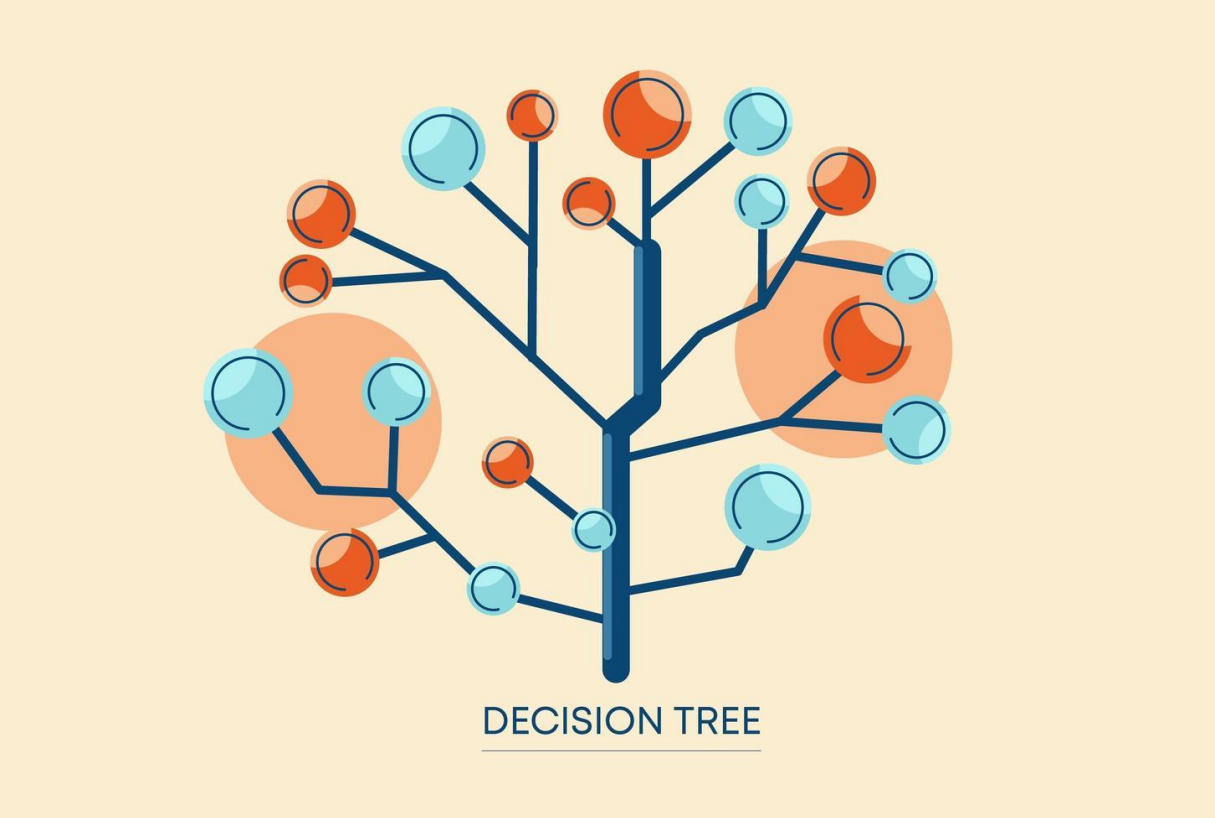
INSTITUTO TECNOLÓGICO BELTRÁN
Centro de Tecnología e Innovación

Informe Árbol de decisión

"Predicción de abandono en estudiantes terciarios"



Integrantes = Nicolas Mesquiatti, Lucas Oviedo, Marco Medina, Coral Tolazzi, Ariel Colatto, Cristian Monzón
Docente= Yanina Scudero
Materia = Aprendizaje Automático
Fecha = Octubre 2025

Objetivo del trabajo

b) **Predicción de abandono en estudiantes terciarios – En Equipo**

El Instituto Tecnológico Beltrán desea identificar a los estudiantes con riesgo de abandonar la carrera para implementar estrategias de acompañamiento. Para ello, se ha recopilado un conjunto de datos históricos de estudiantes que incluye las siguientes variables:

- Edad
- Género
- Carrera elegida
- Promedio del primer cuatrimestre
- Cantidad de materias aprobadas en el primer año
- Cantidad de materias desaprobadas
- Asistencia promedio (%)
- Situación laboral (trabaja/no trabaja)
- Distancia desde el domicilio al instituto
- Participación en tutorías o actividades extracurriculares
- Estado final del estudiante (abandonó / continúa)

Objetivo: Construir un modelo de árbol de decisión que permita predecir si un estudiante abandonará la carrera, en base a los datos de su cursada durante el primer año.

Tareas:

1. Analizar y preparar los datos.
2. Entrenar un árbol de decisión utilizando un conjunto de entrenamiento.
3. Evaluar el modelo con un conjunto de prueba.
4. Interpretar el árbol generado: ¿Qué variables parecen ser más importantes para predecir el abandono?
5. Proponer recomendaciones para la institución basadas en los resultados del modelo.

Descripción del dataset

El conjunto de datos contiene información de estudiantes del Instituto Tecnológico Beltrán, con el objetivo de predecir el abandono académico. Cada fila representa a un estudiante, y se incluyen variables tanto académicas como personales.

Cantidad de registros: 100 estudiantes.

Propósito del análisis:

- Identificar patrones que expliquen el abandono.
- Entrenar un árbol de decisión para predecir el riesgo de deserción.

