambios en los datos de entrada pueden provocar cambios significativos en la estructura del árbol y en las decisiones resultantes.

4.4.2.2. Algoritmo de bosque aleatorio o *random forest*

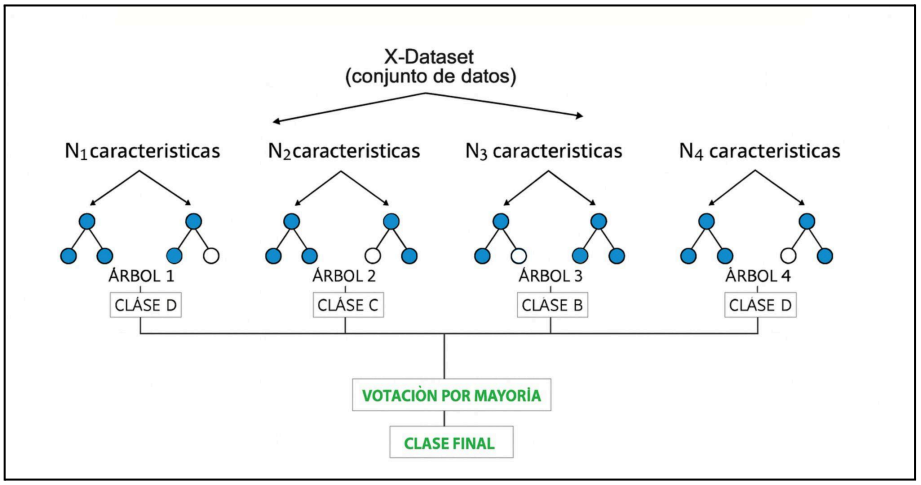
El algoritmo bosque aleatorio se describe como una técnica de aprendizaje automático ampliamente utilizada que combina múltiples árboles de decisión para producir un único resultado. Debido a su implementación sencilla y a su gran flexibilidad, se aplica tanto en problemas de clasificación como de regresión (Cardellino, 2021).

La idea central de este método consiste en construir varios árboles de decisión y luego integrar sus resultados, ya sea a través del voto mayoritario en clasificación o mediante el promedio en problemas de regresión. Esta estrategia permite reducir la varianza y el sesgo del modelo, además de mejorar la capacidad de generalización y la robustez de las predicciones (IBM, s.f.).

En la Figura 10 se observa un esquema de bosque aleatorio.

Figura 10

Bosque aleatorio



Nota. Adaptado de Metodología para la clasificación de cobertura del suelo en Argentina (p. 10), por E. Gaviola et al., 2023, Universidad Nacional del Comahue.

IBM (s.f.) indica las ventajas del algoritmo de bosque aleatorio:

- menor riesgo de sobreajuste: reduce significativamente este riesgo, debido a que, al construir un número considerable de árboles no correlacionados, el modelo logra disminuir la varianza general y el error de predicción mediante el promedio de los resultados obtenidos por cada árbol.
- flexibilidad en su aplicación: se destaca por su capacidad de adaptarse tanto a problemas de clasificación como de regresión, manteniendo un alto nivel de precisión. Además, el método de feature bagging que emplea permite al algoritmo estimar valores faltantes de manera eficaz, conservando la precisión incluso cuando parte de los datos no están disponibles.
- facilidad para determinar la importancia de las variables: puede evaluar la importancia de las distintas características o variables del modelo. Existen diversas métricas para ello, entre las que se destacan la importancia de Gini[14] y la disminución media de la impureza (MDI)[15], que permiten medir cuánto se afecta la precisión del modelo al eliminar una determinada variable.

Por otro lado, resulta relevante señalar las desventajas:

- procesamiento más lento: si bien es capaz de manejar grandes volúmenes de datos y proporcionar predicciones precisas, este proceso puede resultar más lento. Esto se debe a que el algoritmo debe realizar cálculos y generar predicciones a partir de múltiples árboles de decisión de manera individual.
- mayor demanda de recursos: el procesamiento de conjuntos de datos de gran tamaño requiere un consumo considerable de recursos, tanto en términos de almacenamiento como de capacidad de cómputo.
- complejidad en la interpretación: las predicciones generadas resultan más difíciles de interpretar, debido a la cantidad de árboles involucrados en la toma de decisiones.

4.4.2.3. K-vecinos más cercanos

Según IBM (s.f.), el algoritmo k vecinos más cercanos (KNN) “es un clasificador de aprendizaje supervisado no paramétrico[16] que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual”.

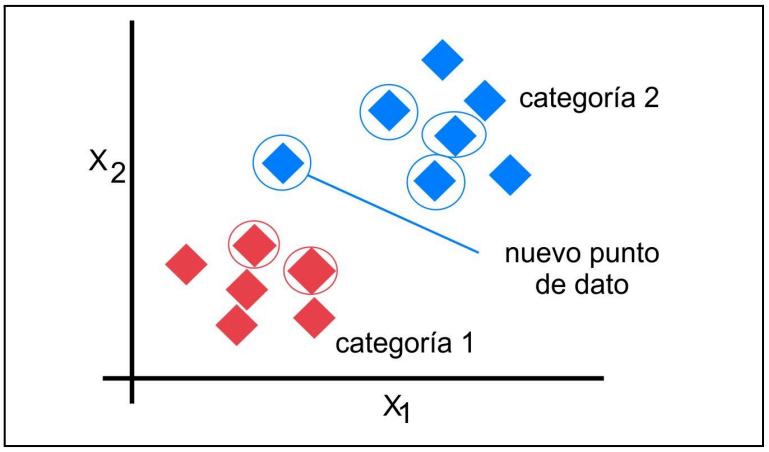
[14] Es una medida de cuán mezclados o impuros son los elementos.
[15] Es un método para medir la reducción de una impureza mediante el cálculo del coeficiente de Gini para cada división de características.
[16] El modelo se entrena a partir de ejemplos etiquetados sin imponer una forma funcional predeterminada ni un número fijo de parámetros

Además agrega que “el objetivo del algoritmo de vecino *k* más cercano es identificar los vecinos más cercanos de un punto de consulta determinado, de modo que podamos asignar una etiqueta de clase a ese punto”.

Por ejemplo, considerar la siguiente Figura 11 de puntos de datos que contienen dos características.

Figura 11

Clasificación de un nuevo punto con K-Vecinos Más Cercanos



Nota. Adaptado de “K-Nearest Neighbors (KNN) Algorithm”, por GeeksforGeeks (2025).

- el nuevo punto se clasifica como categoría 2 porque la mayoría de sus vecinos más cercanos son cuadrados azules. La Figura 11 muestra cómo KNN predice la categoría de un nuevo punto de datos basándose en sus vecinos más cercanos.
- KNN funciona utilizando la proximidad y la votación mayoritaria[17] para hacer predicciones.

En el algoritmo k-vecinos más cercanos, k es solo un número que le dice al algoritmo cuántos puntos o vecinos cercanos debe mirar cuando toma una decisión.

Métodos comunes para elegir k

- Validación cruzada: divide los datos en partes para probar distintos valores de k y elegir el que mejor rendimiento promedio tenga.
- Método del codo: gráfica la precisión o error frente a diferentes k; el "codo" de la curva indica un buen valor.
- Valores impares: usar números impares evita empates en clasificación.

[17] La mayoría de votos,se utiliza la etiqueta que se representa con más frecuencia en torno a un punto de datos determinado.

Métricas de distancia utilizadas en el algoritmo KNN

Según IBM (s.f.), el objetivo de *KNN* es, para un punto de consulta dado, encontrar sus vecinos más próximos y con esa información asignarle una etiqueta de clase. Para determinar cuáles son esos vecinos, el algoritmo compara distancias con alguna de las métricas de distancias habituales son:

- distancia euclidiana
- distancia de Manhattan
- distancia de Minkowski

Las fórmulas matemáticas de estas distancias se detallan en el Anexo B.

IBM (s.f.) indica las siguientes ventajas del algoritmo:

- sencillez de implementación: es fácil de entender y aplicar, por lo que suele ser uno de los primeros algoritmos que se enseñan en ciencia de datos.
- adaptabilidad a nuevos datos: cómo almacena todos los datos de entrenamiento en memoria, puede ajustarse fácilmente a nuevas muestras sin necesidad de reentrenar el modelo.
- pocos hiperparámetros: sólo necesita definir el número de vecinos k y una métrica de distancia, lo que simplifica el proceso de configuración en comparación con otros algoritmos más complejos.

Tiene algunas desventajas, tales como:

- baja escalabilidad: al ser un algoritmo de aprendizaje perezoso *(lazy learning)*, requiere mucho almacenamiento y procesamiento, lo que lo hace lento y costoso en grandes volúmenes de datos.
- maldición de la dimensionalidad: su rendimiento disminuye cuando se trabaja con datos que tienen muchas características, ya que la distancia entre puntos pierde significado y aumenta la tasa de error.
- riesgo de sobreajuste o subajuste: si el valor de k es demasiado bajo, el modelo puede sobreajustarse a los datos; si es muy alto, puede subestimar patrones importantes, afectando la precisión.

4.4.2.4. Redes neuronales

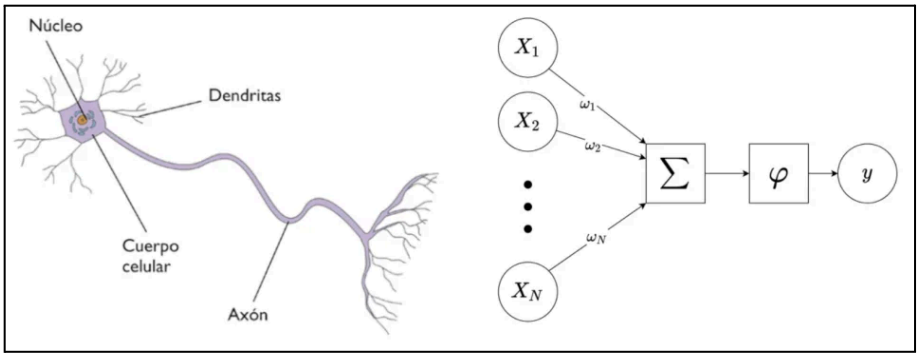
Una red neuronal artificial (RNA) es un modelo de aprendizaje automático[18] que imita el funcionamiento del cerebro humano mediante neuronas artificiales interconectadas; el perceptrón multicapa (MLP) es un tipo de RNA con varias capas y funciones de activación no lineales, capaz de aprender patrones complejos y realizar tareas como clasificación, regresión y reconocimiento de patrones. (DataCamp, s.f.)

[18] El aprendizaje automático es un subcampo de la inteligencia artificial que permite a los sistemas aprender de los datos y mejorar con la experiencia, sin necesidad de programación explícita.

Cada neurona artificial, denominada perceptrón simple, tiene su relación con las partes de una neurona biológica, como se observa en la Figura 12.

Figura 12

Neurona biológica vs Red neuronal artificial

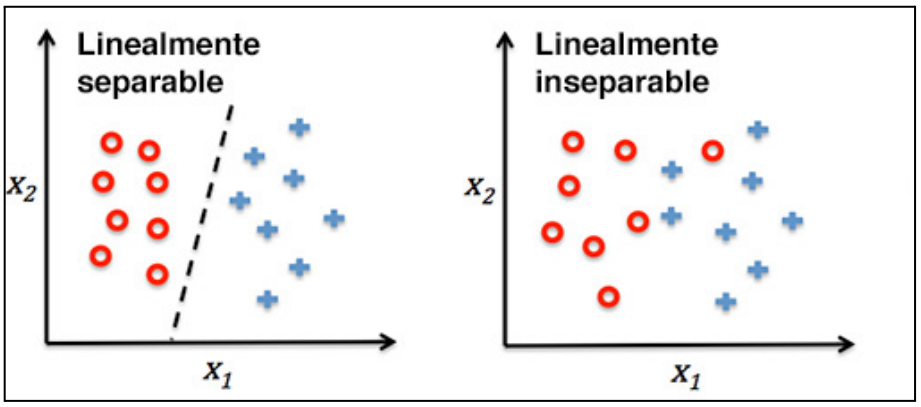


Nota. Adaptado de Perceptrón Simple: Definición matemática y propiedades, por Damavis (2020).

