PRACTICA PRE-PROFESIONAL

TP FINAL

INTRODUCCIÓN

El trabajo tiene como objetivo principal analizar y modelar el comportamiento académico general de los estudiantes de la Universidad del Gran Rosario a partir de datos reales. El enfoque principal consiste en aplicar técnicas de Machine Learning para predecir tanto la probabilidad de egreso como el riesgo de deserción de los alumnos.

Este proyecto integra conocimientos de análisis de datos, estadística, programación y ciencia de datos aplicada, con el propósito de transformar información académica y socioeconómica en conocimiento accionable. A través del procesamiento, limpieza y unificación de distintas fuentes de datos institucionales, se construyó una base consolidada que permitió explorar las relaciones entre variables como el rendimiento académico, la modalidad de cursado, la ubicación geográfica, la situación laboral y la obtención de becas.

Posteriormente, se desarrollaron modelos predictivos supervisados basados en algoritmos de bosques aleatorios (Random Forest), tanto para el egreso como para la deserción. Estos modelos permiten identificar los factores más influyentes en la finalización de los estudios y estimar la probabilidad de que un estudiante abandone o complete su carrera.

1- INTEGRACION Y ANALISIS PRINCIPAL DEL DATASET

Esta primera etapa tuvo como propósito unificar y explorar los datos institucionales provenientes del archivo **Base_Pasantes.xlsx**, con el fin de obtener una base consolidada, limpia y lista para el análisis posterior.

1. Carga de datos: Esta primera etapa tuvo como propósito unificar y explorar los datos institucionales provenientes del archivo **Base_Pasantes.xlsx**, con el fin de obtener una base consolidada, limpia y lista para el análisis posterior.

2. Normalización texto de la variable “alumno”: Se estandarizó la variable **alumno** (nombre del estudiante) convirtiendo todo a minúsculas y eliminando espacios innecesarios. Esto evitó errores por diferencias de formato en la unión entre ambas hojas.

3. Unificación de datos: Se realizó un **inner join** entre las dos hojas, combinando solo los alumnos presentes en ambas. De esta forma se obtuvo una base única llamada unido, que integra los datos personales, académicos y administrativos de cada estudiante.

4. Limpieza: Se eliminaron las columnas duplicadas y se renombró la columna final a **Alumno** para mantener consistencia.

5. Exploración y diagnóstico: Se revisaron dimensiones, tipos de variables, porcentajes de valores nulos y cantidad de valores únicos por columna.

Se identificaron las columnas más incompletas (principalmente las relacionadas al trabajo y al título).

Se calcularon estadísticas descriptivas para las variables numéricas.

Se analizaron distribuciones de variables categóricas relevantes como **sexo**, **Modalidad**, **Ubicación**, **Propuesta**, **Calidad actual**, **Tiene Beca?** y **trabajo_existe**.

Resultados principales: La base consolidada resultó en 40.786 registros y 41 columnas.

Predominan las variables categóricas (≈80%), con presencia moderada de datos numéricos y algunos valores faltantes. La mayoría de los estudiantes son de sexo femenino (≈63%), modalidad presencial (≈59%), y sede Rosario (≈75%). Las carreras están fuertemente desbalanceadas, destacando la Licenciatura en Kinesiología y Fisiatría como la más numerosa.

El promedio general de notas ronda los 6,9 puntos, mientras que la mediana de materias aprobadas es 15.

Esta primera fase permitió obtener una visión global del conjunto de datos, identificar problemas de calidad (valores faltantes o desbalance) y establecer una base sólida para las siguientes etapas de modelado predictivo.

2- CREACIÓN DE VARIABLES OBJETIVO Y CLASIFICACIÓN DEL ESTADO ACADÉMICO

El propósito de esta etapa fue definir y construir las variables objetivo (targets) que servirán para el entrenamiento de los modelos predictivos.

A partir de los campos **“motivo_cambio”** y **“Calidad actual”**, se buscó determinar el estado académico final de cada estudiante dentro de tres categorías principales: Activo, Egresado o Pasivo (abandono o pausa en los estudios).

1. Normalización texto: Se creó una función **normalizar_columna()** para estandarizar los textos, eliminando espacios y pasando todo a mayúsculas. Esto permite comparar valores sin errores por diferencias de formato.

2. Creación de la variable **es_pasivo**: Se evaluó la existencia de las columnas **motivo_cambio** o **Calidad actual**, adaptando el código según cuál estuviera presente.

Si el motivo del cambio correspondía a casos como "ABANDONÓ", "PASIVO", "FALTA DE PAGO" o "SUSPENSIÓN DE CLASES", se asignó el valor 1, indicando deserción o pasividad. En los demás casos, el valor fue 0.

