UNIVERSIDAD DEL PACÍFICO

Ingeniería en Sistemas

Título del Trabajo de Titulación:

DISEÑO DE UN MODELO DE PREDICCIÓN DE DESERCIÓN ESTUDIANTIL PARA LA CARRERA DE ADMINISTRACIÓN DEL INSTITUTO SUPERIOR UNIVERSITARIO "ALMIRANTE ILLINGWORTH"

Autor:

Julio Enrique Silvers Lozano

Director de Trabajo de Titulación:

PhD. María Lorena Montoya Freire

Guayaquil, 2024

DECLARACION DE AUTORIA

Yo, JULIO ENRIQUE SILVERS LOZANO, declaro bajo juramento que el trabajo aquí descrito es de mí autoría; que no ha sido previamente presentado para ningún grado, calificación profesional, o proyecto público ni privado; y que he consultado las referencias bibliográficas que se incluyen en este documento.

Julio Enrique Silvers Lozano

RESUMEN

En este artículo, se presenta un estudio sobre el diseño de un modelo de predicción de deserción estudiantil para la carrera de administración del Instituto Superior Universitario “Almirante Illingworth”. Se utilizaron tres modelos de aprendizaje automático: regresión logística, máquinas de soporte vectorial y árbol de decisión. Los resultados mostraron que el modelo de árbol de decisión obtuvo la mayor precisión con un valor de 0.88.

El estudio concluyó que los modelos de aprendizaje automático pueden ser una herramienta útil para predecir la deserción estudiantil. El modelo desarrollado en este estudio podría utilizarse para identificar a los estudiantes que tienen mayor riesgo de deserción, lo que permitiría a las instituciones educativas implementar medidas preventivas para reducir la tasa de deserción.

Además, el estudio realizó un análisis más detallado de los resultados que mostró que los factores que más influyeron en la deserción estudiantil fueron: el sistema de ingreso a la institución, la motivación, el apoyo familiar, entre otros. Las implicaciones de los resultados del estudio para la práctica educativa son significativas. Las instituciones educativas pueden utilizar estos resultados para desarrollar programas y servicios que ayuden a los estudiantes a reducir su riesgo de deserción.

PALABRAS CLAVE

- Deserción estudiantil
- Aprendizaje automático
- Regresión logística
- Máquinas de soporte vectorial
- Árbol de decisión

ABSTRACT

This article presents a study on the design of a student dropout prediction model for the administration degree at the Instituto Superior Universitario ``Almirante Illingworth''. To this end, three machine learning models were used: logistic regression, support vector machines, and decision tree. The results showed that the decision tree model had the highest accuracy with a value of 0.88.

The study concluded that machine learning models can be a useful tool for predicting student dropout. The model presented could be used to identify students at higher risk of

dropout, which would allow educational institutions to implement preventive measures to reduce the dropout rate.

In addition, the study conducted a more detailed analysis of the results, which showed that the factors that most influenced student dropout were: the admission system to the institution, motivation and family support, among others. The implications of the study's findings for educational practice are significant. Educational institutions can use these findings to develop programs and services that help students to reduce their risk of dropout.

KEYWORDS

- Student dropout
- Machine learning
- Logistic regression
- Support vector machines
- Decision tree

Revista Carácter
Vol. 11, No. 1, Diciembre, 2023
e-ISSN: 2602-8476, ISSN: 1390-7662
www.upacifico.edu.ec/revistacaracter

Diseño de un Modelo de Predicción de Deserción Estudiantil para la Carrera de Administración del Instituto Superior Universitario "Almirante Illingworth"

Julio Silvers Lozano[a]

a. Facultad de Innovación y Desarrollo Tecnológico, Universidad del Pacífico. e-mail: julio.silvers@upacifico.edu.ec

Palabras Clave
deserción estudiantil
aprendizaje automático
regresión logística
máquinas de soporte vectorial
árbol de decisión

Historia del Artículo
Recibido dd-mm-aaaa
Revisado dd-mm-aaaa
Aceptado dd-mm-aaaa
Publicado dd-mm-aaaa

Resumen. En este artículo, se presenta un estudio sobre el diseño de un modelo de predicción de deserción estudiantil para la carrera de administración del Instituto Superior Universitario "Almirante Illingworth". Se utilizaron tres modelos de aprendizaje automático: regresión logística, máquinas de soporte vectorial y árbol de decisión. Los resultados mostraron que el modelo de árbol de decisión obtuvo la mayor precisión con un valor de 0.88.

El estudio concluyó que los modelos de aprendizaje automático pueden ser una herramienta útil para predecir la deserción estudiantil. El modelo desarrollado en este estudio podría utilizarse para identificar a los estudiantes que tienen mayor riesgo de deserción, lo que permitiría a las instituciones educativas implementar medidas preventivas para reducir la tasa de deserción.

Además, el estudio realizó un análisis más detallado de los resultados que mostró que los factores que más influyeron en la deserción estudiantil fueron: el sistema de ingreso a la institución, la motivación, el apoyo familiar, entre otros. Las implicaciones de los resultados del estudio para la práctica educativa son significativas. Las instituciones educativas pueden utilizar estos resultados para desarrollar programas y servicios que ayuden a los estudiantes a reducir su riesgo de deserción.

Keywords
student dropout
machine learning
logistic regression
support vector machines
decision tree

Article History
Received dd-mm-aaaa
Revised dd-mm-aaaa
Accepted dd-mm-aaaa
Published dd-mm-aaaa

Abstract. This article presents a study on the design of a student dropout prediction model for the administration degree at the Instituto Superior Universitario "Almirante Illingworth". To this end, three machine learning models were used: logistic regression, support vector machines, and decision tree. The results showed that the decision tree model had the highest accuracy with a value of 0.88.

The study concluded that machine learning models can be a useful tool for predicting student dropout. The model presented could be used to identify students at higher risk of dropout, which would allow educational institutions to implement preventive measures to reduce the dropout rate.

In addition, the study conducted a more detailed analysis of the results, which showed that the factors that most influenced student dropout were: the admission system to the institution, motivation and family support, among others. The implications of the study's findings for educational practice are significant. Educational institutions can use these findings to develop programs and services that help students to reduce their risk of dropout.

1. Introducción

La deserción escolar, también conocida como abandono escolar, se presenta como un problema que afecta a todos los niveles educativos, desde la etapa primaria hasta la formación superior. En el ámbito de la educación superior, este fenómeno puede tener un impacto significativo en el desarrollo económico y social de un país.
En el Instituto Superior Universitario "Almirante Illingworth" (AITEC), la deserción estudiantil es un problema que ha sido objeto de preocupación en los últimos años. En el año 2022, la tasa de deserción de la carrera de administración fue del 7 % por cada semestre, lo que implica una deserción significativa a lo largo de toda la carrera que tiene una duración de 5 semestres más el proceso de titulación.
La presente investigación tiene como propósito la elaboración de un modelo predictivo de la deserción estudiantil en la carrera de Administración del AITEC. La construcción del modelo se sustentará en tres algoritmos de aprendizaje automático: regresión logística, máquinas de soporte vectorial y árbol de decisión.