- modelo de perceptrón simple: es un tipo de red neuronal perteneciente a las arquitecturas más básicas, y se utiliza para realizar tareas de clasificación binaria, asignando características de entrada a una decisión de salida, generalmente clasificando los datos en una de dos categorías, como 0 o 1. Este modelo consta de una sola capa de nodos de entrada completamente conectados a una capa de nodos de salida, y resulta especialmente eficaz en el aprendizaje de patrones linealmente separables. En la Figura 13 se observa la comparación entre linealmente separable y linealmente inseparable.

Figura 13

Linealmente separable vs linealmente inseparable



Nota. Adaptado de *Python Machine Learning* (p. 23), por S. Raschka y V. Mirjalili, 2017, Packt Publishing.

En este contexto, Haykin (2009) distingue dos tipos de perceptrones:

- el perceptrón de una sola capa: se caracteriza por su capacidad para resolver únicamente problemas linealmente separables. Sin embargo, presenta limitaciones importantes, ya que fracasa al enfrentar problemas no linealmente separables, como el clásico ejemplo del XOR[19]. Además, su salida resulta insuficiente para representar relaciones más complejas entre las variables de entrada.
- el perceptrón multicapa (MLP): es una extensión del perceptrón simple que permite superar su principal limitación: la incapacidad de clasificar patrones no linealmente separables. Está compuesto por una capa de entrada, una o más capas ocultas y una capa de salida, en las que cada neurona combina linealmente las entradas y aplica una función de activación no lineal diferenciable, lo que habilita a la red a modelar relaciones complejas.

En la Figura 14 se presenta el esquema de un perceptrón. Este modelo se compone de elementos fundamentales que, en conjunto, permiten procesar la información y producir predicciones.

A continuación, se describen sus partes principales:

- características de entrada: el modelo recibe múltiples entradas representadas por X_0, X_1, X_2, …, X_n, donde cada una corresponde a una característica específica de los datos de entrada.
- pesos: a cada entrada se le asigna un peso W_0, W_1, W_2, …, W_n, que determina su influencia sobre la salida. Estos pesos se ajustan durante el entrenamiento para optimizar el rendimiento del modelo.
- sesgo (bias): el sesgo se representa con la entrada X_0, que siempre vale 1, y su peso asociado W_0. Su función es permitir que el perceptrón ajuste la salida incluso cuando todas las demás entradas son cero, mejorando así la flexibilidad del modelo.
- función de suma: todas las entradas ponderadas (multiplicadas por sus respectivos pesos) se combinan en el nodo de suma ponderada (Σ), produciendo una salida intermedia z. Este valor se calcula como:

$$z = W_0 \cdot X_0 + W_1 \cdot X_1 + W_2 \cdot X_2 + \cdots + W_n \cdot X_n$$

- función de activación: el valor z se pasa por una función de activación, toma un único número y realiza una determinada operación matemática fija sobre él.

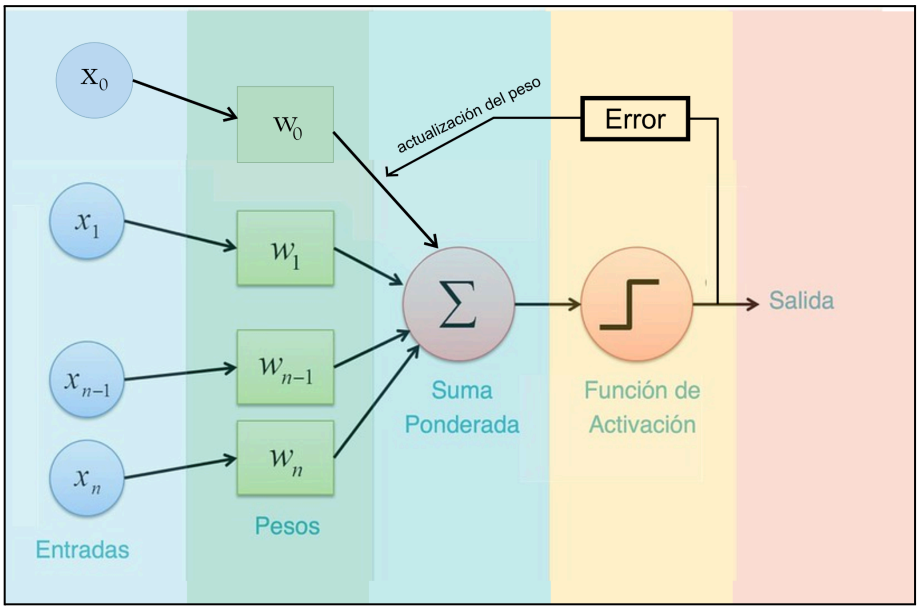
[19] XOR, que significa "OR exclusivo", es una operación lógica binaria que devuelve un resultado verdadero (1) si sus entradas son diferentes, y falso (0) si son iguales.

Hay varias funciones de activación que se pueden encontrar en la práctica, las más comunes son la sigmoide o la ReLU o unidad lineal rectificada.

- salida: es el resultado final del perceptrón, indicado a la derecha del esquema.
- algoritmo de aprendizaje: si la salida generada difiere de la salida esperada, se calcula un error, el cual es utilizado para actualizar los pesos (incluido el sesgo) a través de un algoritmo de aprendizaje, como la regla del perceptrón, con el objetivo de minimizar futuros errores.

Figura 14

Esquema del perceptrón artificial



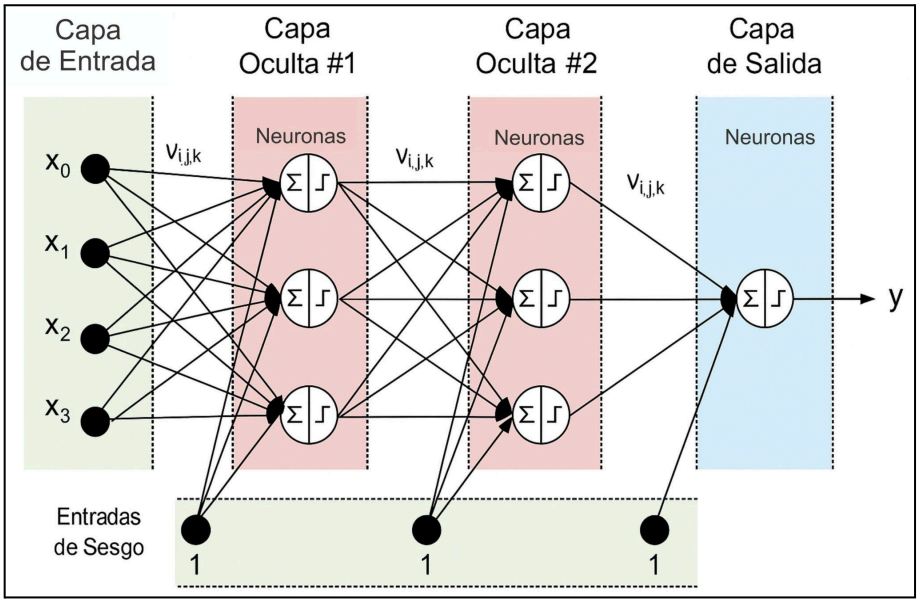
Nota. Adaptado de Qué son las redes neuronales artificiales, por Aprende IA (2021).

En la Figura 15 se presenta el esquema de un perceptrón multicapa (MLP), que es una red neuronal de tipo *feedforward*[20] que extiende la arquitectura del perceptrón simple mediante la inclusión de una o más capas ocultas entre la capa de entrada y la de salida. Cada neurona combina de forma lineal sus entradas y aplica una función de activación no lineal diferenciable, lo que permite que el modelo aprenda relaciones complejas entre las variables de entrada y salida. A diferencia del perceptrón de una sola capa, restringido a patrones linealmente separables, el MLP posee la capacidad de resolver problemas no lineales al transformar los datos en nuevos espacios de representación. De este modo, el procesamiento de la información ocurre de manera secuencial a través de varias capas, hasta producir la salida final (Haykin, 2009, p. 123–127).

[20] *Feedforward* se refiere a un tipo de arquitectura donde la información fluye únicamente en una sola dirección, desde la capa de entrada, pasando por las capas ocultas, hasta la capa de salida, sin existir bucles ni retroalimentación.

Figura 15

Modelo de perceptrón multicapa



Nota. Adaptado de “Mastering the Training of Neural Networks: Key Points to Achieve Success”, por P. Gupta, 2023.

El MLP se entrena utilizando el algoritmo de retropropagación, que ajusta los pesos de la red mediante un proceso de aprendizaje supervisado. Este proceso se desarrolla en dos etapas:

- etapa hacia adelante, las funciones de activación se originan desde la capa de entrada a la capa de salida.
- etapa hacia atrás: el error, calculado como la diferencia entre el valor observado y el valor esperado, se propaga desde la capa de salida hacia las capas anteriores, ajustando los pesos para minimizar dicho error.

El perceptrón multicapa puede definirse como una red compuesta por numerosas neuronas artificiales organizadas en distintas capas. En esta arquitectura, la función de activación deja de ser lineal y, en su lugar, se emplean funciones de activación no lineales, como las funciones sigmoide, tangente hiperbólica (tanh) y ReLU, entre otras, que permiten dotar al modelo de una mayor capacidad de representación.

4.5. Evaluación de modelos

En esta sección se describen los criterios y procedimientos utilizados para evaluar el desempeño de los modelos de clasificación aplicados al riesgo de deserción.

4.5.1 Matriz de confusión

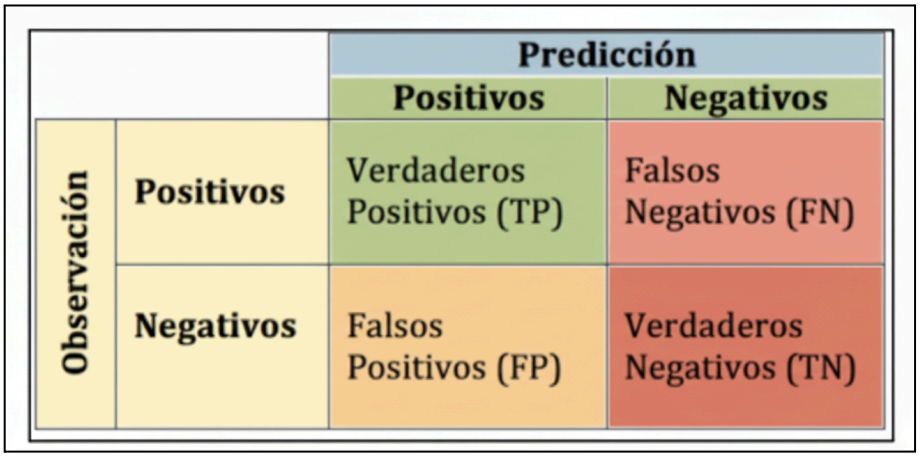
La matriz de confusión es una herramienta que permite evaluar el rendimiento de un modelo de clasificación en *machine learning* mediante la comparación entre los valores predichos y los valores reales de un conjunto de datos. Según IBM (s.f.), puede aplicarse al análisis del desempeño de diversos algoritmos clasificadores, entre ellos Naïve Bayes, regresión logística y árboles de decisión, cuyas definiciones se desarrollaron a lo largo de este capítulo.