Resultado:

- **es_pasivo** = 1 → Alumno pasivo o desertor.
- **es_pasivo** = 0 → Alumno activo o egresado.

Tras la ejecución, se comprobó que la nueva variable fue creada correctamente con las siguientes proporciones:

- 85,6 % no pasivos
- 14,4 % pasivos

3. Clasificación del estado final: Se definió la función **clasificar_estado()** para categorizar a los estudiantes en tres grupos:

- *"Egresado"* → cuando el motivo fue EGRESO o EMISIÓN DE CERTIFICACIÓN DE GRADUADOS/AS.
- *"Pasivo"* → cuando figuraba ABANDONÓ, PASIVO, FALTA DE PAGO o SUSPENSIÓN DE CLASES.
- *"Activo"* → en todos los demás casos.

De esta manera se generó una nueva variable categórica llamada **estado_final**, con la siguiente distribución:

- *Activo:* 22.057 estudiantes (54,1 %)
- *Egresado:* 12.839 estudiantes (31,5 %)
- *Pasivo:* 5.890 estudiantes (14,4 %)

4. Derivación de etiquetas binarias: A partir de **estado_final** se crearon dos variables adicionales:

- **es_egresado** → 1 si el alumno fue clasificado como “Egresado”, 0 en caso contrario.
- **es_pasivo** → 1 si el alumno fue clasificado como “Pasivo”, 0 en caso contrario.

Estas variables binarias se utilizarán como targets en los modelos de clasificación, uno enfocado en predecir egreso y otro en deserción.

3- MODELO PREDICTIVO MULTICLASE PASIVO/ACTIVO/EGRESADO

En esta etapa se desarrolló un modelo de clasificación multiclase con el objetivo de predecir el estado académico final de cada estudiante (Activo, Egresado o Pasivo), en función de sus características académicas, demográficas y socioeconómicas. Este modelo permite detectar patrones de comportamiento estudiantil y comprender qué factores influyen más en el egreso o la deserción.

Se seleccionaron variables predictoras que reflejan tanto el rendimiento académico como el contexto del estudiante:

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

El dataset se limpió de valores faltantes y se dividió en un conjunto de entrenamiento (75%) y otro de prueba (25%), manteniendo la proporción de clases mediante estratificación.

Se implementó un pipeline que combinó los siguientes pasos:

1. Imputación de valores nulos:
   - Para variables numéricas → se reemplazaron por la **mediana**.
   - Para variables categóricas → por la **moda (valor más frecuente)**.

2. Codificación categórica: Se aplicó un **OneHotEncoder** para transformar textos en variables numéricas binarias interpretables por el modelo.

3. Estandarización automática: Todo el preprocesamiento se integró dentro de un **ColumnTransformer** que garantiza un flujo limpio y reproducible.

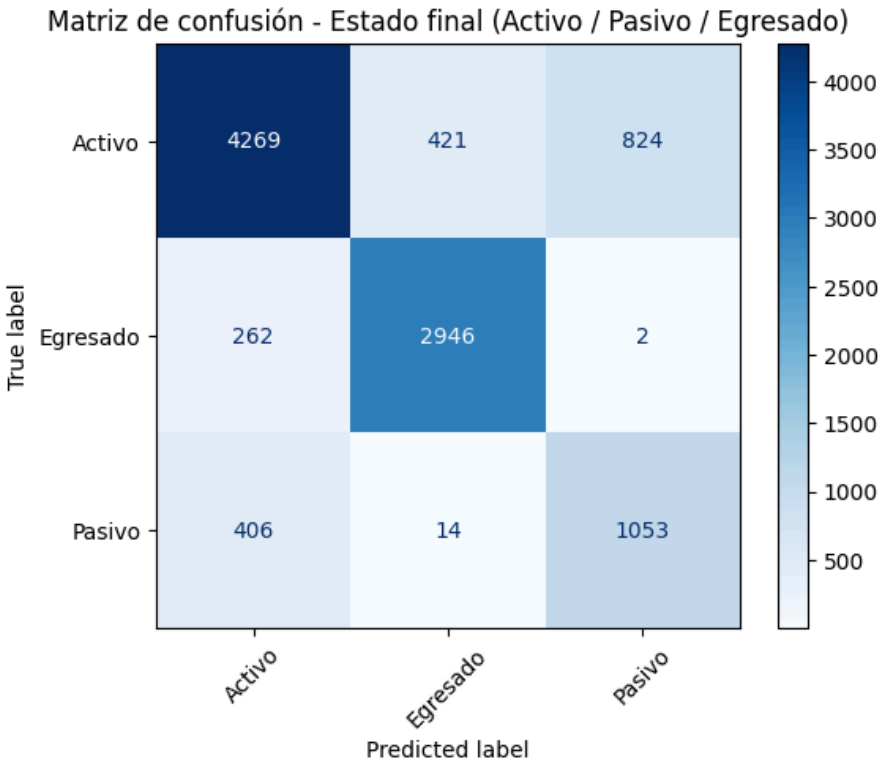
Luego, se utilizó un **Random Forest Classifier**, configurado con los siguientes parámetros:
- **n_estimators** = 300 árboles,
- **class_weight** = "balanced" para compensar el desbalance entre clases,
- **random_state** = 42 para reproducibilidad.