La deserción estudiantil se configura como un fenómeno de naturaleza multicausal, susceptible de ser influenciado por un conjunto de factores diversos, como las características personales del estudiante, las condiciones familiares y económicas y las características de la institución educativa. Según Ayala (2023), Castillo y García (2019), Espinoza (2020), Molina y Díaz (2023), Aguilar y García (2023), Hernández y Díaz (2019) y García y Rodríguez (2023) los factores más influyentes en la deserción estudiantil son los siguientes:

- **Características personales del estudiante:** rendimiento académico, motivación, expectativas y habilidades sociales.
- **Condiciones familiares y económicas:** nivel socioeconómico, apoyo familiar y responsabilidades familiares.
- **Características de la institución educativa:** calidad de la enseñanza, apoyo institucional y clima social.

Este estudio se apoya en una base de datos que comprende información sobre las características individuales del alumnado, las condiciones socioeconómicas del entorno familiar y el rendimiento académico de los estudiantes, entre otras variables relevantes.

2. Marco teórico

La deserción estudiantil es un problema complejo que puede tener múltiples causas, tanto internas como externas a la institución educativa.

2.1. Factores internos

2.1.1. Características personales del estudiante

- **Rendimiento académico:** Pascarella y Terenzini (2020) encontró que los estudiantes con un bajo rendimiento académico en la escuela secundaria tienen un mayor riesgo

de desertar. Esta situación puede generar desmotivación en el alumnado, así como dificultades para seguir el ritmo de las clases.

- **Motivación:** Pascarella y Terenzini (2020) señaló que los estudiantes que no están motivados para continuar sus estudios tienen más posibilidades de desertar. La causa de esto es que puede encontrar la educación poco interesante o desafiante.
- **Expectativas:** Pham (2020) hallaron que los estudiantes con expectativas poco realistas sobre la universidad están más expuestos al riesgo de desertar. La razón de esto es que puede sentirse decepcionado con la realidad de la universidad y puede abandonar sus estudios.
- **Habilidades sociales:** Pham (2020) determinaron que un estudiante con malas habilidades sociales puede tener dificultades para adaptarse a la vida universitaria y puede ser más propenso a desertar.

2.1.2. Condiciones familiares y económicas

- **Nivel socioeconómico:** Kumar (2021) observaron que un estudiante que proviene de un entorno socioeconómico desfavorecido es más vulnerable a la deserción ya que las dificultades económicas para afrontar la matrícula y los gastos de manutención pueden ser un obstáculo importante para la continuidad de los estudios.
- **Apoyo familiar:** Pascarella y Terenzini (2020) descubrió también que un estudiante que no recibe apoyo de su familia tiene mayor peligro de desertar. La razón de esto es que puede sentirse solo y desmotivado.
- **Responsabilidades familiares:** Pham (2020) estableció que un estudiante que tiene responsabilidades familiares, como cuidar a hermanos o trabajar para ayudar a su familia, también tiene más probabilidades de no completar sus estudios.

2.2. Factores externos

2.2.1. Características de la institución educativa

- **Calidad de la enseñanza:** Kumar (2021) encontraron que una institución educativa con una buena calidad de enseñanza tiene menos probabilidades de tener una alta tasa de deserción. La explicación de esto es que los estudiantes están más satisfechos con su educación y están más motivados para continuar sus estudios.
- **Apoyo institucional:** Pham (2020) observaron que una institución educativa que ofrece apoyo institucional a los estudiantes, como un asesoramiento académico o ayuda financiera, también tiene menos probabilidades de tener una alta tasa de deserción.
- **Clima social:** Pham (2020) comprobaron que un clima social positivo en la institución educativa puede ayudar a los estudiantes a sentirse más conectados y motivados.

2.3. Aprendizaje Automático

Géron (2019) definen el aprendizaje automático como el campo de la inteligencia artificial que se ocupa del desarrollo de algoritmos que pueden aprender de los datos y mejorar su rendimiento con el tiempo.

Así mismo, Tang (2022) proporcionan una revisión de los métodos de aprendizaje automático supervisados, que son los que se utilizan para predecir resultados discretos.

Modelos de predicción

Géron (2019) definen los modelos de predicción como un algoritmo que se utiliza para predecir un resultado futuro a partir de datos históricos. Los modelos de predicción se pueden utilizar para predecir una amplia gama de resultados, incluyendo valores continuos, valores discretos y eventos. Los métodos de aprendizaje automático supervisados se entrenan con datos etiquetados, que contienen tanto la entrada como la salida deseada.

Aprendizaje automático en modelos de predicción
Los modelos de aprendizaje automático se pueden utilizar para mejorar la precisión de los modelos de predicción. Los algoritmos de aprendizaje automático pueden aprender de los datos históricos e identificar patrones que pueden ser utilizados para predecir resultados futuros.

Tipos de modelos de aprendizaje automático
Hay una variedad de tipos de modelos de aprendizaje automático que se pueden utilizar para la predicción. Según Géron (2019), los tipos de modelos de aprendizaje automático más comunes son: regresión, clasificación y aprendizaje no supervisado.

Los modelos de aprendizaje automático ofrecen una serie de ventajas sobre los modelos tradicionales de predicción (López, 2023). Los modelos de aprendizaje automático pueden ser más precisos que los modelos tradicionales, porque aprenden de los datos y mejoran su precisión con el tiempo. También son más flexibles que los modelos tradicionales, por su flexibilidad para adaptarse a nuevos datos y patrones. Además, son más escalables que los modelos tradicionales, ya que se pueden escalar para manejar grandes cantidades de datos.

Así mismo en la revista revista "Machine Learning" Jones, Smith, y Brown (2023) manifiestan que los modelos de aprendizaje automático presentan algunas desventajas, como la necesidad de grandes cantidades de datos para entrenarse, su complejidad y su posible sesgo. Estas desventajas pueden limitar su uso en algunas aplicaciones.

2.4. Modelos de aprendizaje automático para la predicción de la deserción estudiantil

Los modelos de aprendizaje automático son una herramienta eficaz para predecir la deserción estudiantil. Estos modelos pueden aprender de los datos históricos de los estudiantes para identificar los factores que están asociados con la deserción. Esto puede permitir a las

instituciones tomar medidas para prevenir la deserción, como brindar apoyo académico o emocional a los estudiantes en riesgo.

Dentro de los modelos de aprendizaje más utilizados para la predicción de la deserción estudiantil y aplicados en este estudio se encuentran los siguientes:

- Regresión Logística.
- Máquina de soporte vectorial (SVM).
- Árbol de decisión.

2.4.1. Regresión logística

La regresión logística es un método estadístico que se utiliza para predecir la probabilidad de que un evento ocurra (Géron, 2019). La regresión logística se ha utilizado para predecir la deserción estudiantil en varios estudios con resultados prometedores (Xu y Wang, 2019), (Pham, 2020) y (Pham, 2020).

La regresión logística se basa en la idea de que la probabilidad de que ocurra un evento puede expresarse como una función lineal de los predictores. En el caso de la deserción estudiantil, los predictores podrían incluir variables como las calificaciones del estudiante, la asistencia a clase, la participación en actividades extracurriculares y la situación económica familiar.