Preparación de datos

Codifique el atributo EstadoFinal del csv como Binarias

- Continúa → 0
- Abandonó → 1

Algunas variables las separe para hacer un análisis por grupo de estas

Edad: segmentada en tres grupos (Joven: ≤30, Adulto: 31–50, Mayor: >50)

Promedio: dividido en bajo (<6), medio (6–8) y alto (>8)

Materias aprobadas/desaprobadas: agrupadas en pocas, moderadas y muchas

Carreras: Las separe por carreras para identificar en cuál de las carreras hay mayor abandono

DistanciaDomicilio(km) = cerca (0–5.99 km), medio (6 a 12km) y lejos (más de 13km)

AsistenciaPromedio = *bajas (0–59% De inasistencia), medio (60-80% de inasistencia) y alto más de 80% de inasistencia*

Calculo Manual De entropía y Ganancias

fórmula que utilice para cada atributo:

Entropía = $-(p_1 * \log_2(p_1) + p_0 * \log_2(p_0))$

Ganancia de información = Entropía total − Entropía ponderada del atributo

<u>Entropía Original del conjunto</u>

Entropía Original

Total, estudiantes 100

- Abandonan: 53 → 53/100= -0,915
- Continúan: 47 → 47/100= -1,083

Log2(0,53) = -0.915

Log2(0,47) = -1,083

H=−(0.47·−1.083+0.53·−0.915) =−(−0.509+−0.485) =0.994

Entropía de cada atributo manualmente

a) Edad

<u>Subgrupo 1= Joven ≤30 ()</u>

29 ejemplos

• 13 "Abandono" = 13/29 = 0,448

• 16 "Continua" = 16/29 = 0,552

• Log2(0,448) =-1,152

• Log2(0,552) =-0,855

Entropía del grupo = -(0,448 * (-1,152) + 0,552 * (-0,855)) =

-(-0,516 –0,472) =

0,516 + 0,472 = **0,988**

Subgrupo 2: Adulto 31–50

42 ejemplos

• 20 “Abandono” = 20/42 = 0,476

• 22 “Continua” = 22/42 = 0,524

• Log2(0,476) =-1,067

• Log2(0,0,524) = -0,930

Entropía del grupo = −(0,476*log2(0,476) +0.524*log2(0.524) =

-(0,476 * (-1,067) + 0,524 * (-0,930)) =

-(-0,508 -0,487)

0,509 + 0,487 =**0,995**

Subgrupo 3: Mayor >50)

29 ejemplo

• 16 “Abandono” = 16/29 = 0,552

• 13 “Continua” = 13/29 = 0,448

• *Log2(0,552) = -0,855*

• *Log2(0,448) = -1,152*

*Entropía del grupo = -(0,552 * (-0,855) + 0,448 * (-1,152)) =*

-(-0,472 – 0,516)

*0,472 +0,516= **0,988***

*Entropía **ponderada del tributo edad***

H(Edad)=29/100(0,988) +42/100(0.995) +20/100(0,988) =0,9909

Ganancia de información del atributo EDAD

Ganancia=Htotal−Hponderada=0.994−0.9909=0.0031

En Python me dio un resultado más redondeado, ya que usan los valores completos y en las cuentas manuales yo los simplifico.

Ej

Log2(0,49) Manualmente: -1,022 Simplifica el calculo

Log2(0,49) Python: -1,021651247…

B) Genero

Femenino

49 ejemplos

“24” Abandono: 24/49 = 0,4898

“25” Continua: 25/49 = 0,5102

Log2(0,4898): -1,0297

Log2(0,5102): -0,9709

Hf=−(0.4898·log2(0.4898) +0,5102·log2(0.5102)) ≈0.9997

-(0,4898 * (-1,0297) + 0,5102 *(-0,9709)) =

-(-0,5043 + (-0,4953) = **0,9996**

Masculino

51 ejemplos

“25” Abandono = 25/51 = 0,4902

“26” Continua = 26/51 = 0,5098

Log2(0,4902): -1.0286

Log2(0,5098): -0,972

Hm=−(0.4902·log2(0.4902) +0.5098·log2(0.5098)) ≈0.9931

*-(0,4902 * (-1,0286) + 0,5098 *(-0,972)) =*

-(-0,5042 + (-0,4955) = ***0,9997***

Entropía **Ponderada del atributo Genero**

Hponderada=0.49·0.9996+0.51·0.9997

0,4898+ 0,5098 = **0,9996**

Ganancia de Información

Ganancia=Htotal−Hponderada=0.994−0.9996=−0.0056

c)PromedioPrimerCuatrimestre

Grupo Bajo Promedio < 6

27 ejemplos

Abandono 13/27: 0,4815

Continua 14/27: 0,5185

Hbajo=−(0.4815·log⁡2(0.4815) +0.5185·log⁡2(0.5185)) ***≈0.9990***

Grupo Medio Promedio Entre 6-8

37 ejemplos

Abandono 18/37: 0,4865

Continua 19/37: 0,5135

Hmedio=−(0.4865·log⁡2(0.4865) +0.5135·log⁡2(0.5135)) ***≈0.9995***

Grupo Alto Promedio > 8

36 ejemplos

Abandono 16/36: 0,4444

Continua 20/36: 0,5555

Halto=−(0.4444·log⁡2(0.4444) +0.5556·log⁡2(0.5556)) ***≈0.9799***

Entropía ***Ponderada del atributo PromedioPrimerCuatrimestre***

Hponderada=0.27·0.9990+0.37·0.9995+0.36·0.9799

Ganancia De información PromedioPrimerCuatrimestre

Ganancia=Htotal−Hponderada=0.994−0.9923=0.0017

D)Atributo categórico: Carrera

Grupo TECNICATURA SUPERIOR EN ADMINISTRACIÓN DE PYMES:

- *Total: 12 estudiantes*
- *Abandonaron: 5/12*
- *Continuaron: 7/12*
- *H=−0.417log2 (0.417) −0.583log2 (0.583) ≈0.980*
- ***Entropía ≈ 0.980***

Grupo TECNICATURA SUPERIOR EN ADMINISTRACIÓN CONTABLE:

- *Total: 15*
- *Abandonaron: 8*
- *Continuaron: 7*
- *H=−0.533log2 (0.533) −0.467log2 (0.467) ≈0.997*
- ***Entropía ≈ 0.997***

Grupo TECNICATURA SUPERIOR EN RADIOLOGIA:

- *Total: 13*
- *Abandonaron: 6*
- *Continuaron: 7*
- *H=−0.462log2 (0.462) −0.538log2 (0.538) ≈0.996*
- ***Entropía ≈ 0.996***

Grupo TECNICATURA SUPERIOR EN DISEÑO INDUSTRIAL:

- *Total: 12*
- *Abandonaron: 8*
- *Continuaron: 4*
- *H=−0.667log2 (0.667) −0.333log2 (0.333) ≈0.918*
- ***Entropía ≈ 0.918***

Grupo TECNICATURA SUPERIOR EN ENFERMERÍA:

- *Total: 11*
- *Abandonaron: 5*
- *Continuaron: 6*
- *H=−0.455log2 (0.455) −0.545log2 (0.545) ≈0.994*
- ***Entropía ≈ 0.994***

Grupo TECNICATURA SUPERIOR EN CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL':

