**Informe: Predicción de Abandono en Estudiantes Terciarios**

**Institución: Instituto Tecnológico Beltrán  
Materia: Procesamiento de Aprendizaje Automático  
Actividad: Arboles de Decision en equipo  
Echo por: Mariano Gordillo, Maximiliano Facundo Stella Zapata, Nicolas Martin, Josue Mena**

**Introducción**

El presente trabajo tiene como objetivo construir un modelo de árbol de decisión capaz de predecir si un estudiante tiene riesgo de abandonar la carrera o continuar sus estudios, utilizando información académica y demográfica obtenida durante su primer año en el Instituto Tecnológico Beltrán.

Esta herramienta busca brindar a la institución una base técnica para detectar casos de riesgo temprano y planificar estrategias de acompañamiento personalizadas.

**1. Datos utilizados**

El conjunto de datos fue obtenido del archivo TablaPrediccionAbandono-DatosFinal.xlsx y contiene la siguiente información por estudiante:

| **Variable** | **Descripción** |
| --- | --- |
| **Edad** | **Edad del estudiante** |
| **Género** | **Masculino o femenino** |
| **Carrera** | **Tecnicatura o especialidad cursada** |
| **PromedioPrimerCuatrimestre** | **Nota promedio del primer cuatrimestre** |
| **CantMateriasAprobadasPrimerCuatrimestre** | **Cantidad de materias aprobadas** |
| **CantMateriasDesaprobadasPrimerCuatrimestre** | **Cantidad de materias desaprobadas** |
| **AsistenciaPromedio(%)** | **Porcentaje de asistencia general** |
| **trabaja/NoTrabaja** | **Si el estudiante trabaja o no** |
| **DistanciaDomicilioAlInstituto(Kms)** | **Distancia desde su casa al instituto** |
| **ActividadesExtracurriculares(Estudio)** | **Si participa o no en tutorías u otras actividades** |
| **EstadoFinal** | **Variable objetivo: “Abandonó” o “Continúa”** |

Antes de entrenar el modelo, los datos fueron normalizados (por ejemplo: convertir “Si” -> “Sí”, “continua” -> “continúa”), y las variables categóricas se transformaron en valores numéricos para que el modelo pudiera interpretarlas.

**2. Preparación y metodología**

1. Limpieza y normalización:  
   Se eliminaron inconsistencias en mayúsculas, tildes y espacios.  
   Se aplicaron transformaciones con .str.lower() y .replace() para unificar los valores de texto.
2. Codificación numérica:
   * M/F → 0/1
   * Sí/No → 1/0
   * Continúa/Abandonó → 0/1
   * Las carreras se mapearon automáticamente con valores enteros.
3. División de datos:  
   Se separaron los datos en:
   * 70 % para entrenamiento
   * 30 % para prueba, asegurando representación equilibrada de ambas clases (*stratify=y*).
4. Modelo:  
   Se utilizó un árbol de decisión (DecisionTreeClassifier) de scikit-learn con los siguientes parámetros:
   * Profundidad máxima (max\_depth=4)
   * Muestras mínimas por división (min\_samples\_split=10)
   * Muestras mínimas por hoja (min\_samples\_leaf=5)  
     Estos valores ayudan a evitar el sobreajuste (*overfitting*).

**3. Resultados del modelo**

* Precisión general: alrededor del 80–90 %, dependiendo de la muestra.
* Matriz de confusión: mostró una buena capacidad para identificar correctamente a los estudiantes que continuaron y los que abandonaron.
* Reporte de clasificación: la métrica F1 para ambas clases fue satisfactoria, indicando equilibrio entre precisión y sensibilidad.

Variables más importantes

Según la métrica de importancia del modelo, los factores con mayor influencia en la predicción fueron:

| **Variable** | **Importancia aproximada** |
| --- | --- |
| **Promedio del primer cuatrimestre** | **Alta** |
| **Asistencia promedio (%)** | **Alta** |
| **Cantidad de materias desaprobadas** | **Moderada** |
| **Participación en tutorías** | **Moderada** |
| **Situación laboral y distancia al instituto** | **Baja a moderada** |

**4. Interpretación del árbol**

El árbol de decisión revela que los principales criterios para predecir abandono son:

* Bajo rendimiento académico (promedio < 6)
* Baja asistencia (< 70 %)
* Alta cantidad de materias desaprobadas (> 2)
* No participar en tutorías o actividades extracurriculares

Cuando un estudiante cumple varias de estas condiciones, la probabilidad de abandono aumenta significativamente.

**5. Recomendaciones para la institución**

A. Alertas tempranas:  
Detectar automáticamente a estudiantes con bajo promedio o asistencia deficiente al finalizar el primer cuatrimestre.

1. Tutorías obligatorias o reforzadas:  
   Priorizar la participación de los alumnos con bajo rendimiento o sin apoyo académico adicional.
2. Flexibilidad para estudiantes trabajadores:  
   Ofrecer modalidades adaptadas a quienes trabajan y viven lejos del instituto.
3. Seguimiento personalizado:  
   Implementar entrevistas o planes de acompañamiento en el primer año, especialmente en los casos que el modelo indique como de alto riesgo.

**Conclusiones**

El modelo desarrollado demostró ser efectivo para identificar patrones de abandono a partir de variables académicas y personales simples.  
Su aplicación práctica podría servir como una herramienta predictiva de apoyo a la gestión institucional, permitiendo anticipar situaciones de riesgo y actuar antes de que el estudiante abandone la carrera.

Además, el uso de técnicas de inteligencia artificial en este contexto refleja la utilidad de la ciencia de datos aplicada a la educación, fomentando una toma de decisiones basada en evidencia.

Tecnologías utilizadas

* Lenguaje: Python 3
* Librerías:
  + pandas, numpy → análisis y limpieza de datos
  + scikit-learn → modelo de árbol de decisión y métricas
  + matplotlib, seaborn → visualización
* Archivo de entrada: TablaPrediccionAbandono-DatosFinal.xlsx