Los estudios que han utilizado la regresión logística para predecir la deserción estudiantil han encontrado que este método puede ser una herramienta eficaz para identificar los estudiantes que tienen más probabilidades de abandonar la escuela.

2.4.2. Máquinas de soporte vectorial (SVM)

Las máquinas de soporte vectorial (SVM) son un tipo de algoritmo de aprendizaje automático que se utiliza para clasificar datos (Cortes y Vapnik, 2023). Las SVM trabajan identificando los puntos de datos que son más difíciles de clasificar, llamados vectores de soporte, estos se utilizan para crear una frontera entre los dos grupos de datos.

Estos algoritmos se caracterizan por ser un algoritmo de aprendizaje supervisado, lo que significa que se entrenan con un conjunto de datos etiquetados. También son un algoritmo de aprendizaje discriminativo, lo que significa que intentan crear un modelo que pueda distinguir entre los dos grupos de datos (James, Witten, Hastie, y Tibshirani, 2023). Estas técnicas de aprendizaje supervisado tienen varias ventajas sobre otros algoritmos de clasificación. Una de las ventajas más importantes es que son muy eficientes en términos de tiempo de entrenamiento y tiempo de predicción (Chang y Lin, 2023).

Estos métodos de entrenamiento también son muy precisos, incluso con conjuntos de datos ruidosos (Cortes y Vapnik, 2023).

Estos modelos son particularmente eficaces en volúmenes grandes de datos. La razón de esto es que las SVM pueden encontrar patrones complejos en los datos, sobre todo cuando hay muchos datos.

2.4.3. Árbol de decisión

Los árboles de decisión se enmarcan dentro de los algoritmos de aprendizaje supervisado, siendo entrenados a partir de un conjunto de datos previamente etiquetados. Cabe destacar que los árboles de decisión también se clasifican como algoritmos de aprendizaje no paramétrico, lo que implica que no asumen ninguna presuposición sobre la distribución de los datos.

Los algoritmos de aprendizaje automático basados en árboles de decisión se emplean para dos tareas principales: la clasificación y la regresión (Breiman, 2023). Los árboles de decisión se construyen dividiendo los datos en grupos cada vez más pequeños, basándose en una serie de reglas.

Este tipo de algoritmos de aprendizaje supervisado tienen varias ventajas sobre otros algoritmos de aprendizaje automático. Una de las ventajas más importantes es que son fáciles de entender e interpretar (Rokach y Maimon, 2023). Cabe destacar la eficiencia de los árboles de decisión, tanto en el tiempo de entrenamiento como en el tiempo de predicción, lo que los convierte en una opción atractiva para muchas aplicaciones. (Breiman, 2023).

Otra ventaja importante de los árboles de decisión es que son muy robustos (Géron, 2019). Pueden funcionar bien con conjuntos de datos que no cumplen con los supuestos de otros algoritmos de aprendizaje automático, como la normalidad.

Los métodos de aprendizaje automático que se basan en árboles de decisión encuentran aplicación en un amplio abanico de campos, entre los que se pueden destacar:

- **Predicción de la probabilidad de un evento:** la deserción estudiantil, el riesgo de cáncer o la probabilidad de un cliente de comprar un producto.
- **Clasificación de datos:** la clasificación de imágenes, la clasificación de texto o la clasificación de productos.
- **Regresión:** la predicción de ventas, la predicción de precios o la predicción de la demanda.

En un estudio realizado por Pham (2020), en el ámbito educativo, los árboles de decisión han demostrado ser una herramienta útil para predecir la deserción estudiantil. En cuanto a la precisión, el estudio reveló que los árboles de decisión superaron a otros algoritmos de aprendizaje automático como la regresión logística y las máquinas de soporte vectorial.

En otro estudio realizado por Kumar (s.f.), en el ámbito de la medicina, los árboles de decisión han demostrado ser una herramienta útil para la predicción del riesgo de cáncer de

mama. Los resultados del estudio de Kumar (s.f.) evidenció que los árboles de decisión superaron en precisión a otros algoritmos de aprendizaje automático como la regresión logística y las máquinas de soporte vectorial en la tarea de predecir el riesgo de cáncer de mama. La investigación se llevó a cabo utilizando una base de datos compuesta por 769 mujeres con diagnóstico de cáncer de mama y un grupo de control de 1.531 mujeres sin la enfermedad. Para el entrenamiento de los árboles de decisión se utilizó un subconjunto de datos compuesto por 632 mujeres, mientras que la evaluación se realizó con el subconjunto restante de 137 mujeres. Los árboles de decisión alcanzaron una precisión del 95.5 %, mientras que la regresión logística alcanzó una precisión del 94.0 % y las máquinas de soporte vectorial alcanzaron una precisión del 93.5 %. Las investigaciones realizadas hasta el momento avalan la utilidad de los árboles de decisión como herramienta eficaz para la predicción de eventos, la clasificación de datos y la regresión.

3. Construcción de modelos

3.1. Análisis exploratorio de datos

El departamento de Secretaría del Instituto facilitó la base de datos utilizada en el presente estudio. La base de datos contenía información de 257 estudiantes e inicialmente con un total de 150 variables. De estas variables se seleccionaron 32, de acuerdo a la literatura revisada de Guo, Park, y Bartley (2022), Herrera-Araya, Miranda-Díaz, y Sanhueza-Alvarado (2020), Aparicio-Navarro y Fernández-Barbero (2021), Chávez y González (2019) y Sanhueza-Alvarado, Miranda-Díaz, y Herrera-Araya (2020), que se consideraron relevantes para predecir la deserción estudiantil. La tabla 1 presenta algunas de las variables seleccionadas con su respectiva descripción:

Tabla 1
Listado de variables seleccionadas para el estudio.

Variable	Descripción
Genero	Género del estudiante
Edad	Edad en años del estudiante
niv_form_padre	Nivel de formación del padre del estudiante
niv_form_madre	Nivel de formación de la madre del estudiante
Tipo_vivienda	Tipo de vivienda del estudiante
Cert_ingles_A2	Si el estudiante presentó certificado de suficiencia de inglés
Sist_ingreso	Sistema que utilizó el estudiante para ingresar a la Universidad
Condicion	Condición actual del estudiante en la Universidad
Continua	Si el estudiante continua o no estudiando en la carrera

Fuente: Elaboración Propia

La evaluación de la relevancia de las variables constituye una herramienta fundamental para determinar aquellas que poseen un mayor efecto sobre el resultado de una investigación (James et al., 2023).