- *Total: 12*
- *Abandonaron: 7*
- *Continuaron: 5*
- *H=−0.583log2 (0.583) −0.417log2 (0.417) ≈0.980*
- ***Entropía ≈ 0.980***

Grupo TÉCNICATURA SUPERIOR EN HIGIENE Y SEGURIDAD EN EL TRABAJO:

- *Total: 6*
- *Abandonaron: 4*
- *Continuaron: 2*
- *H=−0.667log2 (0.667) −0.333log2 (0.333) ≈0.918*
- ***Entropía ≈ 0.918***

Grupo TECNICATURA SUPERIOR EN ANÁLISIS DE SISTEMAS:

- *Total: 10*
- *Abandonaron: 6*
- *Continuaron: 4*
- *H=−0.600log2 (0.600) −0.400log2 (0.400) ≈0.971*
- ***Entropía ≈ 0.971***

Grupo TECNICATURA SUPERIOR EN COMUNICACIÓN MULTIMEDIAL:

- *Total: 9*
- *Abandonaron: 4*
- *Continuaron: 5*
- *H=−0.444log2 (0.444) −0.556log2 (0.556) ≈0.991*
- ***Entropía ≈ 0.991***

Entropía ***Ponderada del atributo carrera***

0.1176+0.1496+0.1295+0.1102+0.1093+0.1176+0.0551+0.0971+0.0892=0.8752

- ***Entropía ponderada*** *≈ 0.8752*

Ganancia de información del atributo Carrera

Ganancia=Htotal−Hponderada=0.994−0.8763=0.1177

E) CantMateriasAprobadasPrimerCuatrimestre

Subgrupo 1= Pocas ≤ 4

44 ejemplos

• 20 “Abandono” = 20/44 = 0,4545

• 24 “Continua” = 24/44 = 0,5455

*Hpocas=−(0.4545 ·log⁡2(0.4545) +0.5455 ·log⁡2(0.5455)) ≈**0.9865***

Subgrupo 2: Moderadas 5–6

50 ejemplos

Abandonaron: 25/50 = 0,5

Continuaron: 25/50 = 0,5

*Hmoderadas=−(0.5 ·log⁡2(0.5) +0.5 ·log⁡2(0.5)) =**1.000***

Subgrupo 3: Muchas >= 7)

6 ejemplo

• 2 “Abandono” = 2/6 = 0,3333

• 4 “Continua” = 4/6 = 0,6667

*Hmuchas=−(0.3333 ·log⁡2(0.3333) +0.6667 ·log⁡2(0.6667)) ≈**0.9183***

Entropía **ponderada del tributo CantMateriasAprobadasPrimerCuatrimestre**

Hponderada=0.44·0.9865+0.50·0.9988+0.06·0.9183 = **0,9886**

Ganancia de información del atributo CantMateriasAprobadasPrimerCuatrimestre

Ganancia=Htotal−Hponderada=0.994−0.9886=**0.0054**

F) CantMateriasDesaprobadasPrimerCuatrimestre

Subgrupo 1: bajas (0–3 desaprobadas)

36 ejemplos

Abandonaron: 18/36 = 0,5

Continuaron: 18/36 = 0,5

*Hbajas=−(0.5 ·log⁡2(0.5) +0.5 ·log⁡2(0.5)) =**1.000***

Subgrupo 2: medias (4–6 desaprobadas)

53 ejemplos

Abandonaron: 26/53 = 0,5 = 0,4906

Continuaron: 27/53 = 0,5 = 0,5094

*Hmedias=−(0.4906 ·log⁡2(0.4906) +0.5094 ·log⁡2(0.5094)) ≈**0.9936***

Subgrupo 3: altas (>7 desaprobadas)

11 ejemplos

Abandonaron: 6/11 = 0,5 = 0,4906 = 0,5455

Continuaron: 5/11 = 0,5 = 0,5094 = 0,4545

*Haltas=−(0.5455 ·log⁡2(0.5455) +0.4545 ·log⁡2(0.4545)) ≈**0.9940***

Entropia ***Ponderada CantMateriasDesaprobadasPrimerCuatrimestre***

Hponderada=0.36·0.9978+0.53·0.9936+0.11·0.9940=**0,9951**

Ganancia de información del atributo CantMateriasDesaprobadasPrimerCuatrimestre

Ganancia=Htotal−Hponderada=0.994−0.9951= -0.0011

G) AsistenciaPromedio(%)

Subgrupo 1: bajas (0–59% De inasistencia)

39 ejemplos

Abandonaron: 19/39 = 0,4872

Continuaron: 20/39 = 0,5128

*Hbaja=−(0.4872 ·log⁡2(0.4872) +0.5128 ·log⁡2(0.5128)) ≈**0.9957***

Subgrupo 2: medias (60–80% De inasistencia)

31ejemplos

Abandonaron: 15/31 = 0,4839

Continuaron: 16/31 = 0,5161

*Hmedia=−(0.4839 ·log2(0.4839) +0.5161 ·log2(0.5161)) ≈**0.9932***

Subgrupo 3: alta (>=81% De inasistencia)

30ejemplos

Abandonaron: 15/30= 0,5

Continuaron: 15/30 = 0,5

*Halta= −(0.5 ·log2(0.5) +0.5 ·log2(0.5)) =**1.000***

*Entropia **Ponderada AsistenciaPromedio(%)***

Hponderada=0.39 ·0.9957+0.31 ·0.9932+0.30 ·1.000= 0,9962

Ganancia de información del atributo AsistenciaPromedio(%)

Ganancia=Htotal−Hponderada=0.994−0.9962= -0.0022

H) trabaja/NoTrabaja

Valor "Sí" (trabaja)

- Total: 43 estudiantes
- Abandonaron: $21/43 = 0{,}4884$
- Continuaron: $22/43 = 0{,}5116$

$HSı' = -(0.4884 \cdot \log_2(0.4884) + 0.5116 \cdot \log_2(0.5116)) \approx$**0.9996**

Valor "No" (no trabaja)

- Total: 57 estudiantes
- Abandonaron: 28/57 = 04912
- Continuaron: 29/57 = 0,5088

Entropía **Ponderada Del atributo Trabaja/No trabaja**

HNo=−(0.4912·log2(0.4912) +0.5088·log2(0.5088)) **≈0.9944**

Ganancia de información del atributo Trabaja/No trabaja

Ganancia=Htotal−Hponderada=0.994−0.9967=-0.0027

I) DistanciaDomicilioAlInstituto(Kms)