En la Figura 16 se muestra la matriz correspondiente a este estudio, donde la clase positiva se define como valor 1. Se adopta la convención: filas = clase real y columnas = clase predicha.

- TP (verdaderos positivos): el número de predicciones correctas para la clase positiva. Ej: desertores correctamente identificados como en riesgo.
- FP (falsos positivos): aquellos casos de clase negativa identificados incorrectamente como casos positivos. Ej: no desertores marcados como desertores.Este tipo de equivocación se conoce como error tipo I (falso positivo).
- FN (falsos negativos): los casos positivos reales que se predijeron erróneamente como negativos. Ej:desertores que el modelo no detectó.Este tipo de equivocación se conoce como error tipo II (falso negativo).
- TN (verdaderos negativos): que son las instancias de clase negativa reales que se predijeron con precisión como negativas. Ej: no desertores correctamente identificados.

Figura 16

Matriz de confusión



Nota. Adaptado de Sistema predictivo de probabilidad de compra en e-commerce (p. 19), por R. Barba Alonso, 2022, Universidad de Valladolid.

4.5.2. Métricas de evaluación

Las métricas se obtienen a partir de la matriz de confusión mencionada con anterioridad en la Figura 16.

- Exactitud (*Accuracy*): es la proporción de predicciones correctas sobre el total de casos evaluados. Indica qué tan bien clasifica el modelo en general, pero puede ser engañosa cuando las clases están desbalanceadas.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

IBM (s.f.) presenta el siguiente ejemplo sintetizado en la Tabla 16: en un conjunto de 100 instancias, el clasificador obtiene 4 TP (verdaderos positivos), 95 TN (verdaderos negativos), 1 FN (falso negativo) y ningún FP (falso positivo). La exactitud se calcula como (TP + TN)/100 = (4 + 95)/100 = 99 %. Este valor tan elevado ilustra cómo una alta exactitud puede resultar engañosa, ya que oculta el costo asociado a los falsos negativos.

Tabla 5

Ejemplo de matriz de confusión con 100 instancias (adaptado de IBM)

	Predicha(1) (positivo)	**Predicha(0) (negativo)**
Real=1	TP: 4	FN: 1
Real = 0	FP: 0	TN: 95

Una exactitud elevada, no garantiza buen desempeño cuando los falsos negativos son críticos. Por ejemplo, detectar una enfermedad contagiosa. Ese 1% puede tener un alto costo. Por lo tanto, se pueden utilizar otras métricas de evaluación para proporcionar un mejor rendimiento del algoritmo de clasificación.

- Precisión (*Precision*): es la proporción de verdaderos positivos entre todos los casos que el modelo predijo como positivos.

$$Precisión = \frac{TP}{TP+FP}$$

Google Developers (2025), explica el ejemplo de clasificación de *spam*, la *precision* mide la fracción de correos electrónicos clasificados como spam que realmente lo eran. No obstante, en conjuntos desequilibrados con muy pocos positivos reales. Por ejemplo, de 1-2 casos, la *precision* aislada resulta menos informativa, ya que puede inflarse prediciendo muy pocos positivos y no refleja cuántos positivos reales se omiten.

La *precision* mejora a medida que disminuyen los FP (falsos positivos), mientras que el *recall* mejora cuando disminuyen los FN (falsos negativos). Como resultado, la *precision* y el *recall* suelen mostrar una relación inversa, en la que mejorar uno de ellos empeora el otro.

- Sensibilidad (*Recall*): es la proporción de casos positivos correctamente identificados por el modelo.

$$Recall = \frac{TP}{TP+FN}$$

Towards Data Science (2020), explica que en problemas como la detección de fraude con tarjetas de crédito, el objetivo central es no pasar por alto ninguna operación fraudulenta, por lo que se busca minimizar los falsos negativos.

En este tipo de escenarios puede aceptarse una precisión relativamente baja, siempre que el *recall* sea alto. De manera análoga, en el ámbito médico se prioriza no dejar sin detectar a ningún paciente que presente la condición analizada, por lo que se pone el foco en maximizar el *recall*.

En nuestro caso de deserción universitaria, un FN (falso negativo), implica clasificar como “no en riesgo” a un estudiante que en realidad abandonará, por lo que también puede ser preferible privilegiar un *recall* elevado, aún a costa de una menor precisión.

- *F1-Score*: es la media armónica entre precisión y sensibilidad. Equilibra ambas métricas, siendo útil cuando hay desbalance de clases y se busca un balance entre detectar positivos y evitar falsos positivos.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

IBM (s.f.), menciona que la puntuación F1 es especialmente útil en conjuntos de datos desequilibrados, porque equilibra *precision* y *recall*.

Por ejemplo, en un clasificador de una enfermedad rara, un modelo que predice que nadie está enfermo puede tener precisión alta pero *recall* nulo, mientras que otro que predice que todos están enfermos logra *recall* perfecto pero precisión muy baja.

F1-Score combina ambos valores y ofrece una visión más equilibrada del rendimiento del clasificador.

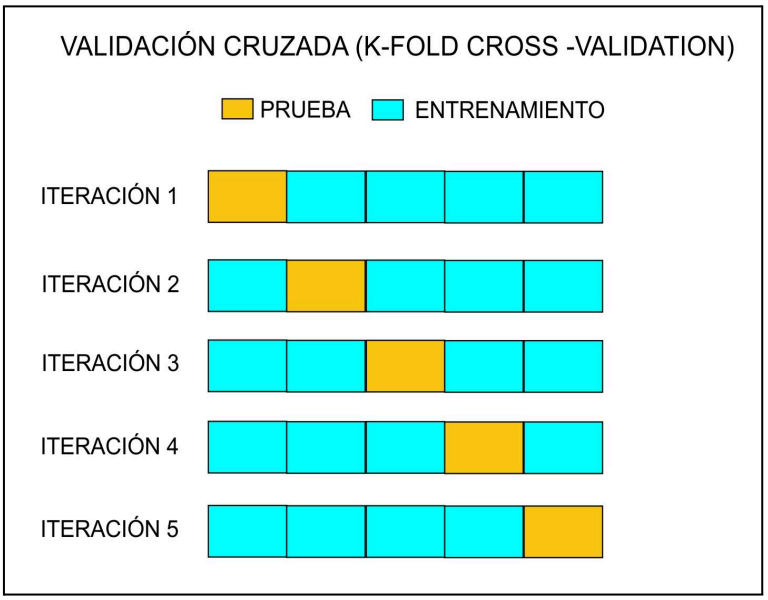
4.5.3. Validación cruzada k-fold

DataCamp (2024) señala que la validación cruzada *k*-fold es “una técnica robusta para evaluar el rendimiento de los modelos de aprendizaje automático y garantizar que generalicen bien”. Por su parte, Codificando Bits (s.f.) resume la idea de manera sencilla: “en lugar de usar diferentes sets para entrenamiento y validación, en la validación cruzada usaremos todos los datos”.

El algoritmo comienza con la mezcla aleatoria de los datos y con la inicialización del parámetro *k* que es simplemente un número entero que definirá el número de particiones de nuestro set de datos así como el número de iteraciones de entrenamiento y validación que se usarán al construir el modelo. Esto se observa en la Figura 17.

Figura 17

Validación cruzada k-fold



Nota. Adaptado de Cross-Validation en Python, por Deepnote (s.f.).

Según Codificando Bits (s.f.), una vez definido el valor de *k* inicia este proceso de entrenamiento y validación. Así que por un total de *k* iteraciones se repiten estos pasos:

- se toma una de las *k* particiones y se mantiene “oculta” al modelo
- se toman las *k-1* particiones restantes y con ellas se entrena el modelo. Se calculan de forma automática los parámetros, a través del algoritmo de entrenamiento.
- una vez entrenado el modelo se almacena su desempeño para con el set de entrenamiento (*k-1* particiones) y con el set oculto.
- luego se repiten los tres pasos anteriores pero cambiando la partición que se mantiene oculta.

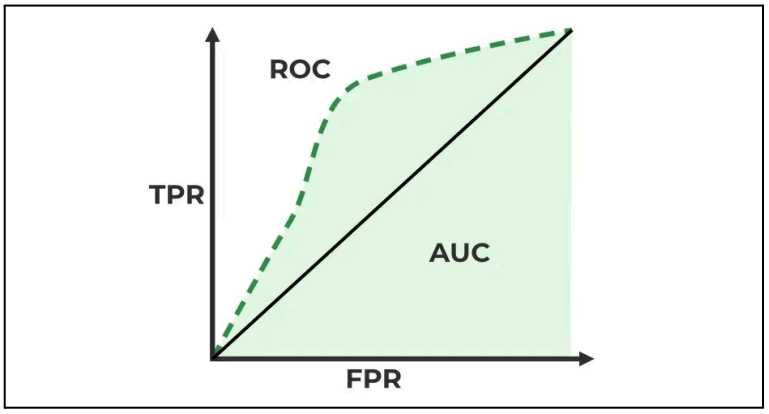
Y una vez terminadas las iteraciones tendremos *k* medidas de desempeño para los sets de entrenamiento y validación usados en cada iteración. Así que el desempeño final del modelo será simplemente el promedio de los desempeños anteriores.

4.5.4. Curva ROC

La curva ROC ofrece una representación visual, la Figura 18 muestra un ejemplo de esta, de las compensaciones entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) con distintos umbrales[21]. Proporciona información sobre lo bien que el modelo puede equilibrar las compensaciones entre detectar instancias positivas y evitar falsos positivos a través de diferentes umbrales (DataCamp, s.f.).

Figura 18

Métrica de evaluación de la clasificación AUC-ROC



Nota. Adaptado de AUC-ROC Curve in Machine Learning, por GeeksforGeeks (2025)

- Tasa de verdaderos positivos (TPR): refleja la capacidad de un modelo para identificar correctamente los casos positivos. Mide la proporción de casos positivos reales que el modelo identifica con éxito.

$$TPR = \frac{TP}{TP+FN}$$

Dónde:

- ➔ Los verdaderos positivos (TP) son el número de registros de clase positivos que el modelo predice correctamente como positivos.
- ➔ Los falsos negativos (FN) son el número de registros de clase positivos que el modelo predice incorrectamente como negativos.

[21] El umbral es un valor de corte a partir del cual decidís si esa probabilidad la considerás clase positiva o negativa.

- Tasa de falsos positivos (FPR): representa la frecuencia con la que nuestro modelo clasifica incorrectamente como positivas las instancias de clase negativas. Mide la proporción de instancias negativas reales que el modelo identifica incorrectamente como positivas, lo que indica la tasa de falsas alarmas.

$$FPR = \frac{FP}{FP+TN}$$

Dónde:

➔ FP o falsos positivos son el número de registros de clase negativos que se predicen incorrectamente como positivos.

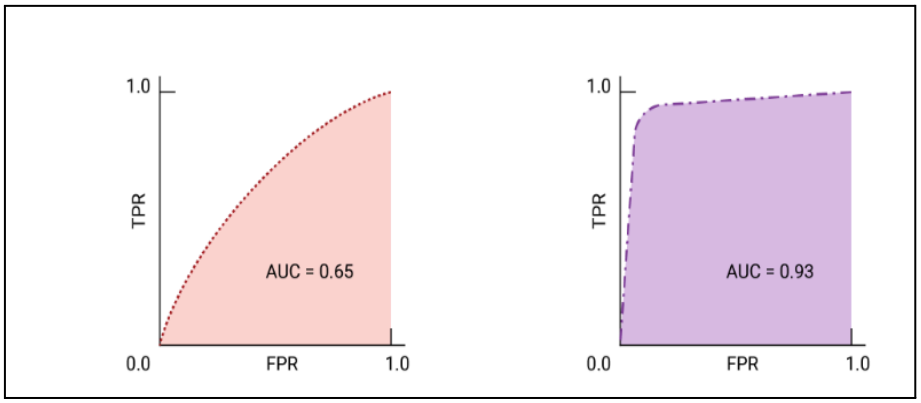
➔ TN o verdaderos negativos son el número de registros de clases negativas que se predicen correctamente como negativos.