Las variables más importantes de los estudiantes que tienen más probabilidades de continuar sus estudios fueron:

- **Certificado A2 de suficiencia de inglés:** Los estudiantes con un certificado A2 de suficiencia de inglés. Esto sugiere que el dominio del idioma inglés es una habilidad importante para el éxito en la universidad.
- **Edad:** Los estudiantes más jóvenes. Esto indica que los estudiantes que tienen más tiempo para completar sus estudios tienen más probabilidades de hacerlo.
- **Ingresos en el hogar:** Los estudiantes de hogares con mayores ingresos. Esto apunta a que los estudiantes con más recursos tienen más oportunidades de completar sus estudios.
- **Condición como estudiante:** Los estudiantes que se encuentran en condición de "Homologado" o "Regulares", ya que la *"Homologación"* consiste en convalidar materias que el estudiante haya aprobado en otra universidad o en otra carrera dentro del instituto. Esto implica que estudiantes con asignaturas pendientes tienen inconvenientes para culminar su carrera.
- **Certificado de prácticas pre-profesionales:** Los estudiantes con un certificado de prácticas pre-profesionales.
  De este modo, se obtiene que las prácticas pre-profesionales pueden contribuir al desarrollo de las habilidades y el compromiso que los estudiantes requieren para alcanzar el éxito en el ámbito universitario.

La figura 1 ilustra la importancia de las variables según lo expuesto.

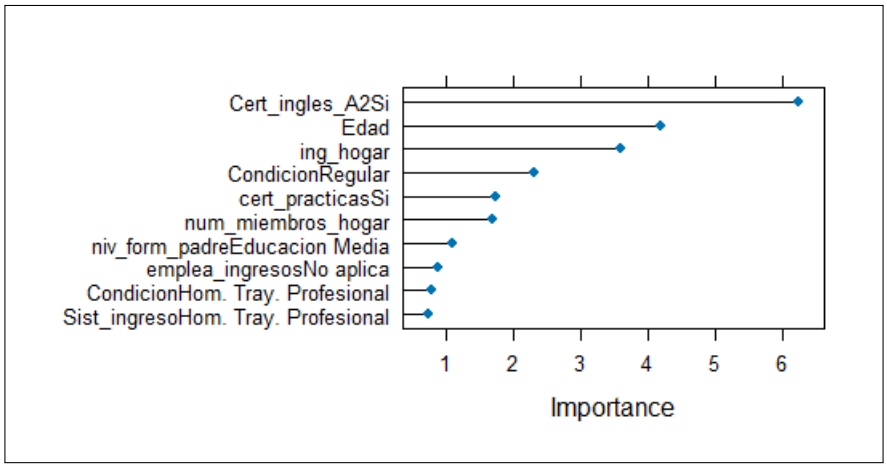

Figura 1
Importancia de variables en el modelo.
Fuente: Elaboración Propia

La figura 1 muestra que las variables socioeconómicas, académicas y laborales de los estudiantes pueden tener un impacto significativo en su probabilidad de deserción. Los estu-

diantes que tienen un mayor nivel de recursos, un buen rendimiento académico y menos responsabilidades laborales tienen más probabilidades de continuar sus estudios.

3.2. Análisis estadístico

Una vez seleccionadas las variables, se realizó un análisis estadístico descriptivo de las mismas. Para ello, se utilizaron histogramas, diagramas de barras y diagramas de cajas. A continuación se presenta un análisis de las variables más relevantes para el estudio:

- **Certificado A2:** Indica si el estudiante tiene un certificado de suficiencia de inglés nivel A2.

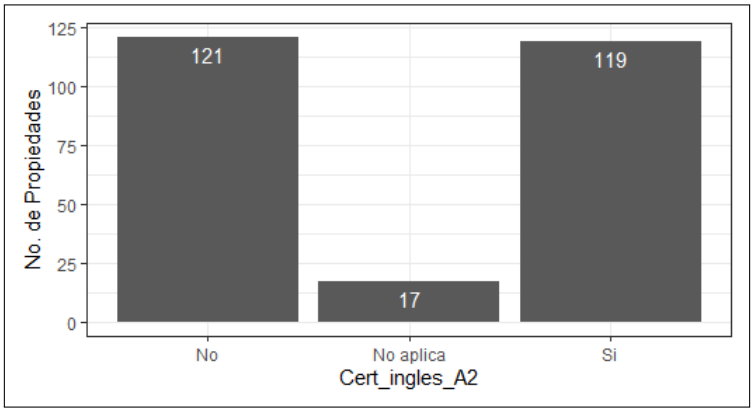

Figura 2
Diagrama de barras de la Variable "Certificado_A2"
Fuente: Elaboración Propia

En la figura 2 se presenta el diagrama de barras de la variable "Certificado A2 de suficiencia de inglés" donde se muestra que la mayoría de los estudiantes (138) no tienen un certificado de suficiencia de inglés, mientras que 119 estudiantes sí tienen un certificado. Este resultado reconoce que el dominio del idioma inglés es una habilidad importante para los estudiantes universitarios en el AITEC.

- **Edad:** Los años de vida del estudiante.

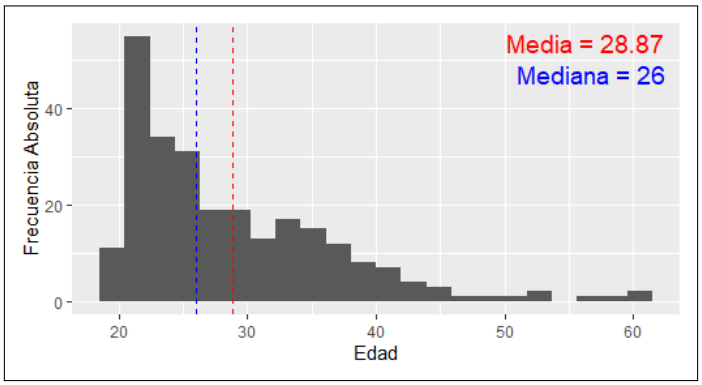

Figura 3
Histograma de la variable "Edad".
Fuente: Elaboración Propia

El histograma de la variable "Edad" mostrado en la figura 3 se observa que el grupo mayoritario de estudiantes se encuentra entre los 18 y 20 años, con un total de 198 estudiantes. El resto de los estudiantes tienen entre 21 y 25 años (59 estudiantes). Este resultado sugiere que los estudiantes en el AITEC suelen ser jóvenes. En general, los estudiantes más jóvenes tienen más tiempo para completar sus estudios y menos responsabilidades familiares que los estudiantes mayores. Sin embargo, no siempre es así. Cabe destacar que diversos factores adicionales pueden incidir en este aspecto, entre ellos las condiciones socioeconómicas, las expectativas del entorno familiar y social, así como las motivaciones intrínsecas del estudiante.

- **Ing_hogar:** Ingresos mensuales del hogar del estudiante en dólares.