Grupo "cerca" (0–5.99 km)

- Total: 24 estudiantes
- Abandonaron: 12/24 = 0,5
- Continuaron: 12/24 = 0,5

Hcerca=−(0.5·log2(0.5) +0.5·log2(0.5)) =**1.000**

Grupo "media" (6–12.99 km)

- Total: 36
- Abandonaron: 18/30 = 0,5
- Continuaron: 18/30 = 0,5

Hmedia=−(0.5·log2(0.5) +0.5·log2(0.5)) =**1.000**

Grupo "lejos" (≥ 13 km)

- Total: 40
- Abandonaron: 19/40 = 0,475
- Continuaron: 21/40 = 0,525

Hlejos=−(0.475·log2(0.475) +0.525·log2(0.525)) **≈0.9837**

Entropia **ponderada del atributo DistanciaDomicilioAlInstituto(Kms)**

Hponderada=0.24·0.9950+0.36·0.9978+0.40·0.9837=**0,9915**

Ganancia de informacion del atributo DistanciaDomicilioAlInstituto(Kms)

Ganancia=Htotal−Hponderada=0.994−0.9915=0.0039

J) ActividadesExtracurriculares(Estudio)

Valor "Sí"

- *Total: 53 estudiantes*
- *Abandonaron: 26/53 = 0,4906*
- *Continuaron: 27/53 = 0,5094*

HSı´=−(0.4906·log⁡2(0.4906) +0.5094·log⁡2(0.5094)) ≈0.9874

Valor "No"

- *Total: 47 estudiantes*
- *Abandonaron: 23/47 = 0,4894*
- *Continuaron: 24/47 = 0,5106*

*HNo=−(0.4894·log⁡2(0.4894) +0.5106·log⁡2(0.5106)) ≈**0.9997***

Entropia **Ponderada del atributo ActividadesExtracurriculares(Estudio)**

Hponderada=0.5233+0.4693=**0.9926**

Ganancia de Informacion para El atributo ActividadesExtracurriculares(Estudio)

Ganancia=0.994−0.9926=0.0014

RESUMEN DE LAS ENTROPIAS Y GANANCIAS DE INFORMACION PARA CADA ATRIBUTO

Atributo	Entropía Ponderada	Ganancia de Información
Carrera	0.8763	**0.1177**
Edad	0.9909	0.0031
PromedioPrimerCuatrimestre	0.9923	0.0017
CantMateriasAprobadasPrimerCuatrimestre	0.9886	0.0054
CantMateriasDesaprobadasPrimerCuatrimestre	0.9951	-0.0009
AsistenciaPromedio(%)	0.9962	-0.0022
trabaja/NoTrabaja	0.9967	-0.0007
DistanciaDomicilioAlInstituto(Kms)	0.9915	0.0039
ActividadesExtracurriculares(Estudio)	0.9932	0.0008
Género	0.9996	−0.0036

Comparación entre cálculos manuales y Python

Durante el análisis, se calcularon manualmente las ganancias de información para cada atributo del conjunto de datos. Posteriormente, se construyó un árbol de decisión en Python utilizando la misma lógica de discretización. Sin embargo, se observaron algunas diferencias en los valores obtenidos.

Atributo	Ganancia Manual	Ganancia Python

carrera	0.1197	0.0742
edad	0.0051	0.0215
PromedioPrimerCuatrimestre	0.0037	0.0169
DistanciaDomicilioAlInstituto(Kms)	0.0059	0.0159
CantMateriasDesaprobadasPrimerCuatrimestre	0.0009	0.0068
CantMateriasAprobadasPrimerCuatrimestre	0.0074	0.0060
AsistenciaPromedio(%)	0.0012	0.0013
trabaja/NoTrabaja	0.0007	0.0001
ActividadesExtracurriculares(Estudio)	0.0028	0.0000
genero	−0.0036	0.0069

Árbol Manual

Diagrama del árbol

CARRERA
- Diseño Industrial / Higiene y Seguridad → ABANDONA
- Administración de PYMES / Enfermería → CONTINÚA
- Otras carreras
  - PROMEDIO
    - < 6 → ABANDONA
    - > 8 → CONTINÚA
    - 6–8
      - EDAD
        - ≤ 30
          - Materias Aprobadas ≤ 4 → ABANDONA
          - Materias Aprobadas > 4 → CONTINÚA
        - > 30 → CONTINÚA

Justificación de cada decisión

- **Nivel 1: CARRERA** Las carreras con mayor tasa de abandono (Diseño Industrial e Higiene y Seguridad) se clasifican directamente como “abandona”. Las carreras con baja tasa (Administración de PYMES y Enfermería) se clasifican como “continúa”. Las demás se analizan con atributos secundarios.
- **Nivel 2: PROMEDIO** Promedios bajos (<6) se asocian a abandono. Promedios altos (>8) se asocian a continuidad. Promedios medios (6–8) requieren un análisis adicional.
- **Nivel 3: EDAD** Estudiantes jóvenes (≤30) con pocas materias aprobadas (≤4) tienden a abandonar. Si tienen más materias aprobadas, se clasifican como

"continúa". Adultos y mayores (>30) se clasifican como "continúa" por mayor estabilidad académica.

Evaluación del árbol

```
4. EVALUACIÓN DEL MODELO:
   • Precisión en conjunto de prueba: 0.6000
   • Reporte de clasificación:

              precision    recall  f1-score   support

    Continúa       0.55      0.67      0.60         9
    Abandona       0.67      0.55      0.60        11

    accuracy                           0.60        20
   macro avg       0.61      0.61      0.60        20
weighted avg       0.61      0.60      0.60        20
```

El árbol de decisión manual fue evaluado sobre un conjunto de prueba compuesto por 20 estudiantes (20% del total). Las métricas obtenidas fueron:

- **Precisión global**: 60.0%
- **F1-score promedio**: 60.0%
- **Recall para clase "Continúa"**: 67%
- **Recall para clase "Abandona"**: 55%
- **Macro promedio de precisión y recall**: 61%

Estas métricas indican que el modelo tiene un rendimiento moderado, con una ligera tendencia a identificar mejor a los estudiantes que continúan que a los que abandonan.

