

Universidad Nacional de Costa Rica

Laboratorio de Procesamiento de Images

Inteligencia Artificial Aplicada
Area de Comportamiento Humano

Coordinador: Msc. Johnny Villalobos Murillo

RF03-04

Modelo de Machine Learning para predicción de deserción estudiantil

1. Introducción

La presente investigación propone el desarrollo de un sistema de análisis predictivo basado en **aprendizaje automático supervisado**, cuyo objetivo es identificar patrones en los datos institucionales universitarios que permitan apoyar la toma de decisiones académicas y administrativas.

Como resultado de la investigación, se obtendrá **un modelo de Machine Learning supervisado (modelo de ML)**, que será capaz de procesar los datos disponibles, aprender de ellos y generar predicciones confiables sobre la trayectoria estudiantil y otros fenómenos académicos de interés, tal es el caso de Deserción Estudiantil.

2. Justificación del modelo de ML

El **modelo de ML supervisado** se desarrollará dentro de la investigación y se selecciona por su capacidad de:

- Manejar grandes volúmenes de datos
- Trabajar con variables heterogéneas (numéricas, categóricas, académicas y administrativas)
- Mantener robustez frente a datos incompletos o ruidosos
- Reducir el riesgo de sobreajuste mediante el aprendizaje combinado de múltiples estimadores
- Permitir la interpretación de la importancia relativa de las variables para la toma de decisiones

En resumen, **el modelo ML deberá poder aprender de los datos históricos y proporcionar información útil y accionable para la universidad.**

3. Fuentes de Datos Institucionales

El modelo se desarrollará utilizando datos provenientes de diferentes sistemas institucionales:

3.1 Sistema Banner de la Universidad Nacional

Datos disponibles:

- Historial académico de estudiantes
- Matrícula de cursos
- Calificaciones
- Avance en planes de estudio
- Información demográfica básica
- Registros de permanencia o retiro

3.2 Vicerrectoría de Vida Estudiantil

Datos sobre el contexto social y bienestar estudiantil:

- Becas socioeconómicas
- Participación en programas de apoyo
- Servicios psicológicos o de orientación
- Actividades extracurriculares
- Programas de acompañamiento estudiantil

3.3 Departamento de Registro

Datos sobre la trayectoria académica formal:

- Procesos de matrícula

- Cambios de carrera
- Reingresos
- Historial de cursos aprobados o reprobados
- Control de planes de estudio

4. Proceso Metodológico

El desarrollo del **modelo de ML** se realizará siguiendo varias etapas:

4.1 Identificación y recopilación de datos

Se identificarán las fuentes disponibles y se recopilarán los datos históricos de:

- Sistema Banner
- Registros del Departamento de Registro
- Bases de datos de la Vicerrectoría de Vida Estudiantil

El modelo de ML deberá poder integrar estos datos de manera coherente y completa.

4.2 Preparación y limpieza de datos

Se procesarán los datos para garantizar su calidad:

- Depuración de registros incompletos
- Tratamiento de valores faltantes
- Codificación de variables categóricas
- Normalización o transformación de variables cuando sea necesario
- Integración de datos de diferentes fuentes

El modelo de ML deberá poder manejar datos heterogéneos y parcialmente incompletos.

4.3 Definición de variables del modelo

Se seleccionarán las **variables predictoras** y la **variable objetivo**:

- Variables predictoras: promedio académico, número de cursos matriculados, número de cursos reprobados, condición de beca, participación en programas de apoyo, historial de

matrícula, variables demográficas generales.

- Variable objetivo: riesgo de deserción, probabilidad de éxito académico o progreso en la carrera.

El modelo de ML deberá poder identificar relaciones significativas entre estas variables y la variable objetivo.

4.4 Construcción del modelo de ML

El modelo se entrenará utilizando los datos históricos:

- División en conjuntos de entrenamiento y prueba
- Entrenamiento y ajuste de hiperparámetros
- Combinación de múltiples estimadores para mejorar precisión y reducir sobreajuste

El modelo de ML deberá poder generar predicciones confiables sobre nuevos datos de estudiantes.

4.5 Evaluación del modelo

El desempeño se evaluará con métricas de aprendizaje supervisado:

- Precisión (accuracy)
- Matriz de confusión
- Recall o sensibilidad
- F1-score

El modelo de ML deberá poder demostrar su capacidad predictiva mediante estas métricas.

4.6 Interpretación de resultados

Se analizará la **importancia relativa de las variables** para identificar:

- Factores que más influyen en los fenómenos estudiados
- Variables institucionales con mayor impacto
- Posibles áreas de intervención institucional

El modelo de ML deberá poder aportar información interpretable y útil para la toma de decisiones.

5. Resultados Esperados

Se espera que, como resultado de la investigación, **el modelo de ML obtenido:**

1. Pueda identificar patrones en la trayectoria académica estudiantil
2. Detecte factores asociados al éxito o riesgo académico
3. Genere información útil para apoyar la toma de decisiones institucionales
4. Contribuya al diseño de políticas universitarias basadas en datos