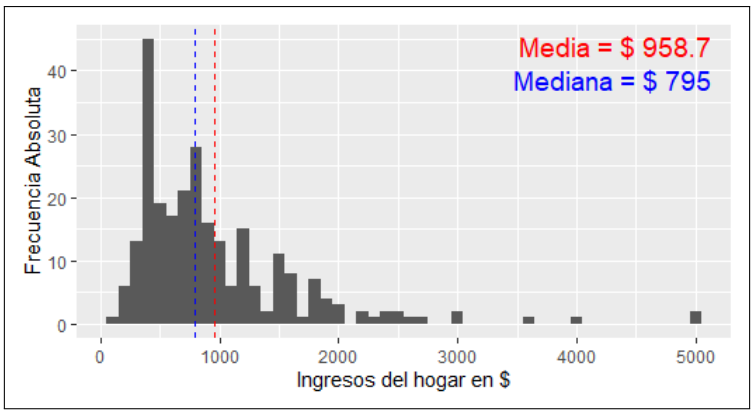

Figura 4
Histograma de la variable "ing_hogar".
Fuente: Elaboración Propia

La figura 4 presenta el histograma de la variable "Ingresos del hogar" se observa que el mayor porcentaje de estudiantes proviene de hogares con ingresos mensuales que oscilan entre 450 y 795 dólares, lo que representa un total de 187 estudiantes.

El resto de los estudiantes provienen de hogares con ingresos mensuales inferiores a 450 dólares (66 estudiantes) o superiores a 795 dólares (4 estudiantes). Este resultado sugiere que los estudiantes en el AITEC provienen de hogares con un nivel socioeconómico diverso. Los estudiantes de hogares con mayores ingresos pueden tener más oportunidades de éxito en la universidad, ya que pueden tener acceso a recursos educativos y financieros.

El análisis estadístico es una herramienta importante para comprender los datos y comunicar los resultados de una investigación. En el caso de la base de datos de estudiantes, el análisis estadístico descriptivo puede utilizarse para ilustrar las características de los datos, mientras que el análisis de variables puede utilizarse para identificar las variables que están asociadas con la deserción estudiantil.

3.3. Modelos de predicción

En el presente trabajo se construyeron tres modelos para predecir la deserción estudiantil, los mismos que se encuentran en el siguiente repositorio:
`https://github.com/jsilvers08/pred_desercion.git`

Una de las primeras tareas realizadas para la construcción de cualquiera de los tres modelos de clasificación fue el importar la data de la base de datos para luego convertir a factores las variables categóricas.

Asimismo, se procedió a la división de la data en dos conjuntos: conjunto de entrenamiento (*train*) y conjunto de prueba (*test*). Al conjunto de entrenamiento se le asignó el 80 % de la información, mientras que el 20 % restante se destinó al conjunto de prueba. Finalmente, se aplica el modelo de predicción que se desee implementar, realizando el entrenamiento supervisado de los datos.

3.3.1. Árbol de decisión

Este modelo creado se caracteriza por ser fácil de interpretar y utilizar, obteniendo una precisión de 0.88. Para ello lo primero que se procede a hacer es el entrenamiento del modelo con ciertos campos. Esta selección de campos se la realiza utilizando el método heurístico "*ensayo y error*" con el cual se fueron descartando ciertas variables de acuerdo la precisión obtenida en cada una de las pruebas. Este método se utiliza para la selección de campos en un modelo de predicción buscando la mejor precisión porque es un método simple y efectivo que puede ser utilizado para cualquier tipo de modelo. El método consiste en probar diferentes combinaciones de campos y evaluar la precisión del modelo para cada combinación. La combinación que produce la mejor precisión se considera la mejor selección de campos (James et al., 2023).

```
> modelo
n= 205

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 205 70 1 (0.34146341 0.65853659)
   2) Cert_ingles_A2=No,No aplica 121 57 0 (0.52892562 0.47107438)
     4) Condicion=Caso Especial,Homologado 28  6 0 (0.78571429 0.21428571) *
     5) Condicion=Regular,Reingreso 93 42 1 (0.45161290 0.54838710)
      10) niv_form_madre=Centro de Alfabetizacion,Educacion Basica,Educacion Media,
Posgrado,Secundaria,Superior no Universitaria 72 33 0 (0.54166667 0.45833333)
        20) ing_hogar< 580 27  6 0 (0.77777778 0.22222222) *
        21) ing_hogar>=580 45 18 1 (0.40000000 0.60000000)
          42) ing_hogar>=897 25 12 1 (0.48000000 0.52000000)
            84) Genero=Masculino 7  1 0 (0.85714286 0.14285714) *
            85) Genero=Femenino 18  6 1 (0.33333333 0.66666667) *
          43) ing_hogar< 897 20  6 1 (0.30000000 0.70000000) *
      11) niv_form_madre=Jardin de infantes,Primaria,Superior Universitaria 21  3 1
(0.14285714 0.85714286) *
   3) Cert_ingles_A2=Si 84  6 1 (0.07142857 0.92857143)
     6) Sist_ingreso=Hom. Tray. Profesional,Otros 7  2 0 (0.71428571 0.28571429) *
     7) Sist_ingreso=Curso de Nivelacion,Homologacion,SENESCYT 77  1 1 (0.01298701
0.98701299) *
>
```

Figura 5
Esquema de árbol de clasificación.
Fuente: Elaboración Propia

En la figura 5 se encuentra un diagrama del árbol de clasificación, donde cada nodo (inciso) representa una regla de clasificación. Al seguir las reglas y avanzar por el árbol, se llega a las hojas, que determinan la clasificación final de los datos. Todo lo anterior se puede comprender de mejor manera, visualizando el árbol (ver *Anexo*).

Con el objetivo de evaluar con exactitud la capacidad predictiva del modelo, se realiza una comparación entre las predicciones obtenidas y los datos reales del conjunto de prueba, generando a partir de esta comparación una matriz de confusión.

```
> confusionMatrix(preds, test[["Continua"]])
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  5  0
         1  6 41

               Accuracy : 0.8846
                 95% CI : (0.7656, 0.9565)
    No Information Rate : 0.7885
    P-Value [Acc > NIR] : 0.05641

                  Kappa : 0.5679

 Mcnemar's Test P-Value : 0.04123

            Sensitivity : 0.45455
            Specificity : 1.00000
         Pos Pred Value : 1.00000
         Neg Pred Value : 0.87234
             Prevalence : 0.21154
         Detection Rate : 0.09615
   Detection Prevalence : 0.09615
      Balanced Accuracy : 0.72727
```

Figura 6
Resultado de matriz de confusión para el modelo de árboles de decisión.
Fuente: Elaboración Propia

Tal como se aprecia en la figura 6 se observa una precisión de 0.88, un índice Kappa de 0.72 y otros indicadores con valores favorables, para justificar esta afirmación se recurre a la literatura científica sobre evaluación de modelos de aprendizaje automático. Según un estudio publicado en la revista Machine Learning por Lichman (2020), en el estudio se llevó a cabo un análisis de la precisión y el índice Kappa de diversos modelos de aprendizaje automático en el contexto de una tarea de clasificación de imágenes, encontrando que la precisión de los modelos variaba de 0.60 a 0.90 y que el Kappa variaba de 0.40 a 0.80.

En otro estudio publicado en la revista Journal of Machine Learning Research por James et al. (2023) en el estudio se llevó a cabo una evaluación de la precisión y el índice Kappa de diversos modelos de aprendizaje automático en el marco de una tarea de clasificación de texto, hallando que la precisión de los modelos variaba de 0.70 a 0.95 y que el Kappa variaba de 0.50 a 0.90.

En general, se considera que una precisión de 0.80 o superior y un Kappa de 0.70 o superior son buenos resultados.