Matriz de confusión

```
• Matriz de Confusión:
  Verdaderos Positivos: 6 (Abandonan - correcto)
  Falsos Positivos:    3 (Continúan - predicho abandono)
  Verdaderos Negativos: 6 (Continúan - correcto)
  Falsos Negativos:    5 (Abandonan - predicho continuación)
```

- **Verdaderos positivos (abandona correctamente)**: 6
- **Falsos positivos (predice abandono, pero continúa)**: 3
- **Verdaderos negativos (continúa correctamente)**: 6
- **Falsos negativos (predice continuidad, pero abandona)**: 5

El modelo logra identificar correctamente a más de la mitad de los estudiantes en ambas clases, aunque presenta un margen de error que puede ser crítico en contextos institucionales

Interpretación del árbol

El árbol fue construido priorizando la variable `carrera`, que mostró la mayor ganancia de información. Las decisiones se refinan con `PromedioPrimerCuatrimestre` y `edad`, lo que permite identificar patrones claros de abandono:

- Carreras como Diseño Industrial e Higiene y Seguridad presentan alta tasa de abandono.
- Estudiantes con promedio bajo (<6) o jóvenes con pocas materias aprobadas (≤4) también muestran riesgo elevado.

```
5. INTERPRETACIÓN DEL ÁRBOL:
   • Variable más importante: CARRERA
   • Variables secundarias: PROMEDIO y EDAD
   • El árbol identifica patrones claros de abandono
```

Variables más influyentes

El análisis de ganancia de información permitió identificar las variables con mayor poder predictivo sobre el abandono estudiantil. Las más influyentes fueron:

- **Carrera**: con la mayor ganancia de información (0.0742), esta variable permite segmentar directamente a los estudiantes según su trayectoria académica.
- **Promedio del primer cuatrimestre**: refleja el rendimiento académico inicial, clave para detectar riesgo de abandono.
- **Edad**: aporta contexto sobre la etapa de vida del estudiante, influenciando su compromiso y continuidad.
- **Materias aprobadas**: complementa el promedio, indicando el progreso real en la cursada.

Patrones detectados

El modelo identificó varios patrones relevantes:

- **Carreras con alta tasa de abandono**: Diseño Industrial y Higiene y Seguridad en el Trabajo presentan mayor probabilidad de deserción, posiblemente por exigencia técnica o falta de orientación vocacional.
- **Rendimiento académico bajo (<6)**: se asocia fuertemente con abandono, especialmente en el primer cuatrimestre.
- **Estudiantes jóvenes (≤30 años) con pocas materias aprobadas (≤4)**: muestran mayor vulnerabilidad, lo que sugiere la necesidad de acompañamiento temprano.
- **Carreras como Administración de PYMES y Enfermería**: presentan mayor estabilidad, con predominancia de estudiantes que continúan.

Recomendaciones

Acciones institucionales basadas en recomendaciones

Gestión institucional

- Enfocar recursos en carreras con alta tasa de abandono: *Diseño Industrial* y *Higiene y Seguridad*.

- Implementar programas de acompañamiento desde el primer cuatrimestre en dichas carreras.
- Reforzar el proceso de orientación vocacional para mejorar la elección inicial de carrera.

Apoyo académico

- Estudiantes con promedio < 6 deben ser priorizados para tutorías obligatorias.
- Jóvenes (≤30 años) con pocas materias aprobadas (≤4) requieren seguimiento personalizado.
- Monitorear asistencia y desempeño en el primer cuatrimestre como indicadores tempranos.

Conclusión Árbol manual

La lógica jerárquica basada en `carrera`, `promedio`, `edad` y `materias aprobadas` refleja patrones reales observados en los datos.
Sin embargo, su precisión (60%) indica que, si bien útil, no es infalible. La interpretación debe complementarse con criterio humano y acciones pedagógicas.

Limitaciones del Modelo

El modelo no considera interacciones entre variables ni efectos acumulativos y Además, La muestra es limitada (100 estudiantes), lo que puede afectar la generalización.

Árbol Automático

En la construcción de árboles de decisión existen dos criterios principales para medir la calidad de una división:

- **Ganancia de información (Entropía):** mide la reducción de la incertidumbre. Es más sensible a distribuciones desbalanceadas y permite explicar de manera intuitiva cuánto “orden” se gana al dividir un nodo.
- **Impureza de Gini:** mide la probabilidad de clasificar incorrectamente un elemento si se asigna al azar según la distribución de clases del nodo. Es más rápido de calcular y suele dar resultados muy similares a la entropía.

Árbol con **Entropía (Ganancia de Información)**

criterion="entropy"

CARRERA
├── Diseño Industrial / Higiene y Seguridad → ABANDONA
├── Administración de PYMES / Enfermería → CONTINÚA
└── Otras carreras
└── PROMEDIO ├── < 6 → ABANDONA
├── > 8 → CONTINÚA
└── 6–8
└── EDAD
├── ≤ 30

│ └── **Materias Aprobadas ≤ 4 → ABANDONA**
│ └── **Materias Aprobadas > 4 → CONTINÚA**
└── **> 30 → CONTINÚA**

Árbol con **Gini (Impureza)**

criterion="gini"

el árbol cambió porque el algoritmo buscó divisiones que maximizan la pureza de los nodos, no la reducción de entropía.

