Análisis y Modelado del Comportamiento Académico Estudiantil

Este trabajo aplica Machine Learning para predecir el egreso y riesgo de deserción en la Universidad del Gran Rosario, transformando datos académicos y socioeconómicos en conocimiento accionable.

Paso 1: Integración y Análisis

01

Normalización de Variables

Se realizó una normalización de la variable "alumno" para estandarizar los nombres .

02

Unificación de Datos

Se ejecutó una operación de *inner join* para integrar diferentes fuentes de datos, creando un dataset consolidado y coherente.

03

Proceso de Limpieza

Se llevó a cabo una limpieza del dataset, abordando valores ausentes, corrigiendo inconsistencias y eliminando duplicados

04

Exploración y Diagnóstico

Se realizó un análisis exploratorio de datos (EDA) para entender la distribución de las variables, identificar patrones y detectar anomalías antes de la fase de modelado.

Resultados Principales del Análisis

La base consolidada final cuenta con 40.786 registros y 41 columnas

Sexo

63% femenino

Modalidad

59% presencial

Ubicación

75% Rosario

Paso 2: Creación de Variables Objetivo

Se definieron variables clave (Activo, Egresado o Pasivo) para el modelado predictivo.

01

Normalización de Texto

Función creada para estandarizar textos (minúsculas, sin caracteres especiales ni acentos), asegurando comparaciones precisas.

02

Definición de 'Es Pasivo'

Identificación de alumnos inactivos (sin actividad académica en los últimos 2 años). Resultado: 14.4% pasivos (5,890 alumnos).

03

Clasificación de Estado Final

Alumnos categorizados en 3 estados: Activo (54.1% / 22,057), Egresado (31.5% / 12,839), Pasivo (14.4% / 5,890).

04

Creación de Etiquetas Binarias

Derivación de `es_egresado` y `es_pasivo` como variables binarias a partir de `estado_final` para el modelado.

Paso 3: Modelo Predictivo Multiclase

Desarrollamos un modelo de Random Forest para predecir el estado académico final (Activo, Egresado o Pasivo).

Variables Predictoras

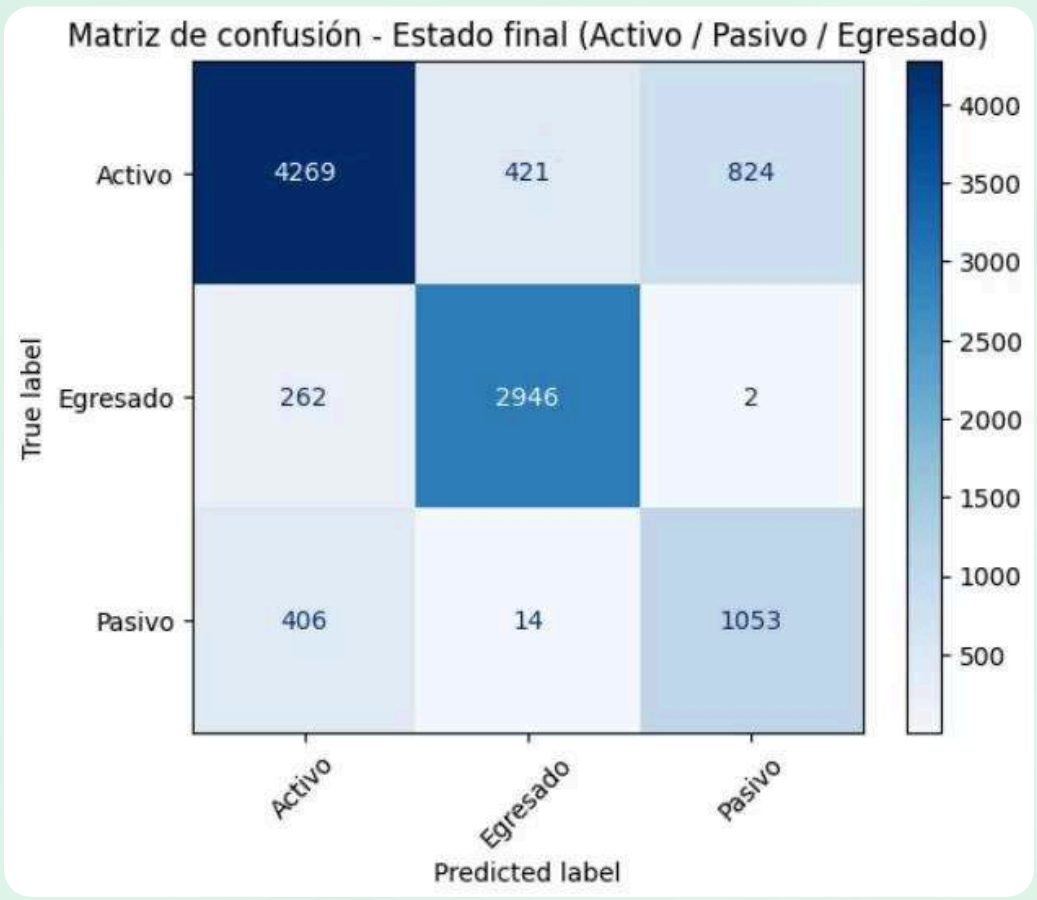
- **Académicas:** total_actas, cant_aprobado, Nota cursada.
- **Institucionales:** Propuesta, Modalidad, Ubicación.
- **Socioeconómicas:** trabajo_existe, Tiene Beca?, sexo.