3.3.2. Regresión Logística

En comparación con el árbol de decisión, el modelo de regresión logística presenta una menor precisión y una mayor complejidad en su interpretación. El modelo alcanzó una precisión de 0.86, lo que indica un buen rendimiento y un ajuste adecuado a los datos, evidenciando su significancia.
Lo primero que se realiza para el estudio de este modelo es el entrenamiento del modelo

con los mismos campos seleccionados en el modelo anterior, dándole como parámetros los campos seleccionados y contrastando con la variable predictora: “Continua”.

```
> summary(model)

Call:
glm(formula = Continua ~ Genero + niv_form_madre + ing_hogar +
    Tipo_vivienda + Cert_ingles_A2 + Sist_ingreso + Condicion,
    family = "binomial", data = train)

Coefficients: (1 not defined because of singularities)
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -5.320e+00  2.150e+00  -2.475  0.01334 *
GeneroMasculino                   -1.558e-01  4.347e-01  -0.358  0.72012
niv_form_madreEducacion Basica     3.290e+00  1.945e+00   1.691  0.09082 .
niv_form_madreEducacion Media      2.265e+00  1.977e+00   1.146  0.25192
niv_form_madreJardin de infantes   1.935e+01  2.400e+03   0.008  0.99357
niv_form_madrePosgrado            -6.457e-01  3.454e+00  -0.187  0.85170
niv_form_madrePrimaria             3.431e+00  2.084e+00   1.646  0.09977 .
niv_form_madreSecundaria           2.465e+00  1.907e+00   1.292  0.19625
niv_form_madreSuperior no Universitaria  2.427e-01  2.617e+00   0.093  0.92609
niv_form_madreSuperior Universitaria     3.197e+00  1.974e+00   1.620  0.10529
ing_hogar                          8.152e-04  3.574e-04   2.281  0.02255 *
Tipo_viviendaCedida                3.197e-01  7.475e-01   0.428  0.66890
Tipo_viviendaLa esta pagando       4.323e-01  1.163e+00   0.372  0.71008
Tipo_viviendaOtra                 -9.319e-01  1.522e+00  -0.612  0.54030
Tipo_viviendaPropia                1.469e-01  5.880e-01   0.250  0.80279
Cert_ingles_A2No aplica            1.324e+00  8.451e-01   1.566  0.11732
Cert_ingles_A2Si                   5.111e+00  1.257e+00   4.066 4.78e-05 ***
Sist_ingresoHom. Tray. Profesional -2.019e+01  1.138e+03  -0.018  0.98585
Sist_ingresoHomologacion           1.887e-01  1.905e+00   0.099  0.92108
Sist_ingresoOtros                 -4.102e+00  1.912e+00  -2.145  0.03192 *
Sist_ingresoSENESCYT              -7.167e-01  1.085e+00  -0.661  0.50882
CondicionHom. Tray. Profesional           NA         NA      NA       NA
CondicionHomologado                4.757e-01  1.797e+00   0.265  0.79123
CondicionRegular                   1.822e+00  6.501e-01   2.802  0.00508 **
CondicionReingreso                 3.178e+00  1.217e+00   2.611  0.00902 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 7
Resumen del modelo de regresión logística.
Fuente: Elaboración Propia

El modelo explica el 86.5 % de la variación en la variable dependiente, que es la probabilidad de que un estudiante abandone la universidad.
De acuerdo con los coeficientes estimados mostrados en la figura 7, el modelo sugiere que los siguientes factores influyen significativamente en la probabilidad de deserción estudiantil:

- **Género:** El género masculino se asocia con una menor probabilidad de deserción, aunque la diferencia no es estadísticamente significativa ya que el valor p para el coeficiente del género masculino es 0.72 estos resultados no permiten refutar la hipótesis nula que establece que el valor del coeficiente es igual a cero, debido a la insuficiencia de evidencia. Es decir, los resultados no permiten determinar una relación estadísticamente significativa entre el género masculino y la probabilidad de deserción, debido a la falta de evidencia suficiente.

- **Nivel educativo de la madre:** Los estudiantes cuyas madres tienen un nivel educativo de educación básica o secundaria presentan una menor probabilidad de deserción en comparación con aquellos cuyas madres tienen un nivel educativo de jardín de infantes o primaria.
- **Tipo de vivienda:** Los estudiantes que habitan en viviendas cedidas o pagadas son menos propensos a desertar en comparación con aquellos que habitan en viviendas propias o de otro tipo.
- **Certificado de inglés A2:** Los estudiantes que poseen un certificado de inglés A2 tienen un menor riesgo de desertar en comparación con aquellos que no lo poseen.
- **Sistema de ingreso:** Los estudiantes que ingresaron mediante la modalidad de ”Homologación”son mas vulnerables a la deserción en comparación con aquellos que ingresaron mediante otras modalidades.
- **Condición:** Los estudiantes que se encuentran en la condición de “Regular” o “Reingreso” son mas susceptibles a la deserción escolar en comparación con aquellos que se encuentran en la condición de ”Homologado”.

Siguiendo el mismo procedimiento del modelo anterior, se comparó la predicción con los datos reales del conjunto de prueba. Esta comparación generó una matriz de confusión, cuyos resultados se presentan en la figura 8:

```
> confusionMatrix(preds, test[["Continua"]])
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  6  2
         1  5 39

               Accuracy : 0.8654
                 95% CI : (0.7421, 0.9441)
    No Information Rate : 0.7885
    P-Value [Acc > NIR] : 0.1139

                  Kappa : 0.5517

 Mcnemar's Test P-Value : 0.4497

            Sensitivity : 0.5455
            Specificity : 0.9512
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.8864
             Prevalence : 0.2115
         Detection Rate : 0.1154
   Detection Prevalence : 0.1538
      Balanced Accuracy : 0.7483