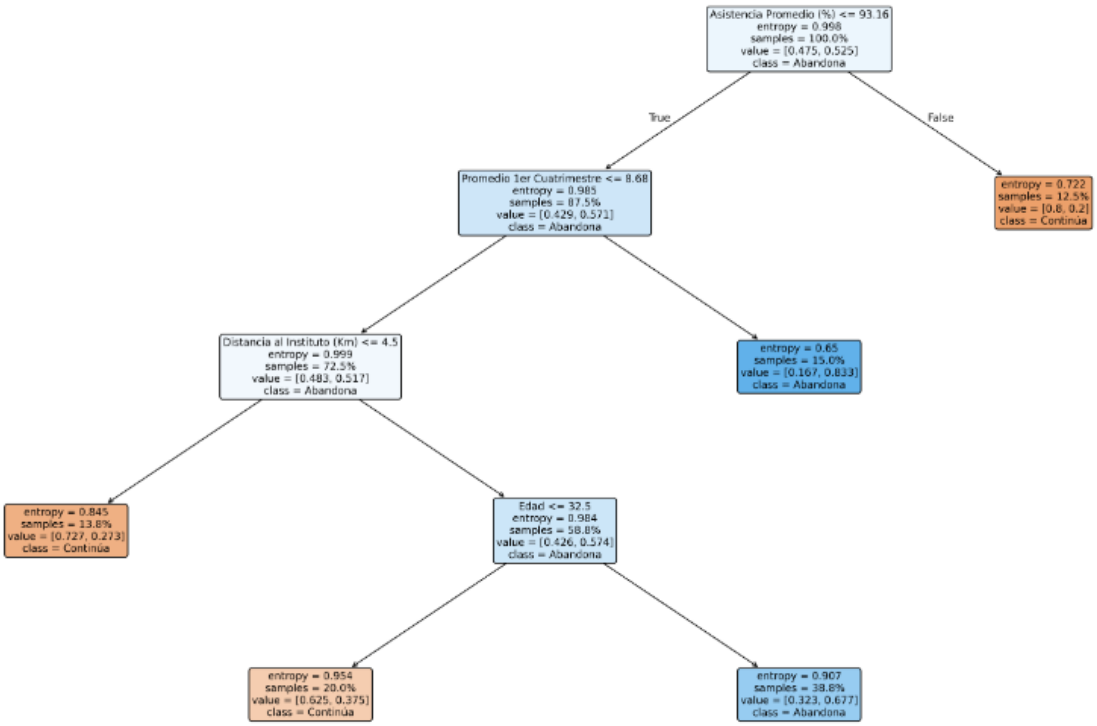
ASISTENCIA PROMEDIO (%) ≤ 83.5
├── Sí
│ └── PROMEDIO 1er Cuatrimestre ≤ 6.48
│ ├── Distancia al Instituto ≤ 6.5 km → ABANDONA
│ └── Distancia al Instituto > 6.5 km → ABANDONA (menos puro)
│ └── PROMEDIO > 6.48
│ ├── Edad ≤ 22.5 → CONTINÚA
│ └── Edad > 22.5 → ABANDONA
└── No (Asistencia > 83.5)
└── CONTINÚA

La raíz ya no es **Carrera**, sino **Asistencia Promedio (%)**, porque Gini detectó que esa variable separa más rápido los grupos.

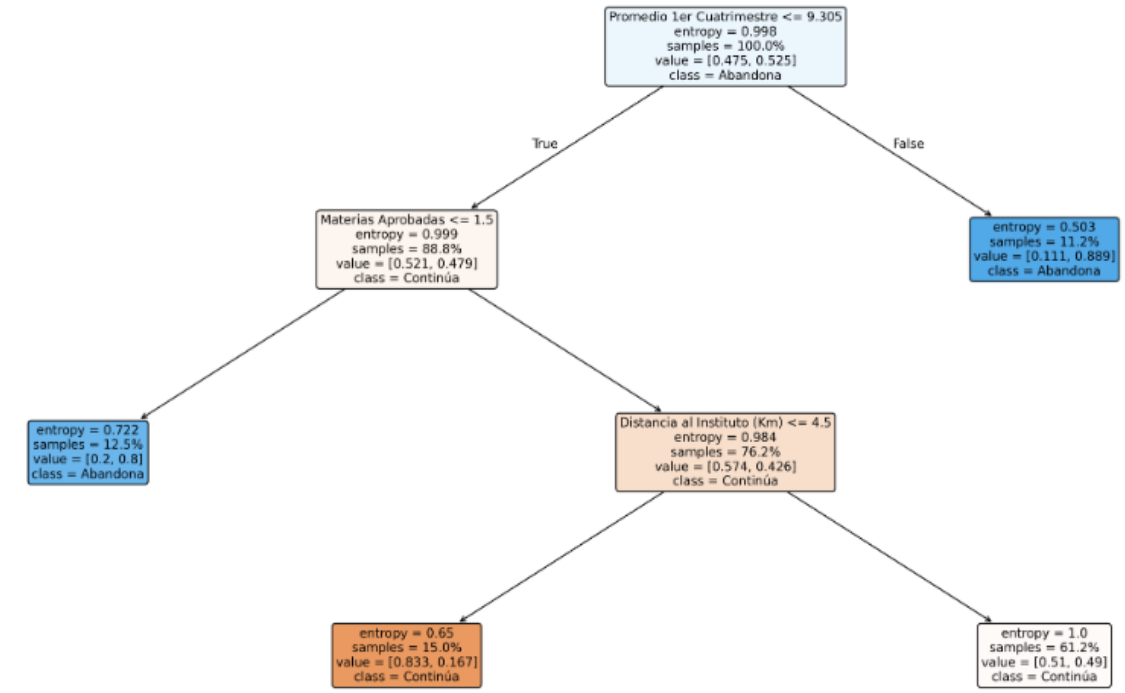
EVALUACION DE MODELOS Y SELECCION FINAL

Evalué distintas configuraciones del árbol de decisión para predecir el abandono estudiantil, comparando precisión, profundidad, número de hojas y capacidad explicativa.