El AUC (área bajo la curva) es un valor escalar único que va de 0 a 1, y que da una instantánea del rendimiento del modelo. Sólo se calcula el AUC después de generar la curva ROC, puesto que el AUC corresponde al área bajo dicha curva (DataCamp, s.f.).

La AUC es una medida útil para comparar el rendimiento de dos modelos diferentes, siempre que el conjunto de datos esté aproximadamente equilibrado. Por lo general, el modelo con mayor área debajo de la curva (más cercano a 1), es el mejor (Google Developers, s.f.). En la Figura 19 se observa la comparativa de rendimiento entre dos modelos utilizando la AUC. La curva situada a la derecha, con un AUC mayor, corresponde al modelo con mejor rendimiento.

Figura 19

Comparación de curvas ROC y AUC entre dos modelos hipotéticos.



Nota. Adaptado de Clasificación: ROC y AUC del Machine Learning Crash Course (Google Developers, s.f.)

El $AUC - ROC$ resume el rendimiento global del modelo considerando todos los umbrales posibles, a diferencia de métricas como la exactitud, la precisión o el $F1$, que dependen de un único punto de corte. Además a ser es un valor escalar único que facilita la comparación de varios modelos, independientemente de sus umbrales de clasificación.

Su naturaleza independiente del umbral hace que sea la mejor opción para establecer una comparación justa entre modelos con diferentes umbrales óptimos (DataCamp, s.f.).

Capítulo 5: Resultados del estudio y análisis de los modelos

Este capítulo presenta los resultados obtenidos en la fase de modelado, siguiendo la metodología *CRISP-DM* expuesta en la sección 3.6. Se describe el proceso realizado, desde la preparación del conjunto de datos hasta la aplicación de distintas técnicas de aprendizaje supervisado, cuyo marco conceptual se desarrolló en la sección 4.1. Asimismo, se evalúa el desempeño de los modelos mediante métricas comparativas y visualizaciones, definidos en la sección 4.5.2. Una vez obtenidos los desempeños, se realizó el análisis de las variables más influyentes para cada uno.

5.1. Comprensión del negocio

En la fase de comprensión del negocio se precisó el objetivo central del estudio: identificar, mediante técnicas de minería de datos, los patrones y factores asociados a la deserción de los estudiantes de las carreras de Analista de Sistemas y Licenciatura en Sistemas de Información de la Facultad de Ciencia y Tecnología (UADER), con el propósito de apoyar la toma de decisiones académicas y de gestión orientadas a su reducción.

A partir de esta definición se establecieron dos puntos importantes:

- seleccionar modelos que permitan estimar la probabilidad de deserción de cada estudiante.
- priorizar modelos con buen *recall*, que logren detectar la mayor cantidad posible de estudiantes en riesgo.

5.2. Comprensión de los datos

En esta fase se analizó la información disponible en la base de datos del SIU GUARANÍ[22] de la facultad. El conjunto de datos incluye registros de alrededor de diez años (2013-2023) y reúne información de aproximadamente 1.600 estudiantes de las dos carreras mencionadas.

De forma general, las variables pueden agruparse en tres bloques:

- académicas: año de ingreso, porcentaje de avance en la carrera, cantidad de materias aprobadas, promedio con y sin aplazos, años de inactividad, entre otras.
- sociodemográficas: lugar de origen, lugar de residencia durante el cursado, tipo de vivienda.
- económicas y laborales: si el estudiante trabaja, cantidad de horas de trabajo semanales, forma de costeo de los estudios, presencia de becas, etc.

[22] SIU Guaraní es un sistema de gestión académica en línea que se utiliza en universidades argentinas para administrar todas las actividades de los estudiantes, desde la inscripción hasta la obtención del título

Para más detalles, la descripción completa de las variables utilizadas y sus definiciones operativas se presentan en las Tablas 2, 3 y 4 de la sección 3.5.

Estas primeras búsquedas permitieron detectar valores faltantes, posibles errores de carga y categorías con muy pocos casos, aspectos que se tuvieron en cuenta en la fase posterior de preparación de los datos.

5.3. Preparación de los datos

La fase de preparación de los datos tiene como objetivo dejar la base de datos en condiciones para poder aplicar los modelos de minería de datos. Para realizar estas tareas se utilizó el lenguaje de programación Python, empleando principalmente las librerías de código abierto Pandas y NumPy para la manipulación del conjunto de datos, y las herramientas de Scikit-learn para la imputación de valores, el reescalado de variables y la codificación de atributos categóricos.

Los pasos principales fueron:

- selección de variables: se eliminaron campos que no aportan información relevante al problema y se conservaron aquellas variables con mayor relación potencial con la deserción estudiantil. Por ejemplo, se contaban con variables que eran comunes para todos los alumnos (como el plan de estudio) o que no aportan información relevante (como la ID del alumno o la carrera de egreso); por lo tanto, estas variables fueron eliminadas.
- definición de la variable objetivo[23]: se creó un atributo llamado deserto con dos valores posibles:
    - 0: estudiante no desertor.
    - 1: estudiante considerado desertor.

      Se consideró que un alumno desertó, si cumple estas condiciones:
        - menos del 20% de avance en la carrera.
        - menos del 25% de materias aprobadas.
        - 2 años o más sin reinscribirse.

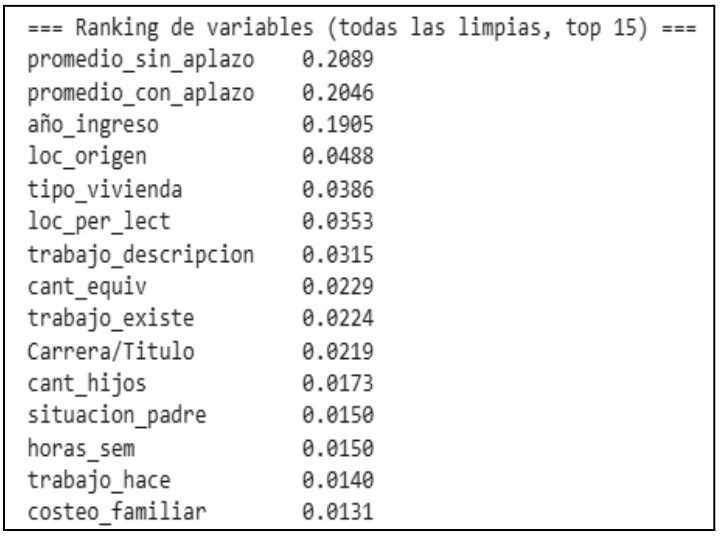
      La categoría “alumno desertó” se define, por tanto, de manera operativa a partir de estos tres umbrales, establecidos por los autores en función de las trayectorias típicas observadas en las carreras de Analista en Sistemas y Licenciatura en Sistemas de Información de la FCyT-UADER, y se apoyan en las perspectivas teóricas sobre cuantificación de la deserción y del rendimiento académico desarrolladas en el sección 2.2.

[23] La variable objetivo es la variable de interés principal en un estudio o análisis, es el resultado que se busca predecir o explicar, y suele ser una variable dependiente.

- tratamiento de valores faltantes: en las variables categóricas, los valores faltantes se imputaron utilizando la estrategia más frecuente o moda mencionado en la sección 3.7.4 y en las variables numéricas se aplicaron estrategias sencillas de imputación, procurando no distorsionar la distribución original de los datos. Por ejemplo, en la variable *tipo de vivienda* se completaron los registros faltantes asignando la categoría más frecuente observada en la base.
- variables categóricas: se transformaron en variables binarias mediante técnicas de codificación (como la codificación *one-hot*), de modo que pudieran ser utilizadas por los algoritmos de modelado. Por ejemplo, en la variable *sexo* los valores originales, registrados como texto, se recodificaron en forma binaria (1 para masculino y 0 para femenino).
- variables numéricas: se reescalaron cuando fue necesario, para que quedaran en rangos comparables entre sí. Por ejemplo, se normalizaron el porcentaje de avance en la carrera (originalmente expresado entre 0 y 100) y la cantidad de materias aprobadas, de manera que ambas quedaran en una escala similar antes de ingresar a los modelos de clasificación.
- variables más importantes: los modelos suelen desempeñarse mejor cuando trabajan con un conjunto reducido y relevante de variables, tal como se mencionó en la sección 3.5; por ello, la selección de las características más significativas resulta fundamental. En este sentido, se empleó un bosque aleatorio como herramienta para obtener un ranking de importancia de las variables a través de su clasificador. En la Figura 20 se presenta el listado de las 15 variables más relevantes. Una de las ventajas del bosque aleatorio, expuesta en la sección 4.4.2.2, es su facilidad para estimar la importancia de las variables mediante métricas como la importancia de Gini y la disminución media de la impureza (MDI), que permiten cuantificar en qué medida se ve afectado el rendimiento del modelo al eliminar cada característica.

Figura 20

Ranking de importancia de las principales variables



```
=== Ranking de variables (todas las limpias, top 15) ===
promedio_sin_aplazo    0.2089
promedio_con_aplazo    0.2046
año_ingreso            0.1905
loc_origen             0.0488
tipo_vivienda          0.0386
loc_per_lect           0.0353
trabajo_descripcion    0.0315
cant_equiv             0.0229
trabajo_existe         0.0224
Carrera/Titulo         0.0219
cant_hijos             0.0173
situacion_padre        0.0150
horas_sem              0.0150
trabajo_hace           0.0140
costeo_familiar        0.0131
```

Nota. Elaboración propia a partir de los resultados obtenidos mediante la ejecución del modelo en Google Colab.

Finalmente, se generó un conjunto de datos depurado y transformado, listo para ser utilizado en la fase de modelado.

5.4. Modelado

En la fase de modelado se aplicaron distintas técnicas de aprendizaje supervisado que se desarrollaron en el capítulo 4 para predecir la deserción estudiantil. El objetivo fue comparar modelos de aprendizaje supervisado para contrastar el desempeño predictivo de diversos algoritmos y optimizar la selección del modelo.

En esta fase se trabajó con cinco modelos de aprendizaje supervisado, cuya fundamentación teórica se presentó en las secciones 4.3.1.1 a 4.4.2.4: regresión logística, máquinas de vectores de soporte (SVM), árbol de decisión, bosque aleatorio y un perceptrón multicapa (red neuronal simple).

En la práctica, estos algoritmos se implementaron en Python utilizando la biblioteca scikit-learn. Las clases utilizadas son:

- LogisticRegression
- RandomForestClassifier
- LogisticRegression
- SVC (Support Vector Classifier)
- Red neuronal

Los modelos se entrenaron utilizando validación cruzada estratificada con 5 *folds*[24] o particiones, dividiendo el conjunto de datos en cinco particiones distintas. Esto permitió evaluar el rendimiento de cada algoritmo en diferentes subconjuntos y evitar depender de una única división entrenamiento–prueba, reduciendo el riesgo de obtener resultados sesgados por una partición particular. Cabe recordar que los conceptos de validación cruzada se desarrolló previamente en la sección 4.5.3.

En todos los casos se utilizó el mismo conjunto de variables predictoras, de manera que las diferencias de desempeño se atribuyan a las características propias de cada algoritmo y no a cambios en los datos de entrada. Esta estrategia facilita la comparación entre modelos y la selección de aquellos que resultan más adecuados para el problema de predicción de la deserción.