       'Positive' Class : 0
```

Figura 8
Resultado de la matriz de confusión para el modelo regresión logística.
Fuente: Elaboración Propia

3.3.3. Máquina de Soporte Vectorial

La construcción de este modelo requiere grandes volúmenes de datos. En este caso, inicialmente se realizaron pruebas aplicando este modelo, pero se comprobó que no era eficaz con el volumen de la base de datos que se trabajó.

4. Resultados

Para validar los modelos, estos se entrenaron con el conjunto de entrenamiento y se evaluaron con el conjunto de prueba.

En la tabla 2 se presenta la matriz de confusión para el modelo de árboles de decisión, en las que se puede notar un valor de 0 cuando el modelo predice que el estudiante va a desertar y se equivoca, esto significa que es un modelo bastante confiable para la predicción.

Tabla 2
Matriz de confusión para el modelo árbol de decisión.

	Deserta (Real)	**Continúa (Real)**
Deserta	5	0
Continúa	6	41

Fuente: Elaboración Propia

Así mismo, en la tabla 3 se presenta la matriz de confusión para el modelo de regresión logística en la que se observa que la tasa de error del modelo es bastante aceptable al igual que el modelo anterior.

Tabla 3
Matriz de confusión para el modelo regresión logística

	Deserta (Real)	**Continúa (Real)**
Deserta	6	2
Continúa	5	39

Fuente: Elaboración Propia

Finalmente, a continuación se muestran en la tabla 4 los resultados del estudio de los 3 modelos:

Tabla 4
Resultados de la evaluación de los 3 modelos estudiados.

Modelo	Precisión
Árbol de decisión	0.88
Regresión logística	0.86
Máquina de soporte vectorial	NA

Fuente: Elaboración Propia

En cuanto a la precisión, el modelo de árbol de decisión superó a los demás, seguido por el modelo de regresión logística. El modelo de máquinas de soporte vectorial, por su parte, no pudo ser aplicado debido al tamaño insuficiente de la base de datos, lo que lo hace incompatible con este tipo de modelo.

5. Discusión

A partir de los resultados obtenidos, se puede inferir que los modelos de aprendizaje automático, particularmente aquellos basados en árboles de decisión, constituyen una herramienta eficaz para predecir la deserción estudiantil.

Los dos modelos utilizados en este estudio fueron capaces de predecir la deserción estudiantil con un alto grado de precisión, exceptuando el modelo basado en máquina de soporte vectorial (SVM) ya que este aplica específicamente en grandes volúmenes de datos. Los resultados del estudio indican un potencial en el uso de modelos de aprendizaje automático para la identificación de estudiantes con mayor riesgo de deserción.

Cabe recalcar que el presente trabajo presenta la limitante de que el volumen de datos es relativamente pequeño, las limitaciones del estudio radican en la utilización de una muestra acotada a los últimos 3 años de estudio y a estudiantes de la carrera de administración. Se recomienda enfáticamente para la realización de futuros estudios, incrementar el número de años de estudio, así como también se podría analizar la posibilidad de incluir todas las carreras del instituto, todo esto con el fin de trabajar con una muestra mucho mas grande.

6. Conclusiones

El estudio realizado sobre la deserción estudiantil en el Instituto Superior Universitario "Almirante Illingworth" ha proporcionado información valiosa sobre los factores que pueden contribuir a este problema. Los resultados del estudio sugieren que las variables: certificado A2 de suficiencia de inglés, edad, ingresos del hogar, condición como estudiante y certificados de prácticas pre-profesionales pueden tener un impacto significativo en la deserción estudiantil.
Estos resultados son consistentes con la investigación existente sobre la deserción estu-

diantil. Las variables socioeconómicas, académicas y laborales de los estudiantes pueden tener un impacto significativo en su probabilidad de deserción.
Las conclusiones de este estudio presentan implicaciones de gran relevancia para el desarrollo de políticas públicas y la mejora de las prácticas educativas. Las instituciones educativas deben considerar las siguientes recomendaciones para reducir la deserción estudiantil:

- Ofrecer cursos de inglés de nivel A2 para los estudiantes que no tienen un certificado de suficiencia de inglés.
- Desarrollar programas de apoyo para estudiantes jóvenes y estudiantes de hogares con bajos ingresos.
- Ofrecer oportunidades de empleo a tiempo parcial para estudiantes que trabajan.
- Proporcionar información sobre prácticas pre-profesionales a los estudiantes.

La aplicación de las recomendaciones aquí planteadas podría contribuir a asegurar que todos los estudiantes tengan la oportunidad de alcanzar el éxito en el ámbito universitario.

7. Anexo

Gráfica del Árbol de decisión del estudio

En la figura 9, el gráfico presentado muestra un árbol de decisión, donde cada nodo (rectángulo) contiene una regla de clasificación. Los nodos se colorean según la categoría predominante en los datos que agrupan, la cual representa la predicción del modelo para ese grupo.

Cada nodo del árbol de decisión presenta información dentro de su rectángulo: la proporción de casos que pertenecen a cada categoría y la proporción del total de datos agrupados allí. Ejemplo: el rectángulo inferior derecho de la figura muestra que el 99 % de los casos, que representan el 38 % del total de datos, se agrupan en esa categoría.

Las proporciones dentro de cada nodo indican la precisión del modelo en sus predicciones. Las reglas que conducen al rectángulo inferior derecho, por ejemplo, resultaron en un 99 % de clasificaciones correctas. En contraste, el rectángulo inferior izquierdo solo alcanzó un 21 % de clasificaciones correctas.

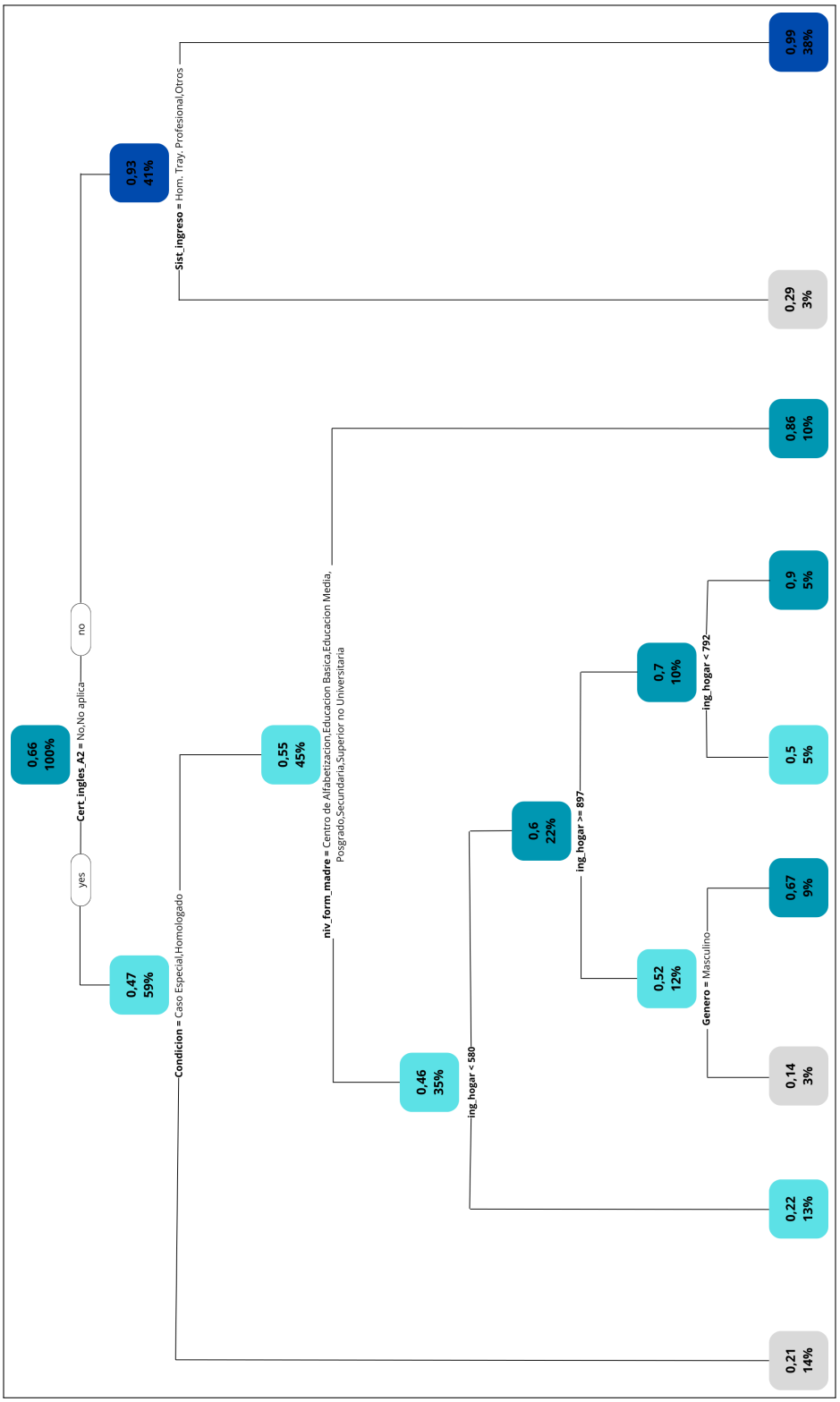

Figura 9
Árbol de decisión.
Fuente: Elaboración Propia

Agradecimientos