Pipeline de Preprocesamiento

- **Imputación:** Se manejaron los valores faltantes usando estrategias adecuadas para cada tipo de variable.
- **Codificación:** Variables categóricas transformadas a formato numérico (One-Hot Encoding).

Configuración del Modelo Random Forest

- **Estimadores:** 300 árboles para mejorar la robustez y reducir el sobreajuste.
- **class_weight = "balanced"** : para compensar el desbalance entre clases
- **random_state = 42**: para reproductibilidad



Resultados del Modelo

	Precisión	Recall	F1-Score
Activo	0.86	0.77	0.82
Egresado	0.87	0.92	0.89
Pasivo	0.56	0.71	0.63

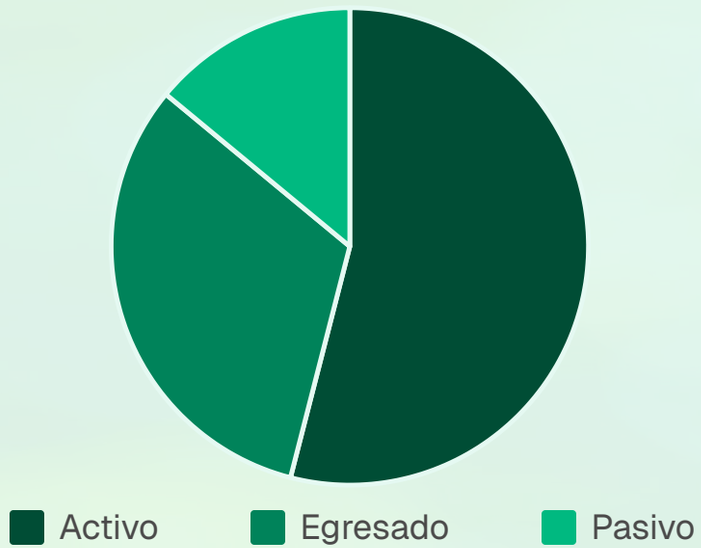
Accuracy Global: 0.81

Análisis de Importancia de Características

El rendimiento académico, reflejado en variables como `Nota cursada` y `cant_aprobado`, demostró ser el principal determinante del estado académico final del alumno.

Paso 4: Análisis de la Distribución

Comprender la distribución de las clases objetivo es fundamental para evaluar el sesgo del conjunto de datos y su posible impacto en el rendimiento del modelo, especialmente para las clases minoritarias como la de "Pasivo".



Para mitigar el impacto potencial de esta distribución en el rendimiento predictivo, especialmente para las clases minoritarias, se pueden aplicar técnicas de balanceo de datos, como el sobremuestreo (oversampling), submuestreo (undersampling) o el ajuste de pesos de clase, para asegurar un entrenamiento robusto y una capacidad de generalización adecuada del modelo.

Validación del Modelo: Split de Datos y Análisis de Distribución

Para la evaluación del modelo, los datos fueron divididos en un conjunto de entrenamiento (75%) y uno de prueba (25%) utilizando muestreo estratificado para asegurar que las proporciones de las clases objetivo (Activo, Egresado, Pasivo) se mantuvieran en ambos subconjuntos.

La distribución observada de las clases objetivo en el conjunto de datos es la siguiente:

- **Activo:** 54.1%
- **Egresado:** 31.5%
- **Pasivo:** 14.4%

Paso 5: Modelo General de Egreso

Se desarrolló un modelo de clasificación binaria para predecir si un estudiante egresa (1) o no egresa (0), utilizando un enfoque integral que incorpora variables académicas y contextuales.