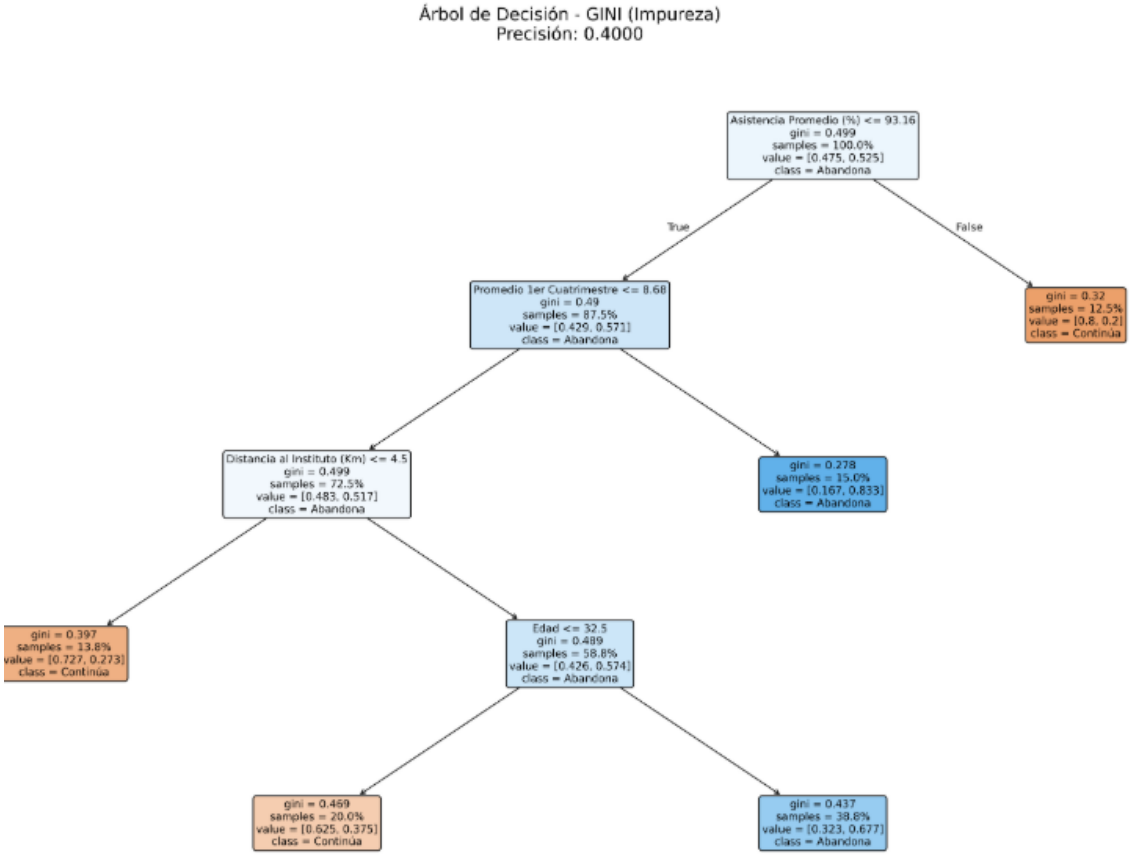
Árbol De desicion:Entropía Con 40% De precisión



Árbol De desicion:Entropía Con 55% De precisión

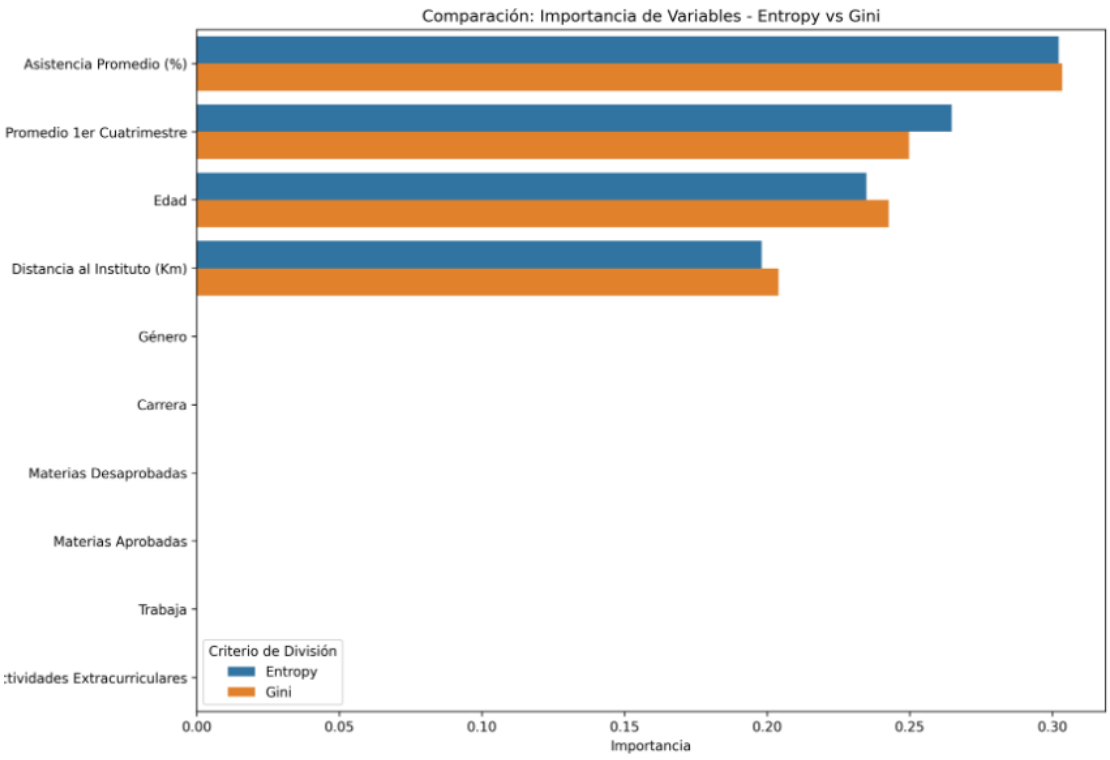


Árbol De decisión: Impureza (GINI) Con 55% De precisión



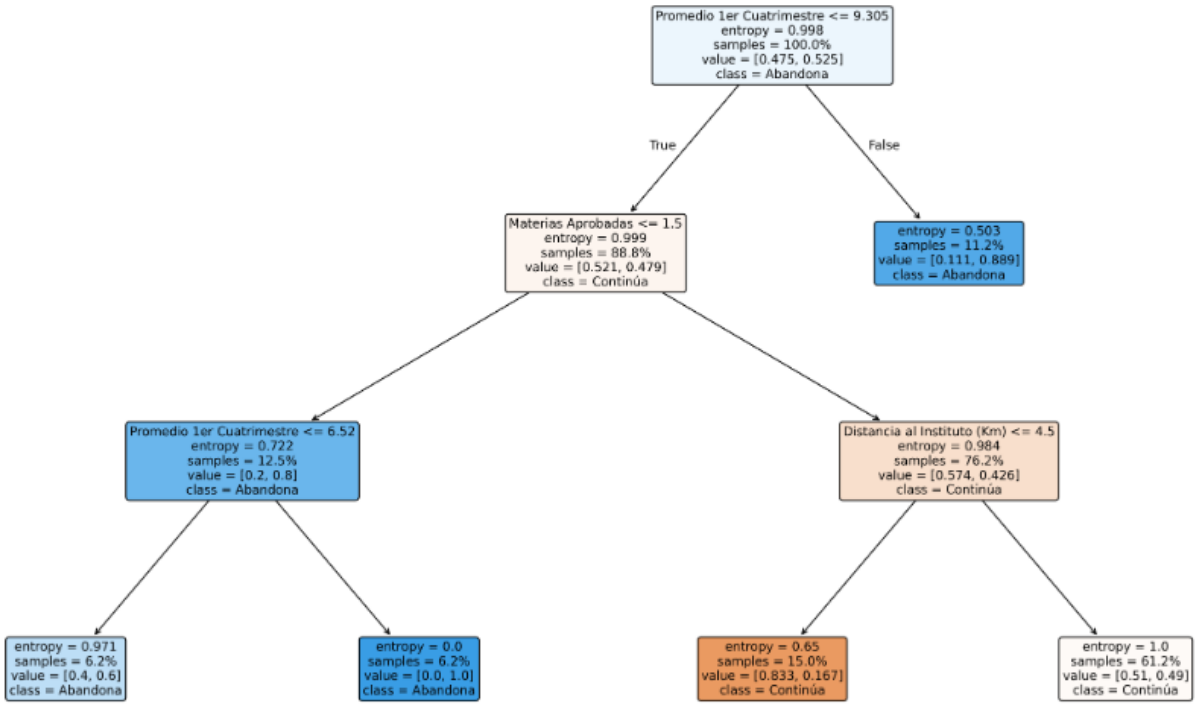
Cuadro comparativo entre importancia de variables