5.5. Evaluación de los modelos

La evaluación de los modelos predictivos constituye una etapa esencial dentro de la metodología *CRISP-DM*, ya que permite estimar su capacidad de generalización y su utilidad práctica como herramienta de apoyo a la toma de decisiones institucionales.

El desempeño se presenta con las métricas explicadas en la sección 4.5.2: exactitud*, precisión, recall*, *F1-score* y AUC-ROC.

El *recall* es prioritario en contextos de retención, porque penaliza los falsos negativos (estudiantes en riesgo no detectados). Los valores obtenidos se encuentran en la Tabla 10.

Tabla 6

Resultados de desempeño de los algoritmos aplicados

Modelo	Accuracy	Precisión	Recall	F1-Score	Auc-Roc
Árbol de decisión	0.821	0.723	0.807	0.763	0.894
Bosque aleatorio	0.806	0.762	0.718	0.711	0.873
Regresión logística	0.781	0.740	0.680	0.673	0.873
Máquinas de vectores de soporte	0.806	0.870	0.536	0.664	0.866
Perceptrón multicapa	0.788	0.681	0.772	0.722	0.887

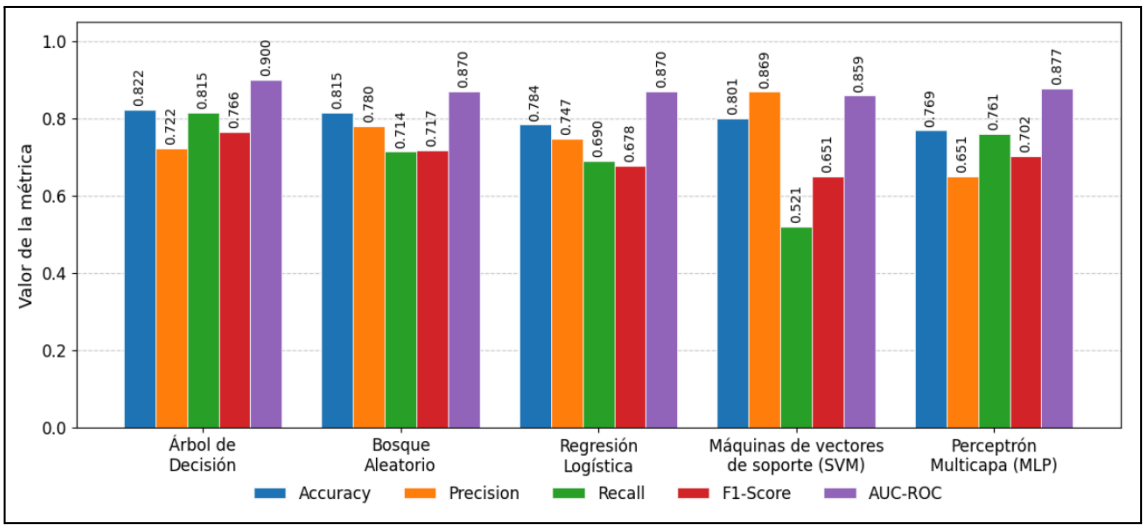
Nota. La tabla presenta las métricas de evaluación obtenidas para cada modelo aplicado.

[24] *fold* es un subconjunto del conjunto de datos original.En la validación cruzada k-fold, el conjunto de datos se divide en k folds (k partes del mismo tamaño aproximado).

Para complementar la tabla anterior, la Figura 21 sintetiza el desempeño de cinco clasificadores: árbol de decisión, bosque aleatorio, regresión logística, SVM y perceptrón multicapa, sobre el conjunto de prueba, considerando *accuracy, precision, recall, F1* y AUC-ROC.

Figura 21

Comparación de los modelos de clasificación



Nota. Elaboración propia a partir de los resultados obtenidos mediante la ejecución del código en Google Colab, posterior al análisis de la matriz de confusión.

5.4. Análisis de los resultados

Los resultados obtenidos indican que la predicción del riesgo de deserción a partir de la información disponible en el SIU-Guaraní es factible. En particular, los modelos alcanzan valores elevados de *recall* en la mayoría de los casos, métrica priorizada en este estudio por su importancia para la detección de estudiantes en riesgo de abandono, tal como se detalló en la sección 4.5.2. Asimismo, aunque se trata de algoritmos de naturaleza diferente evaluados sobre el mismo conjunto de datos, presentan patrones de desempeño consistentes, con niveles similares y aceptables de rendimiento global[25], lo que refuerza la confiabilidad de los resultados obtenidos.

El árbol de decisión presentó un *recall* elevado (0.807) y un rendimiento global equilibrado (*F1* = 0.763; *AUC-ROC* = 0.894; *accuracy* = 0.821), por lo que resulta una opción especialmente valiosa para un sistema de alerta temprana (implementar medidas de apoyo oportunas y eficaces para estudiantes en riesgo de abandono de los estudios), ya que permitiría detectar a la mayoría de los desertores reales sin generar un número excesivo de falsas alarmas.

[25] El rendimiento global es cómo se comporta el modelo en conjunto, mirando varias métricas a la vez y no solo una.

El bosque aleatorio, por su parte, ofrece un buen compromiso entre rendimiento predictivo (precisión = 0.762; AUC-ROC = 0.873) y estabilidad del desempeño, por lo que también constituye una alternativa adecuada para apoyar un sistema de alerta temprana: mantiene una capacidad razonable para diferenciar entre estudiantes que desertan y quienes continúan, sin generar un número excesivo de falsas alarmas.

El perceptrón multicapa se muestra como una segunda alternativa adecuada, al combinar un *recall* adecuado (0.772) con buena discriminación global porque (AUC-ROC = 0.887) es alto en consecuencia detecta bastantes casos de deserción y separa bien entre quienes desertan y quienes no.

La regresión logística obtuvo resultados sólidos y consistentes (precisión = 0.740; AUC-ROC = 0.873), aunque con un recall inferior (0.680) al del árbol de decisión y el perceptrón multicapa, por lo que no es la alternativa más adecuada como filtro principal de alerta temprana.

En contraste, las máquinas de vectores de soporte alcanzaron la mayor precisión (0.870), lo que vuelve muy confiables sus positivos predichos, pero a costa de un recall reducido (0.536), dejando sin detectar a una proporción importante de estudiantes que finalmente desertan, por lo que no resultan adecuadas como primer filtro en escenarios de alerta temprana.

Una vez obtenidos los desempeños de los distintos modelos, se realizó el análisis de las variables más influyentes para cada uno. Se encontró que las variables como año de ingreso, promedio (con y sin aplazo), tipo de vivienda y la carrera elegida se ubicaron sistemáticamente entre las más influyentes para distinguir entre estudiantes desertores y no desertores. Este patrón sugiere que la deserción se vincula principalmente con la cantidad de años cursados, el rendimiento académico y el costo económico de la vivienda; mientras que otros factores como cantidad de equivalencias, cantidad de hijos, localidad de origen y si trabaja, si bien relevantes desde la teoría, muestran un aporte cuantitativo menor en la cohorte estudiada.

5.5. Estrategias para reducir la deserción universitaria

Los hallazgos del estudio muestran que el riesgo de deserción se explica por la combinación de:

- extensión en el tiempo
- rendimiento
- condicionantes económicos

Gestión académica

- Sistema de Alerta Temprana (SAT): se activará una alerta cuando el estudiante presente simultáneamente: (i) inactividad académica ≥ 1 año; (ii) porcentaje de avance < 20 %; y (iii) al segundo año de cursado, una proporción de materias aprobadas < 25 % del plan de estudios. Ante la alerta, se realizará un contacto inicial (WhatsApp, correo electrónico o llamada breve) y se ofrecerá una tutoría de acompañamiento con plan de cursado personalizado.
- Ofertas académicas flexibles: implementar cursos intensivos y modalidades híbridas para favorecer la permanencia y el avance académico.
- Tutorías y planes de cursado personalizados: priorizar las asignaturas de mayor complejidad y tasa de desaprobación y ofrecer un acompañamiento sistemático a través de tutores.
- Mesas extraordinarias: habilitar mesas extraordinarias acotadas para estudiantes con alta carga laboral.
- Regularidad flexible: establecer criterios diferenciados para estudiantes que trabajan.

Bienestar y apoyos económicos

- Becas económicas.
- Convenio con distintas instituciones o servicio para descuentos a estudiantes.
- Residencias estudiantiles: de bajo costo, especialmente a aquellos que viven lejos de los centros educativos.

Comunidad, sentido de pertenencia y comunicación

- Canales de comunicación: ampliar y diversificar los medios para garantizar que el estudiantado reciba información sobre actividades académicas, deportivas y de bienestar, fortaleciendo su sentido de pertenencia y permanencia.
- Charlas de graduados/as: organizar encuentros con estudiantes de los primeros años para compartir experiencias académicas y profesionales en primera persona.
- Visitas anuales a empresas: promover el vínculo entre estudiantes, la facultad y las empresas del sector.
- Exposiciones, hackatones y ferias de proyectos: fomentar la participación y la visibilización de aprendizajes.

Inserción laboral compatible

- Pasantías con franjas horarias compatibles con el cursado.
- Acuerdos con empresas locales para tolerancia de horarios en parciales y finales.

- Bolsa de trabajo segmentada por carga horaria y remoto.

Estrategias pedagógicas y curriculares

- Nivelación y acompañamiento en los primeros cuatrimestres: la idea es atender a las dificultades que muchos estudiantes traen desde la secundaria o que surgen al ingresar a la universidad.
- Consultas programadas: habilitar espacios de consulta 2–3 semanas previas a los exámenes parciales y finales.

5.6. Limitaciones del estudio

Como todo estudio basado en registros administrativos y modelado predictivo, los resultados deben interpretarse considerando ciertas limitaciones que condicionan su alcance.

- La investigación se llevó a cabo en el marco de dos carreras específicas de la UADER. Esto reduce la validez externa y limita la posibilidad de extrapolar los resultados a otras carreras o contextos.
- Se identificaron valores faltantes, posibles errores de carga, como por ejemplo, coma/punto en números y categorías demasiado amplias que pueden afectar la precisión de los modelos.
- Análisis del tiempo de aprobación por asignatura (no realizado). No se modeló el tiempo hasta aprobar cada materia ni se identificaron asignaturas con alta dificultad o elevada tasa de desaprobación que funcionen como instancias críticas en el recorrido formativo. Esta omisión limita la comprensión de tramos críticos del plan de estudios y de cómo la secuencia de cursado impacta en el riesgo de deserción.
- Aunque se usaron cohortes históricas, la validación fue interna. No se aplicó aún una validación temporal externa (entrenar en años previos y testear en cohortes posteriores) para confirmar la capacidad de generalización en el tiempo.

Conclusiones

El estudio confirma que la minería de datos aplicada a los registros académicos permite identificar patrones y factores asociados a la deserción en las carreras de Análisis de Sistemas y Licenciatura en Sistemas de Información de la FCyT-UADER, cumpliendo el objetivo general propuesto.

En este estudio se revisaron las técnicas de minería de datos, se analizó la problemática de la deserción estudiantil, se procesaron los datos institucionales para poder utilizarlos en el entrenamiento de los modelos y se interpretaron resultados para proponer estrategias de reducción de la deserción, en línea con los objetivos específicos del trabajo.

Desde un enfoque analítico, se definió un conjunto de criterios para determinar si un alumno desertó. Los criterios fueron: bajo porcentaje de avance, pocas materias aprobadas y al menos dos años de inactividad. Esto fija un umbral claro para distinguir trayectorias de abandono dentro de la base de datos del SIU-Guaraní, que organiza la información en torno al rendimiento y la trayectoria (porcentaje de avance en la carrera, cantidad de materias aprobadas, promedios, años de inactividad), al contexto de vida (tipo de vivienda: propia o alquilada, localidad de residencia y de origen) y a las condiciones de sostenimiento y trabajo (formas de costeo, empleo, hora semanales, presencia de becas).