El modelo se construyó utilizando un algoritmo de **Random Forest**, configurado con 300 árboles y pesos de clase balanceados. Se implementó un pipeline robusto, asegurando la calidad y preparación de los datos para el entrenamiento.

0.94	0.93	0.97
Accuracy	Recall	ROC AUC

El análisis de importancia de las características reveló que las variables relacionadas con el rendimiento académico son los principales predictores del egreso. Estos resultados subrayan la relevancia de los indicadores académicos en la capacidad predictiva del modelo.

Paso 6: Análisis Descriptivo de Desertores

El objetivo de este paso es comprender las características de los estudiantes que abandonan sus estudios, identificando factores clave a través de un análisis descriptivo.

Criterios de Filtrado

El análisis se centró en estudiantes de Tecnicaturas y Licenciaturas que tuvieran al menos 3 materias aprobadas y notas válidas (>0) para asegurar un historial académico relevante.

Reducción de Casos

Aplicando estos filtros, el número de desertores "reales" se redujo de 5540 a 2246 estudiantes, permitiendo un análisis más preciso.

Enfoque del Análisis

Se exploraron tres áreas principales: la situación laboral de los estudiantes, sus condiciones de convivencia y los aspectos académicos que pudieron influir en su decisión de desertar.

Estos filtros y enfoque nos permitieron concentrarnos en los casos más relevantes, revelando que el 58% de los desertores trabajaba y el 52% vivía con su familia de origen, mientras un 27% lo hacía con pareja e hijos, lo que sugiere posibles influencias externas en la deserción.

Paso 6: Análisis Descriptivo de Desertores - Rendimiento Académico

El análisis se realizó sólo sobre carreras de Tecnicatura y Licenciatura, pero para un análisis más preciso y relevante, se aplicaron filtros de calidad (mínimo 3 materias aprobadas y notas válidas > 0) para enfocarse en estudiantes con progreso académico significativo antes de la deserción, reduciendo el grupo a 2,246 desertores "reales".

Nota Cursada

Promedio de **6.83**, notablemente inferior al de los egresados.

Materias Aprobadas

Un promedio de **11.85**, lo que indica un progreso limitado en sus planes de estudio.

Materias Reprobadas

Un promedio de **2.39**, sugiriendo dificultades persistentes en ciertas asignaturas.

Materias Ausentes

Promedio de **2.43**, un indicador de abandono de cursos.