Gini vs Entropía



Modelo Seleccionado

Se eligió el modelo con criterio `entropía`, profundidad máxima de 3, mínimo de 10 muestras para dividir y 5 por hoja. Este modelo alcanzó una precisión del 55%, con una estructura clara y explicativa, ideal para comunicar resultados a públicos no técnicos



Evaluación del modelo seleccionado

```
4. EVALUACION DEL MODELO OPTIMIZADO:
   • Reporte de clasificación:

              precision    recall  f1-score   support

    Continúa       0.50      0.89      0.64         9
    Abandona       0.75      0.27      0.40        11

    accuracy                           0.55        20
   macro avg       0.62      0.58      0.52        20
weighted avg       0.64      0.55      0.51        20

   • Matriz de Confusión:
     Verdaderos Positivos: 3 (Abandonan - correcto)
     Falsos Positivos:    1 (Continúan - predicho abandono)
     Verdaderos Negativos: 8 (Continúan - correcto)
     Falsos Negativos:    8 (Abandonan - predicho continuación)
```

El modelo entrenado con criterio de entropía alcanzó una precisión general del **55%** sobre el conjunto de prueba. El análisis del reporte de clasificación y la matriz de confusión revela que el modelo tiene un buen desempeño para identificar a los estudiantes que **continúan** (recall de 0.89), pero presenta dificultades para detectar correctamente los casos de **abandono** (recall de 0.27).

- **8 verdaderos negativos**: estudiantes que continúan y fueron correctamente clasificados.
- **3 verdaderos positivos**: estudiantes que abandonan y fueron correctamente identificados.
- **8 falsos negativos**: estudiantes que abandonan, pero el modelo predijo que continuarían.
- **1 falso positivo**: estudiante que continúa, pero fue clasificado como abandono

El modelo tiende a ser conservador, priorizando la clasificación de continuidad. Esto puede ser útil para evitar alarmas innecesarias, pero también implica que **más de la mitad de los abandonos reales no son detectados**. Por eso, se recomienda usar el modelo como herramienta de apoyo, complementado con seguimiento académico personalizado para los casos con probabilidad moderada o alta de abandono.

SECCION FINAL

Descripción General Del código Abandono.py

1. **Carga y exploración** Se importa el historial estudiantil, se codifica la variable objetivo (`EstadoFinal`) y se genera un resumen estadístico.
2. **Preprocesamiento** Se codifican variables categóricas y se divide el conjunto en entrenamiento (80%) y prueba (20%).
3. **Entrenamiento del modelo** Se ajusta un árbol de decisión con criterio `entropía` y parámetros manuales para evitar sobreajuste.
4. **Evaluación** El modelo alcanza una precisión del 55%. Se analiza el rendimiento con métricas y matriz de confusión.

5. **Visualización** Se genera una imagen del árbol (`arbol_optimizado.png`) con atributos legibles para facilitar la interpretación.
6. **Predicción sobre nuevos datos** Se aplica el modelo a estudiantes actuales (`TablaPrediccionAbandono-DatosFinal.xlsx`) y se exportan los resultados a `Predicciones_Optimizadas.xlsx`.
7. **Ejemplo de uso** Se simula un estudiante ficticio y se predice su riesgo de abandono junto con una recomendación institucional.

Archivos generados

Archivo	Contenido
`arbol_optimizado.png`	Visualización del árbol entrenado con parámetros óptimos
`Predicciones_Optimizadas.xlsx`	Resultados de predicción sobre estudiantes actual

Recomendación institucional

A partir del modelo entrenado con criterio de entropía y una precisión del 55%, se recomienda implementar un sistema de seguimiento académico para estudiantes con riesgo moderado o alto de abandono.

El modelo permite identificar patrones asociados al abandono, como bajo promedio, alta cantidad de materias desaprobadas, baja asistencia y ciertas combinaciones de carrera y situación laboral. Se sugiere:

- Contactar proactivamente a estudiantes con probabilidad de abandono superior al 70%.
- Ofrecer tutorías, asesoramiento académico o flexibilización horaria para quienes trabajan.
- Monitorear mensualmente a estudiantes con riesgo moderado (30%–70%) y revisar su evolución.
- Evaluar la incorporación del modelo como herramienta complementaria en el sistema de gestión institucional.

Si bien la precisión puede mejorarse con modelos más complejos o más datos, este trabajo sienta las bases para una intervención temprana, basada en evidencia, que puede contribuir a reducir el abandono y mejorar la permanencia estudiantil.

Conclusión Final

El proceso de construir y seleccionar un modelo predictivo para el abandono estudiantil no fue solo técnico, sino también estratégico. A lo largo del trabajo, se exploraron múltiples configuraciones de árboles de decisión, comparando criterios como entropía y Gini, variando profundidad, tamaño mínimo de hojas y divisiones. Cada ajuste implicó un equilibrio entre precisión, interpretabilidad y utilidad institucional.

Aunque algunos modelos alcanzaron precisiones más altas, también mostraban señales de sobreajuste o estructuras difíciles de interpretar. Por eso, se optó por un modelo con **criterio de entropía**, **profundidad controlada** y una **precisión del 55%**, que, si bien no es perfecta, ofrece una base sólida para la toma de decisiones educativa