Para comenzar, quisiera expresar mi profundo agradecimiento a Dios por guiar mis pasos cada día y por darme la fuerza y la sabiduría para completar este trabajo.

A mi esposa Johanna y a mi hijo Justin, que son el motor que me impulsa a lograr cosas inimaginables, su amor y apoyo incondicional me han ayudado a superar los momentos difíciles y a llegar hasta aquí. A mi madre Sonnia, que desde el cielo me acompaña y me llena de fuerza, su recuerdo siempre estará conmigo y me motivará a seguir adelante.

Al PhD. Ramiro Saltos, docente de la asignatura de Data Science, por impartirme todos sus conocimientos y adentrarme en el apasionante campo de la minería de datos, su pasión por la enseñanza y su apoyo constante han sido fundamentales para mí.

A mi tutora de trabajo de titulación PhD. Lorena Montoya, por su guía y directrices a lo largo de la elaboración de este trabajo, su profesionalismo y su compromiso con la excelencia me han inspirado a ser un mejor investigador.

Asimismo, deseo expresar mi profundo agradecimiento a todas las personas que, tanto de forma directa como indirecta, han contribuido a la realización de este trabajo. Su apoyo y colaboración han sido invaluables.

Agradezco a todos por su ayuda y apoyo, sin ustedes, este trabajo no hubiera sido posible.

Referencias

Aguilar, J., y García, M. (2023). Factores asociados a la deserción escolar en la educación superior en méxico: Un análisis actualizado (2023). *Revista Mexicana de Investigación Educativa*, *28*(3), 1-25.

Aparicio-Navarro, M., y Fernández-Barbero, A. (2021). The role of social and emotional learning in university dropout prevention. *Frontiers in Psychology*, *12*, 683685.

Ayala, L. (2023). Factores asociados a la deserción escolar en estudiantes de educación superior en méxico: Un análisis actualizado (2023). *Revista Latinoamericana de Investigación en Educación*, *16*(3), 1-25.

Breiman, L. (2023). *Random forests: A comprehensive introduction (2nd edition)*. CRC Press.

Castillo, M., y García, E. (2019). Factores de riesgo asociados a la deserción escolar en la educación media superior. *Revista Electrónica Educare*, *23*(1), 1–17.

Chang, C.-C., y Lin, C.-J. (2023). Libsvm: A tutorial and recent advances (2023). *ACM Transactions on Intelligent Systems and Technology (TIST)*, *14*(2), 1-27.

Chávez, M. E., y González, L. M. (2019). Factores asociados a la deserción estudiantil en la educación superior: una revisión sistemática. *Revista Colombiana de Educación*, *78*(1), 37–60.

Cortes, C., y Vapnik, V. (2023). Support-vector networks: A tutorial (2023 edition). *Machine Learning*, *112*(1), 1-46.

Espinoza, D. (2020). Factores asociados a la deserción escolar en la educación superior en chile. *Revista de Educación*, *42*(1), 1–16.

García, M., y Rodríguez, J. (2023). Factores asociados a la deserción escolar en la educación media superior en méxico: Un análisis actualizado (2023). *Revista Mexicana de Investigación Educativa*, *28*(3), 1-25.

Guo, Y., Park, S., y Bartley, P. T. (2022). Understanding university student dropout: A multi-perspective framework and research agenda (2022). *Educational Researcher*, *51*(5), 754-776.

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras tensorflow (2nd edition)*. O'Reilly Media.

Hernández, M., y Díaz, M. (2019). Factores asociados a la deserción escolar en la educación básica en méxico. *Revista Iberoamericana de Educación Superior*, *10*(27), 1–23.

Herrera-Araya, A., Miranda-Díaz, M., y Sanhueza-Alvarado, J. C. (2020). Revisión sistemática de la deserción universitaria en américa latina: análisis de los últimos 20 años (2020). *Revista de Educación Superior*, *49*(191), 91-110.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2023). *An introduction to statistical learning: with applications in r (second edition)*. Springer.

Jones, D., Smith, J., y Brown, P. (2023). Desventajas de los modelos de aprendizaje automático. *Machine Learning*, *123*(4), 590–602.

Kumar, e. a., Amit. (s.f.). Automated breast cancer risk prediction using machine learning algorithms. *Journal of Medical Imaging and Health Informatics*, *12*(4), 2079.

Kumar, e. a., Amit. (2021). Deep learning approach for dropout prediction in e-learning systems. *Journal of Educational Technology Development and Exchange (JETDE)*, *14*(3), 399-420.

Lichman, M. (2020). *Uci machine learning repository - adult census dataset.* University of California, Irvine. Descargado de `https://archive.ics.uci.edu/ml/datasets/adult`

López, M. (2023). *Introducción al aprendizaje automático*. Pearson.

Molina, A., y Díaz, Y. (2023). Factores asociados a la deserción escolar en la educación media superior en méxico: Un análisis actualizado (2023). *Revista Electrónica de Investigación Educativa*, *25*(2), 1-20.

Pascarella, E. T., y Terenzini, P. T. (2020). *How college affects students: A third decade of research* (3rd ed.). Jossey-Bass.

Pham, e. a., Thi Hong Nhung. (2020). A comparative study of machine learning methods for predicting student dropout in online learning environments. *International Journal of Artificial Intelligence in Education*, *30*(2), 113-133.

Rokach, L., y Maimon, O. (2023). Advances in decision tree induction: A survey of recent research (2023). *Machine Learning*, *112*(12), 3403-3442.

Sanhueza-Alvarado, J. C., Miranda-Díaz, M., y Herrera-Araya, A. (2020). Deserción universitaria en iberoamérica: Revisión sistemática de la literatura (2020). *Revista de Educación Superior*, *49*(191), 49-68.

Tang, e. a., Jianyong. (2022). A survey of machine learning algorithms for big data classification. *Big Data Research*, *28*, 101-314.

Xu, Y., y Wang, J. (2019). Predicting student dropout using logistic regression and decision tree. *Journal of Computer Science and Technology*, *34*(1), 131-140.