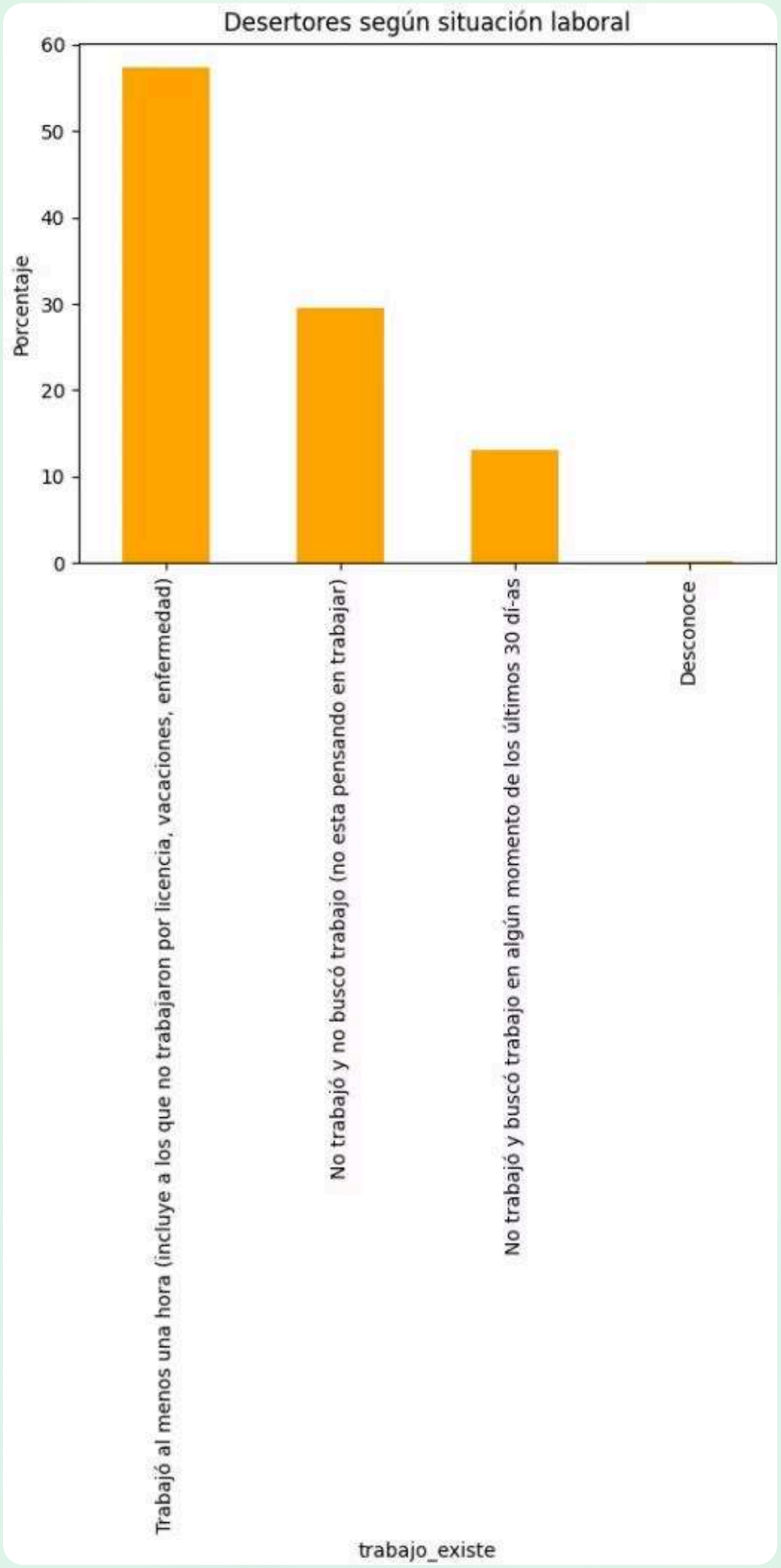
La distribución de calificaciones de los desertores mostró un patrón bimodal, indicando dos perfiles principales: estudiantes con un bajo rendimiento constante y aquellos que inicialmente tenían un rendimiento aceptable pero que luego experimentaron un descenso significativo. Este fenómeno subraya la complejidad de los factores que contribuyen a la deserción.

Comparativa Académica: Egresados vs. Desertores

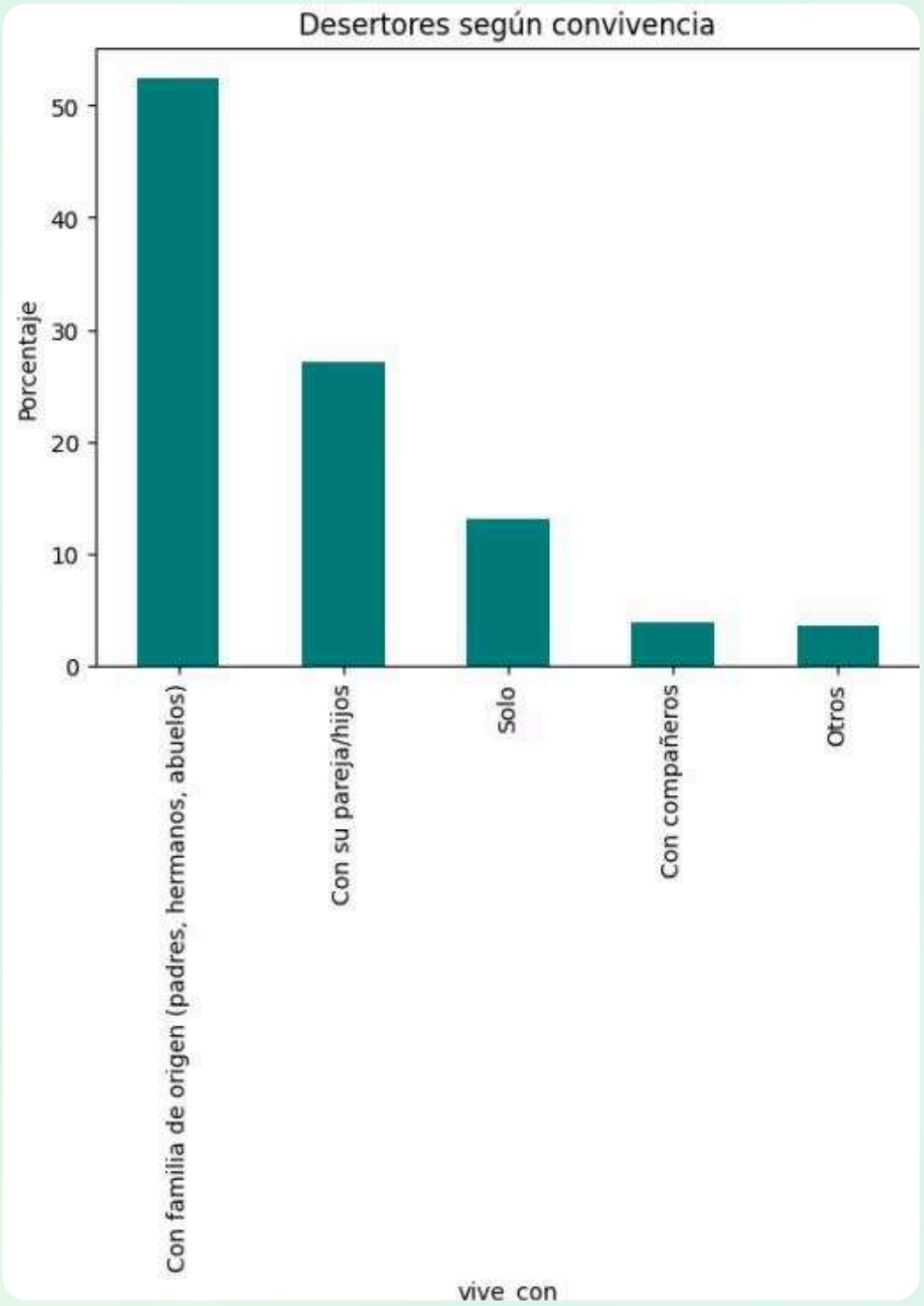
Nota Cursada	7.75	5.42
Materias Aprobadas	26.79	7.66
Materias Reprobadas	0.71	1.69