Cuando se comparan estos resultados con la teoría, se observan coincidencias claras. El modelo de Tinto se refleja en la importancia del rendimiento académico y la continuidad en la trayectoria (año de ingreso, inactividad). El modelo de Bean y Metzner, también se confirma en variables ambientales como trabajo, horas semanales y situación familiar. Finalmente, los enfoques psicopedagógicos resaltan la necesidad de fortalecer hábitos de estudio y acompañamiento académico, lo que coincide con los factores destacados por los modelos.

Considerando el desempeño de los distintos modelos evaluados, se determinó que el modelo de árbol de decisión fue el que presentó un rendimiento superior en comparación con las demás alternativas analizadas.

De manera complementaria, tras evaluar el desempeño de los modelos se procedió a examinar las variables con mayor impacto en la predicción de la deserción. El análisis de importancia manifestó que el año de ingreso, el promedio general (con y sin aplazos), el tipo de vivienda y la carrera cursada aparecen de forma recurrente entre los predictores más destacados al separar estudiantes desertores de quienes permanecen. En conjunto, estos resultados indican que el abandono se explica principalmente por la combinación entre la trayectoria temporal en la carrera y el rendimiento académico, en un contexto marcado por las condiciones económicas y habitacionales. En cambio, otros

factores como la cantidad de equivalencias, la presencia de hijos, la localidad de origen o si posee empleo, aunque respaldados por la teoría, presentan un peso cuantitativo menor en la muestra analizada.

En síntesis, la integración de minería de datos en el ámbito educativo abre la posibilidad de transformar la gestión universitaria: pasar de diagnósticos generales a acciones preventivas personalizadas. Con ello, se puede fortalecer tanto la capacidad predictiva, contribuyendo a reducir la deserción y mejorar las oportunidades de los estudiantes para finalizar sus estudios.

Bibliografía

Libros

- Haykin, S. (2009). *Neural networks and learning machines* (3rd ed.). Pearson.
- Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la minería de datos.* Pearson Prentice Hall. Marqués, M. P. (2015). Minería de datos a través de ejemplos (1ra ed.). Alfaomega.
- Lara Torralbo, J. A. (2014). *Minería de datos*. UDIMA.
- Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2.* Packt Publishing.
- Sampieri, R. H. (2014). *Metodología de la investigación* (6ta ed.).
- Sanseau, M., Sánchez Cestona, J., & Calio, S. (2023). *Permanencia de las y los estudiantes en la universidad. Comisión Nacional de Evaluación y Acreditación Universitaria (CONEAU) y Organización de Estados Iberoamericanos (OEI)*.

Artículos académicos

- Alaminos-Fernández, A. F. (2022). *Árboles de decisión en R con Random Forest*. Alicante. https://rua.ua.es/dspace/handle/10045/133067
- Cabrera, L., Bethencourt, J., Alvarez Pérez, P., & González Afonso, M. (2006). *El problema del abandono de los estudios universitarios.* Revista Electrónica de Investigación y Evaluación Educativa. https://turia.uv.es/index.php/RELIEVE/article/view/4226/3833
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery in databases.* AI Magazine, 17(3), 37–54. https://www.researchgate.net/figure/Figura-21-Etapas-del-proceso-de-KDD-adaptado-de-Fayyad-et-al-1996_fig1_49911537
- Fonseca, G., & García, F. (2016). *Permanencia y abandono de estudios en estudiantes universitarios: Un análisis desde la teoría organizacional.* Revista de la Educación Superior, 45(179), 25–39. https://doi.org/10.1016/j.resu.2016.06.004
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Paramo, G., & Correa Maya, C. (1999). *Deserción estudiantil universitaria. Conceptualización.* Revista Universidad Eafit. https://repository.eafit.edu.co/server/api/core/bitstreams/4222d0d2-39d3-4ccf-b8b6-de7e3e3f4231/content

- Revista Espacios. (2019). *Aplicación de redes neuronales en sistemas de predicción*. Espacios, 40(20). https://www.revistaespacios.com/a19v40n20/19402006.htm
- Seminara, M. P. (2020). Deserción y demora universitaria: lo que los indicadores y los rankings dejan afuera. El caso de la carrera de Bioingeniería de la U.N.S.J – Argentina. Miradas, 15(1), 87–106. https://doi.org/10.22517/25393812.24471
- Calloni, S. A. (2025). Modelos de Aprendizaje Automático para identificación de estudiantes en riesgo de abandono en la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires [Tesis de licenciatura, Universidad de Buenos Aires]. https://lcd.exactas.uba.ar/modelos-de-aprendizaje-automatico-para-identificacion-de-estudiantes-en-riesgo-de-abandono-en-la-facultad-de-ciencias-exactas-y-naturales-de-la-universidad-de-buenos-aires/

Recursos web

- Amazon Web Services. (s.f.). ¿Qué es la regresión logística? Amazon. https://aws.amazon.com/es/what-is/logistic-regression/
- Aprende IA. (2020, 8 de septiembre). Algoritmo Agrupamiento Jerárquico – Teoría. AprendeIA. https://aprendeia.com/2020/09/08/algoritmo-agrupamiento-jerarquico-teoria/
- AprendeIA. (2021, 5 de octubre). ¿Qué es el perceptrón? Perceptrón simple y multicapa. AprendeIA. https://aprendeia.com/2021/10/05/que-es-el-perceptron-simple-y-multicapa/
- Calvo, D. (2018). Clúster Jerárquicos y No Jerárquicos. DiegoCalvo.es. https://www.diegocalvo.es/cluster-jerarquicos-y-no-jerarquicos/
- Calvo, D. (2018, marzo 17). Clúster jerárquicos: estrategia aglomerativa vs. divisiva. DiegoCalvo.es. https://www.diegocalvo.es/cluster-jerarquicos-estrategia-aglomerativa-vs-divisiva/
- Cardellino, F. (2021, marzo 22). Tutorial para un clasificador basado en bosques aleatorios. freeCodeCamp. https://www.freecodecamp.org/espanol/news/random-forest-classifier-tutorial-how-to-use-tree-based-algorithms-for-machine-learning/
- Chugh, V. (2024, octubre 1). AUC y curva ROC en aprendizaje automático. DataCamp. https://www.datacamp.com/es/tutorial/auc

- Codificando Bits. (s.f.). La imputación: ¿inventar datos?. Codificando Bits. https://codificandobits.com/blog/guia-manejo-datos-faltantes/
- Codificando Bits. (s.f.). Validación cruzada y k-fold cross-validation. Codificando Bits. Recuperado el 15 de noviembre de 2025, de https://codificandobits.com/blog/validacion-cruzada-k-fold-cross-validation/
- DataCamp. (s.f.). SVM Classification in Python using Scikit-learn. Recuperado el 21 de junio de 2025, de https://www.datacamp.com/es/tutorial/svm-classification-scikit-learn-python
- DataCamp. (2024, 5 de marzo). Tutorial sobre máquinas de vectores de soporte con Scikit-learn. https://www.datacamp.com/es/tutorial/svm-classification-scikit-learn-python
- DataCamp. (2019, 27 de diciembre). Support Vector Machines with scikit-learn. DataCamp. https://www.datacamp.com/es/tutorial/svm-classification-scikit-learn-python
- DataScientest. (2021, diciembre 16). ¿Qué es la regresión logística? https://datascientest.com/es/que-es-la-regresion-logistica
- Fernández Morís, I. (2015). Aplicaciones SIG en proyectos de ordenación de montes [Trabajo Fin de Máster, Universidad de Oviedo]. Repositorio institucional. https://digibuo.uniovi.es/dspace/bitstream/handle/10651/29817/TFM_Fern%C3%A1ndez_Mor%C3%ADs.pdf?sequence=6
- Ferrero, R. (2017). ¿Cómo lidiar con los datos atípicos (Outliers)?. Máxima Formación. https://www.maximaformacion.es/blog-dat/como-lidiar-con-los-datos-atipicos-outliers/
- GeeksforGeeks. (2025, 14 de mayo). K-Nearest Neighbor (KNN) Algorithm. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/
- GeeksforGeeks. (2025, 25 de agosto). Naive Bayes Classifiers. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/naive-bayes-classifiers/
- GeeksforGeeks. (s.f.). What is Perceptron: The Simplest Artificial Neural Network. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/what-is-perceptron-the-simplest-artificial-neural-network/
- Google Cloud. (s.f.). Aprendizaje supervisado y no supervisado: ¿en qué se diferencian? Google Cloud. https://cloud.google.com/discover/supervised-vs-unsupervised-learning?hl=es

- Google Developers. (2025). Clasificación: Exactitud, recuperación, precisión y métricas relacionadas. Google. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=es-419
- Guía completa para el Manejo de Datos Faltantes. (s.f.). Codificando Bits. https://www.codificandobits.com/blog/manejo-datos-faltantes/
- IBM. (s.f.). Clustering: cómo funciona y tipos comunes. IBM. https://www.ibm.com/mx-es/think/topics/clustering
- IBM. (s.f.). ¿Qué es el algoritmo de k vecinos más cercanos (KNN)? IBM Think. Recuperado el 15 de noviembre de 2025, de https://www.ibm.com/es-es/think/topics/knn
- IBM. (s.f.). ¿Qué es big data? IBM. https://www.ibm.com/es-es/topics/big-data
- IBM. (s.f.). ¿Qué es el algoritmo K-Vecinos más cercanos (KNN)?. IBM Think. https://www.ibm.com/es-es/think/topics/knn
- IBM. (s.f.). ¿Qué es Naive Bayes? https://www.ibm.com/es-es/think/topics/naive-bayes
- IBM. (s.f.). ¿Qué es una matriz de confusión? IBM Think. Recuperado el 12 de noviembre de 2025, de https://www.ibm.com/es-es/think/topics/confusion-matrix
- IBM. (2021). Conceptos básicos de ayuda de CRISP-DM. IBM Docs. Recuperado de https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview
- IBM. (2021, septiembre 21). ¿Qué es la ciencia de datos? IBM. https://www.ibm.com/es-es/topics/data-science
- IBM. (s.f.). ¿Qué es la integración de datos? IBM. https://www.ibm.com/es-es/think/topics/data-integration
- IBM. (s.f.). ¿Qué es machine learning? IBM Think. Recuperado el 21 de noviembre de 2025, de https://www.ibm.com/mx-es/think/topics/machine-learning
- IBM. (2021, 2 de noviembre). ¿Qué es un árbol de decisión?. IBM. https://www.ibm.com/mx-es/think/topics/decision-trees
- Ministerio de Justicia y Derechos Humanos de la Nación Argentina. (s.f.). Datos personales.Argentina.gob.ar. https://www.argentina.gob.ar/justicia/derechofacil/leysimple/datos-personales
- Naukri.com. (s.f.). Linear vs non-linear classification. Naukri.com. https://www.naukri.com/code360/library/linear-vs-non-linear-classification
- Normas APA – 7ma (séptima) edición. (s.f.). https://normas-apa.org/
- Numiqo. (s.f.). Regresión lineal – Explicación sencilla. Recuperado el 4 de noviembre de 2025, de https://numiqo.es/tutorial/linear-regression