Análisis del Perfil y Base para la Predicción de Desertores

Este análisis detalla el perfil de los estudiantes que desertan, identificando características académicas, sociales y laborales clave que servirán de base para el desarrollo de un modelo predictivo de deserción.



La actividad laboral es un factor importante en la deserción.



La convivencia con la familia o responsabilidades propias pueden influir.

Paso 7: Modelo Predictivo de Deserción

Se desarrolló un modelo predictivo para identificar estudiantes en riesgo de deserción, utilizando características académicas, sociales y laborales. La variable objetivo, **"es_pasivo"**, indica si un estudiante ha desertado o no.

Variables Predictoras Clave

Variables académicas

Rendimiento, notas, materias aprobadas.

Variables laborales

Condición de actividad laboral.

Variables socio-demográficas

Entorno familiar y responsabilidades personales.

Configuración del Modelo

Se empleó un modelo de **Random Forest** (300 árboles) con pesos balanceados. Los datos se dividieron en 75% para entrenamiento y 25% para evaluación.

Métricas de Rendimiento

Accuracy

Proporción de predicciones correctas.

Recall (clase pasivo)

Capacidad para identificar desertores.

ROC AUC

Capacidad discriminatoria entre clases.

Conclusiones Principales

Rendimiento Académico

Principal predictor de deserción (notas, materias aprobadas).

Variables Contextuales

Actividad laboral y convivencia familiar también son significativas.

Sistema de Alerta Temprana

Base para identificar estudiantes en riesgo y aplicar intervenciones dirigidas.

Conclusión Final Detallada

Este proyecto desarrolló modelos predictivos de egreso y deserción estudiantil para una institución universitaria, utilizando análisis de datos y Random Forest.

Hallazgos Clave y Perfil del Desertor

- El **rendimiento académico** es el predictor más influyente para egreso y deserción.
- Factores **socio-demográficos y laborales** impactan significativamente la deserción.

Métricas de Rendimiento de los Modelos

Modelo de Egreso

- **Accuracy:** 0.94
- **ROC AUC:** 0.97

Modelo de Deserción

- **Accuracy:** 0.90
- **ROC AUC:** 0.92

Aplicaciones Prácticas y Valor Institucional

Sistema de Alerta Temprana

Identificación proactiva de estudiantes en riesgo.

Intervenciones Personalizadas

Diseño de programas de apoyo dirigidos.

Optimización de Recursos

Focalización de esfuerzos de retención.

Cultura de Mejora Continua

Promoción del éxito y permanencia estudiantil.