- Ortega, C. (2024, abril 16). Árbol de decisión: Qué es, tipos, ventajas y ejemplo. QuestionPro. https://www.questionpro.com/blog/es/arbol-de-decision/
- Síntesis de información Estadísticas Universitarias 2022-2023. (2024). https://www.argentina.gob.ar/sites/default/files/sintesis_2022-_2023.pdf
- The Data Visualisation Catalogue (s.f.). Diagrama Cajas y Bigotes. Datavizcatalogue.com. https://datavizcatalogue.com/ES/metodos/diagrama_cajas_y_bigotes.html
- Towards Data Science. (2020, 14 de septiembre). Performance metrics: Confusion matrix, precision, recall, and F1 score. Medium. Recuperado el 12 de noviembre de 2025, de https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262
- Xunta de Galicia. (s.f.). Árbol de decisiones. Espazo Abalar. https://espazoabalar.edu.x

Imágenes

- Aprende IA. (2020, 8 de septiembre). Algoritmo Agrupamiento Jerárquico – Teoría. https://aprendeia.com/2020/09/08/algoritmo-agrupamiento-jerarquico-teoria/
- Aprende IA. (2021, 2 de noviembre). Qué son las redes neuronales artificiales. https://aprendeia.com/2021/11/02/que-son-las-redes-neuronales-artificiales/
- Barba Alonso, R. (2022). Sistema predictivo de probabilidad de compra en e-commerce [Trabajo Fin de Máster, Universidad de Valladolid]. https://uvadoc.uva.es/bitstream/handle/10324/55580/TFM-G1544.pdf?sequence=1&isAllowed=y
- Damavis. (2020). Perceptrón Simple: Definición matemática y propiedades. Recuperado de https://blog.damavis.com/perceptron-simple-definicion-matematica-y-propiedades/
- DataCamp. (2023). Clasificación SVM con Scikit-Learn en Python. https://www.datacamp.com/es/tutorial/svm-classification-scikit-learn-python
- Deepnote. (s.f.). Cross-Validation en Python. Recuperado el 3 de noviembre de 2025, de https://deepnote.com/app/a_mas/Cross-Validation-en-Python-685fa851-b5b2-4c5b-b5fb-3dc5ae64838f
- freeCodeCamp. (2025, 6 de enero). SVM kernels explained: How to tackle nonlinear data in machine learning.

https://www.freecodecamp.org/news/svm-kernels-how-to-tackle-nonlinear-data-in-machine-learning/

- Gaviola, E., et al. (2023). Metodología para la clasificación de cobertura del suelo en Argentina [Trabajo presentado en las Jornadas IDERA 2023, Universidad Nacional del Comahue]. https://opendata.fi.uncoma.edu.ar/jornadasIDERA/trabajos2023/Gaviola_etal.pdf
- GeeksforGeeks. (2025). AUC-ROC Curve in Machine Learning. Recuperado de https://www.geeksforgeeks.org/machine-learning/auc-roc-curve/
- GeeksforGeeks. (2025). K-Nearest Neighbors (KNN) Algorithm. Recuperado de https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/
- Google Developers. (s.f.). Clasificación: Curva ROC y AUC. Machine Learning Crash Course. Recuperado el 13 de noviembre de 2025, de https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419#area_under_the_curve_auc
- Google Developers. (s.f.). Curso intensivo de aprendizaje automático: Regresión lineal. https://developers.google.com/machine-learning/crash-course/linear-regression?hl=es-419
- Gupta, P. (2023, 3 de agosto). Mastering the Training of Neural Networks: Key Points to Achieve Success. Medium. https://medium.com/@palakgupta33939/mastering-the-training-of-neural-networks-key-points-to-achieve-success-1b4cd27d8f80
- JMP. (s.f.). Media, mediana y moda. Statistics Knowledge Portal. https://www.jmp.com/es/statistics-knowledge-portal/measures-of-central-tendency-and-variability/mean-median-and-mode
- Ministerio de Educación de la Nación Argentina, Secretaría de Políticas Universitarias. (2023). Síntesis de información estadística universitaria 2022-2023. https://www.argentina.gob.ar/sites/default/files/sintesis_2022-_2023.pdf
- Muñoz Ramos, M. (2025). Trabajo de Fin de Grado en Ingeniería Informática [Trabajo de Fin de Grado, Universidad Politécnica de Madrid]. Repositorio UPM. https://oa.upm.es/90222/1/TFG_MIGUEL_MUNOZ_RAMOS.pdf
- Numiqo. (2025). Regresión Logística – Explicación sencilla. Recuperado de https://numiqo.es/tutorial/logistic-regression
- Raschka, S., & Mirjalili, V. (2017). Python Machine Learning (2nd ed.). Packt Publishing.

Anexos

Anexo A: Glosario

- **Análisis avanzado:** es la aplicación de técnicas complejas de aprendizaje automático (ML) e inteligencia artificial (IA) para descubrir patrones ocultos.
- **Algoritmo:** conjunto de instrucciones o reglas definidas que permiten resolver un problema o realizar un proceso, como clasificar estudiantes según su riesgo de deserción.
- **Árboles de Decisión (DT):** modelo predictivo que utiliza una estructura en forma de árbol para tomar decisiones basadas en reglas simples y sucesivas divisiones de los datos.
- **Big Data:** conjunto de datos masivos, complejos y en constante crecimiento que no pueden ser gestionados eficazmente por los sistemas tradicionales de bases de datos. Se caracteriza principalmente por su volumen, velocidad y variedad, y su análisis adecuado permite descubrir patrones, tendencias y relaciones útiles para la generación de conocimiento y valor.
- **Bosque Aleatorio (RF):** conjunto de árboles de decisión en el que cada árbol se entrena con un ruido aleatorio específico.
- **Ciencia de Datos (Data Science):** disciplina interdisciplinaria que combina estadística, matemáticas, programación, análisis avanzado, inteligencia artificial y machine learning con conocimientos específicos de un dominio para extraer información valiosa de grandes volúmenes de datos y apoyar la toma de decisiones.
- **Clasificador Bayes ingenuo:** algoritmo probabilístico que clasifica datos asumiendo independencia entre las variables predictoras.
- **Descubrimiento de conocimiento en bases de datos (KDD):** proceso completo que abarca desde la recolección y limpieza de datos hasta la interpretación de los patrones extraídos mediante minería de datos.
- **Deserción estudiantil:** situación en la que un estudiante abandona sus estudios de manera temporal o definitiva antes de completar el trayecto académico previsto.
- **Educación superior:** nivel educativo que abarca la educación terciaria y universitaria, incluyendo carreras de grado y posgrado.
- **Inteligencia Artificial (IA):** campo de la informática que busca desarrollar sistemas capaces de realizar tareas que requieren inteligencia humana, como el razonamiento, la percepción, el reconocimiento de voz o imágenes y la toma de decisiones. La IA abarca múltiples técnicas, entre ellas el machine learning, y

constituye un área clave para el desarrollo de aplicaciones en la industria, la ciencia y la vida cotidiana.

- **K-vecinos más cercanos (KNN):** algoritmo que clasifica una instancia según la mayoría de las categorías de sus K vecinos más cercanos en el espacio de características.
- **Machine Learning (ML):** rama de la inteligencia artificial que utiliza algoritmos y modelos estadísticos para que los sistemas “aprendan” automáticamente a partir de los datos, sin ser programados de manera explícita. Su objetivo principal es identificar patrones y realizar predicciones basadas en la experiencia previa contenida en los datos.
- **Máquinas de Vectores de Soporte (SVM):** modelo que busca un hiperplano óptimo para separar datos en diferentes clases con el máximo margen posible.
- **Matriz de confusión:** tabla que resume el desempeño de un modelo de clasificación, mostrando el número de aciertos y errores cometidos en cada clase. Permite comparar las predicciones del modelo con los valores reales.
- **Minería de datos:** proceso de exploración y análisis automatizado de grandes volúmenes de datos con el objetivo de descubrir patrones ocultos, relaciones significativas y tendencias útiles para la toma de decisiones.
- **Modelo:** son programas que han sido entrenados con un conjunto de datos para reconocer ciertos patrones o tomar decisiones sin intervención humana adicional.
- **Muestreo:** técnica estadística que consiste en seleccionar una parte representativa de una población con el fin de realizar inferencias sobre el total.
- **Patrones:** se refieren a una configuración regular, repetitiva o predecible encontrada en los datos. Estos patrones pueden manifestarse como tendencias, secuencias, correlaciones u otras formas de estructuras discernibles que emergen del análisis de datos.
- **Perceptrón multicapa (MLP):** tipo de red neuronal artificial formada por varias capas de neuronas.
- **Predicción:** estimación o pronóstico de un valor o evento futuro basado en datos históricos. Se usa en minería de datos para anticipar, por ejemplo, la probabilidad de deserción de un estudiante.
- **Problemática:** conjunto de problemas o desafíos relacionados con un tema. En este caso, se refiere a las causas estructurales, sociales y personales que explican la deserción.
- **Programación especializada:** es un área de la programación informática que se enfoca en el desarrollo de aplicaciones y soluciones adaptadas a las necesidades

específicas de una industria o un problema particular, en lugar de ser de propósito general.

- **Redes Neuronales Artificiales (RNA):** sistemas inspirados en el cerebro humano que consisten en nodos interconectados para modelar relaciones complejas y patrones no lineales.
- **Regresión Logística (RL):** técnica estadística que modela la probabilidad de una clase binaria en función de variables independientes.
- **Sobreajuste (overfitting):** situación en la que un modelo se ajusta demasiado a los datos de entrenamiento y falla al generalizar a nuevos datos.
- **Técnica:** método o procedimiento concreto dentro de un enfoque más amplio, por ejemplo, el uso de árboles de decisión dentro de las técnicas de clasificación.
- **Valores atípicos (outliers):** datos que se alejan significativamente del resto del conjunto. Pueden indicar errores o casos especiales y afectan el análisis si no se tratan adecuadamente.
- **Variables:** características o atributos medibles de los estudiantes (como edad, ingresos, rendimiento académico) que se analizan para encontrar relaciones con la deserción.

Anexo B: Fórmulas matemáticas

Media, mediana y moda

Dentro de la estadística descriptiva existen tres formas principales de representar un conjunto de datos con un solo valor: media, mediana y moda (JMP, s.f.):

- media aritmética (promedio): La media mide el centro de un conjunto de valores de datos. Para datos continuos, la media es el promedio de los valores de los datos. La media se utiliza a menudo como una simple estadística de resumen de un conjunto de datos.
    - Ej.: notas de un examen = 2, 4, 6 y 8. Media = (2+4+6+8)/4 = 5
- mediana: corresponde al percentil de una muestra de datos; es decir, el 50 % de los valores se ubica por encima de ella y el otro 50 % por debajo. Se trata de una medida de tendencia central que ofrece una estimación del "centro" de los datos de la muestra.
    - Si la cantidad de datos es par, se calcula el promedio de los dos del centro. Ej.: serie = 2, 6, 11, 19, 33. La mediana es 11 (divide al conjunto en dos mitades).
    - Si añadimos otro valor (45): 2, 6, 11, 19, 33, 45. La mediana es (11+19) /2 = 15
- moda: es el valor que aparece con más frecuencia en los datos. Un conjunto de datos que no contiene valores que se repiten no tiene moda. Un conjunto de datos con múltiples valores que se repiten con la misma frecuencia puede tener múltiples modas. La moda es otra estadística utilizada para estimar el centro de los datos.
    - Ej.: edades de un equipo de vóley = 14, 15, 13, 15, 16,15.Ordenando: 13 – 14 – 15 – 15 – 15 – 16 → la moda es 15.

Regresión lineal

La regresión lineal es una técnica estadística utilizada para modelar la relación entre una variable dependiente y una o más variables independientes, mediante el ajuste de una recta a los datos observados (Google Developers, s.f.).

En su forma más simple, el modelo puede expresarse algebraicamente como:

$$y = m.\ x_1 + b$$

Donde:

- ➔ y el valor que queremos predecir.
- ➔ m es la pendiente de la recta.
- ➔ x nuestro valor de entrada.
- ➔ b intersección con el eje y.

La ecuación para un modelo de regresión lineal se escribe de la siguiente manera:

$$y' = b + w_1 x_1$$

Donde:

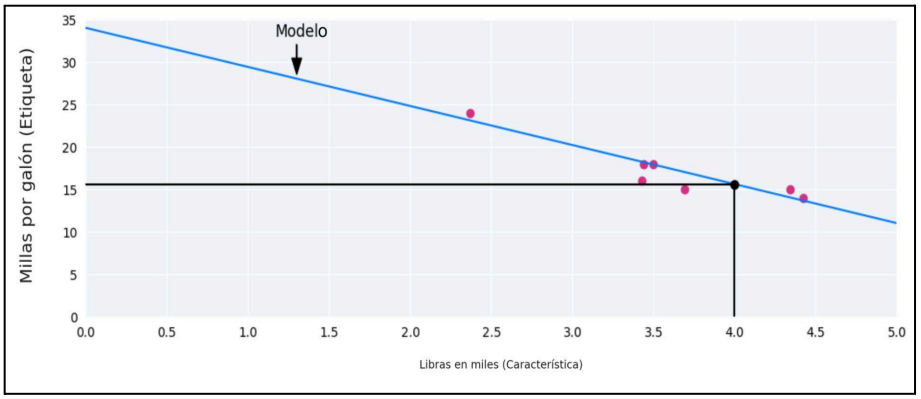
- ➔ y' es la etiqueta predicha, es decir, el resultado
- ➔ b es el sesgo del modelo
- ➔ w_1 es el peso del atributo
- ➔ x_1 es un atributo, la entrada

Durante el entrenamiento, el modelo calcula el peso y el sesgo que producen el mejor modelo.

En este ejemplo, se calcula el peso y la desviación a partir de la línea trazada. El sesgo es 34 (donde la línea se cruza con el eje Y) y el peso es -4.6 (la pendiente de la línea). Por ejemplo, con este modelo, se predeciría que un automóvil de 1,814 kg tendría una eficiencia de combustible de 6.6 km/l.

Figura 22

Regresión lineal



Nota. Adaptado de Linear regression, en Machine Learning Crash Course, por Google Developers, 2025

Con el modelo, se predice que un automóvil de 4,000 libras tiene una eficiencia de combustible de 15.6 millas por galón.

Función logística

La función logística es también llamada función sigmoide

$$f(x) = \frac{1}{1+e^{-z}}$$

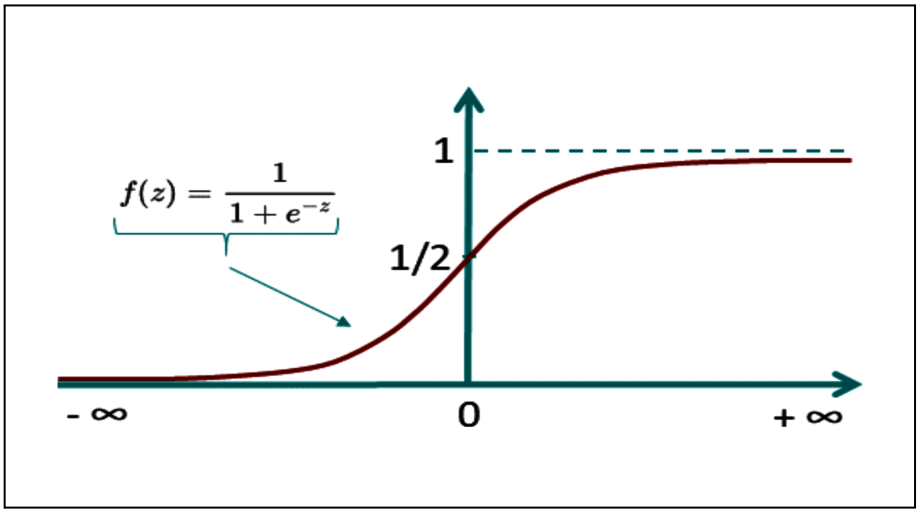
Donde:

- ➔ $f(x)$ resultado de aplicar la función sigmoide a z.
- ➔ z es una combinación lineal de variables de entrada
- ➔ e número de Euler, base de los logaritmos naturales

La gráfica de esta función se observa en la siguiente figura.

Figura 23

Función de regresión logística



Nota. Adaptado de Regresión Logística, por Numiqo, 2025.

Según Numiqo (s.f.), para construir un modelo de regresión logística, se parte de la ecuación de regresión lineal.

$$\widehat{Y} = b_1 . x_1 + b_2 . x_2 + ... + b_k . x_k + a$$

Donde:

- ➔ $\widehat{Y}$ es la variable dependiente
- ➔ $x_1, x_2 \ldots x_k$ variables independiente
- ➔ $b_1 . , b_2 . b_k , a$ coeficiente de regresión
- ➔ a término constante

El modelo logístico se basa en la función lógica. Lo especial de la función logística es que para valores entre menos y más infinito, siempre asume sólo valores entre 0 y 1.

Así es que la función logística es perfecta para describir la probabilidad $p(y = 1)$. Si ahora se aplica la función logística a la ecuación de regresión lineal, el resultado es:

$$f(x) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a)}}$$

Esto garantiza que, independientemente del intervalo en que se encuentren los valores de x, sólo resultarán valores entre 0 y 1.

La probabilidad p de que para unos valores dados de la variable independiente la variable dependiente dicotómica y sea 0 ó 1 viene dada por:

$$p(y = 1 \mid x_1, \ldots, x_n) = \frac{1}{1 + e^{-(b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a)}}$$

$$p(y = 0 \mid x_1, \ldots, x_n) = \frac{1}{1 + e^{-(b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a)}}$$

Teorema de Bayes

El teorema de Bayes proporciona un método basado en principios para invertir las probabilidades condicionales.

Se define como:

$$P(y \mid X) = P(X \mid y) \cdot P(y)/P(X)$$

Donde:

- ➔ $P(y \mid X)$: es la probabilidad posterior: la probabilidad de la clase y dado el conjunto de características X
- ➔ $P(X \mid y)$: es la verosimilitud: la probabilidad de observar X dado que la clase es y
- ➔ $P(y)$: es la probabilidad a priori de la clase y
- ➔ $P(X)$: es la probabilidad marginal o evidencia

Métricas de distancia utilizadas en el algoritmo KNN

Según IBM (s.f.), el objetivo de *KNN* es, para un punto de consulta dado, encontrar sus vecinos más próximos y con esa información asignarle una etiqueta de clase. Para determinar cuáles son esos vecinos, el algoritmo compara distancias con alguna de las métricas de distancias habituales como se detallan a continuación:

Distancia euclidiana

La distancia euclidiana se define como la distancia en línea recta entre dos puntos en un plano o espacio. Puedes pensar en ella como el camino más corto que recorrería si fueras directamente de un punto a otro.

$$Distancia\ euclidiana\ = d(x,y) = \sqrt{\sum_{i=1}^{n} (yi - xi)^2}$$

Donde:

- ➔ $d(x,y)$: es la distancia euclidiana entre los puntos x y y.
- ➔ xi: es el valor del i-ésimo componente del vector x.
- ➔ yi: es el valor del i-ésimo componente del vector y.
- ➔ n: es el número de dimensiones del espacio en el que se encuentran los puntos x y y.
- ➔ $\sum_{i=1}^{n} (yi - xi)^2$: representa la suma de los cuadrados de las diferencias entre las coordenadas correspondientes de los puntos x y y.

Distancia de Manhattan

Esta es la distancia total que recordarías si solo pudieras moverte a lo largo de líneas horizontales y verticales, como una cuadrícula o las calles de una ciudad. También se llama "distancia de taxi" porque un taxi solo puede circular por las calles de una ciudad en cuadrícula.

$$Distancia\ Manhattan\ = d(x,y) = \sum_{i=1}^{m} |xi - yi|$$

donde:

- ➔ $d(x,y)$: es la distancia Manhattan entre los puntos x y y.
- ➔ xi: es el valor del i-ésimo componente del vector x.
- ➔ yi: es el valor del i-ésimo componente del vector y.
- ➔ m: es el número de dimensiones del espacio. Sería 2 si es un plano.
- ➔ $|xi - yi|$: es el valor absoluto de la diferencia entre las coordenadas xi y yi, esto es, qué tan separados están sin importar la dirección.

Distancia de Minkowski

La distancia de Minkowski es como una familia de distancias, que incluye las distancias euclidianas y de Manhattan como casos especiales.

$$Distancia\ Minkowski\ = d(x,y) = \left(\sum_{i=1}^{n} |xi - yi| \right)^{1/p}$$

donde:

- ➔ $d(x, y)$: es la distancia Minkowski entre los puntos x y y.
- ➔ xi: es el valor del i-ésimo componente del vector x.
- ➔ yi: es el valor del i-ésimo componente del vector y.
- ➔ n: es el número de dimensiones del espacio.
- ➔ $|xi - yi|$: es el valor absoluto de la diferencia entre las coordenadas xi y yi, esto es, qué tan separados están sin importar la dirección.
- ➔ p: es el parámetro que determina el tipo de distancia.

De la fórmula anterior, cuando $p = 2$, se convierte en la fórmula de la distancia euclidiana, y cuando $p = 1$, se convierte en la fórmula de la distancia de Manhattan. La distancia de Minkowski es esencialmente una fórmula flexible que puede representar la distancia euclidiana o la distancia de Manhattan según el valor de p.

Aval del Codirector de tesis – FCyT 2025

Estudiante/s: A.S. Arrejin, Martin Raúl -A.S. Richard, Rodrigo Adrián

Título del trabajo de investigación: Minería de datos aplicada para la identificación de patrones de deserción universitaria.

Programa de estudios: Licenciatura en Sistemas de Información.

Fecha de entrega: 9/12/2025.

Por medio de la presente, me dirijo a ustedes en calidad de Director de Tesis a fin de dar consentimiento respecto al cumplimiento de los requisitos de escritura y calidad académica del trabajo final del cual presido, en conformidad con lo establecido en la “Resolución CD FCyT Nro. 503/23”.

Además, quiero dejar en claro que tanto los objetivos propuestos en el pre-proyecto aprobado en el año lectivo 2024 fueron respetados y llevados a cabo en todo el trabajo de investigación desarrollado.

Bioing. Hadad, Alejandro
hadad.alejandro@uader.edu.ar

Aval del co director de tesis – FCyT 2025

Estudiante/s: A.S. Arrejin, Martin Raúl -A.S. Richard, Rodrigo Adrián

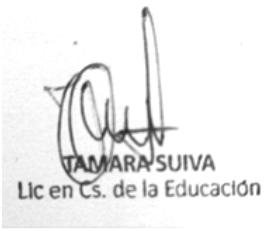
Título del trabajo de investigación: Minería de datos aplicada para la identificación de patrones de deserción universitaria.

Programa de estudios: Licenciatura en Sistemas de Información.

Fecha de entrega: 9/12/2025.

Por medio de la presente, me dirijo a ustedes en calidad de Director de Tesis a fin de dar consentimiento respecto al cumplimiento de los requisitos de escritura y calidad académica del trabajo final del cual presido, en conformidad con lo establecido en la “Resolución CD FCyT Nro. 503/23”.

Además, quiero dejar en claro que tanto los objetivos propuestos en el pre-proyecto aprobado en el año lectivo 2024 fueron respetados y llevados a cabo en todo el trabajo de investigación desarrollado.



Esp. Lic. Suiva, Tamara

Correo electrónico:suiva.tamara@uader.edu.ar