El modelo fue entrenado sobre el conjunto de entrenamiento (**X_train, y_train**) y luego evaluado sobre el conjunto de prueba (**X_test, y_test**).

Los resultados fueron los siguientes:

CLASE	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	-	-

El gráfico muestra que el modelo clasifica correctamente la mayoría de los egresados y activos, aunque los pasivos presentan más confusiones con la clase “Activo”, lo cual es esperable dado su menor representación en la base.



El análisis de las importancias de características del Random Forest mostró los siguientes resultados:

VARIABLE	IMPORTANCIA
Nota Cursada	0.335
Cant. Aprobadas	0.229
Total Actas	0.205
Carrera (Propuesta)	≈ 0.017
Ubicación	≈ 0.011
Modalidad	≈ 0.008

Esto confirma que el rendimiento académico es el principal determinante del estado final, mientras que las variables de contexto (como la carrera, la sede o la modalidad) tienen un papel complementario.

El modelo multiclase alcanzó un rendimiento general del 81% de exactitud, mostrando alta capacidad para distinguir entre estudiantes egresados y activos, y un desempeño moderado en la identificación de pasivos.

Los resultados permiten afirmar que:

- Los alumnos con mejor nota promedio y mayor cantidad de materias aprobadas tienen alta probabilidad de egresar.
- Los pasivos tienden a presentar menor rendimiento y menor número de actas registradas.
- El tipo de carrera (Propuesta) y la modalidad de cursado también inciden en los resultados.

4- ANÁLISIS DE DISTRIBUCIÓN DE CLASES Y VALIDACIÓN DE LOS CONJUNTOS DE ENTRENAMIENTO Y PRUEBA

Antes de avanzar con el modelado binario y las predicciones específicas, se realizó un análisis de la distribución de clases para comprobar que los subconjuntos de entrenamiento y prueba conservaran la misma proporción de estudiantes activos, egresados y pasivos. Esto es fundamental para evitar sesgos de aprendizaje y garantizar que el modelo generalice correctamente.

Una vez creada la variable **estado_final**, se dividió el dataset total en dos subconjuntos mediante la función **train_test_split()**:

- 75 % de los registros se destinaron al entrenamiento del modelo (train).
- 25 % restantes se reservaron para la prueba (test).

Durante la división, se utilizó el parámetro **stratify=y** para mantener la proporción de clases original dentro de cada subconjunto.

Distribución global del dataset:

ESTADO	CANTIDAD	PORCENTAJE
ACTIVO	22.057	54,1%
EGRESADO	12.839	31,5%
PASIVO	5.890	14,4%

Distribución en conjuntos:

CONJUNTO	ACTIVO	EGRESADO	PASIVO	TOTAL
Train (75%)	16.543	9.629	4.417	30.589
Test (25%)	5.514	3.210	1.473	10.197

Las proporciones del conjunto de test quedaron prácticamente idénticas a las globales:

- Activo → 54,1 %
- Egresado → 31,5 %
- Pasivo → 14,4 %

Este resultado demuestra que el procedimiento de muestreo estratificado funcionó correctamente, manteniendo la representatividad de todas las clases en ambos subconjuntos. Gracias a eso, las métricas de desempeño del modelo (accuracy, recall, AUC, etc.) pueden interpretarse de forma confiable y comparativa.

Se confirma también que el dataset presenta un ligero desbalance entre clases (con menor presencia de la categoría Pasivo), razón por la cual se empleó el parámetro **class_weight="balanced"** en el modelo Random Forest de la parte anterior.

5- MODELO GENERAL DE EGRESO

En esta etapa se entrenó un modelo de clasificación binaria para predecir si un estudiante logra egresar (1) o no (0), utilizando variables académicas y contextuales. Se creó una nueva variable binaria **es_egresado**, derivada de **estado_final**. Las variables predictoras elegidas fueron:

- Sexo
- Modalidad
- Ubicación
- Propuesta (tipo de carrera)
- Tiene Beca
- Trabajo existente
- Total de actas
- Cantidad de materias aprobadas
- Nota cursada

Estas variables reflejan tanto el rendimiento académico como el contexto personal del estudiante.

Se aplicó un Random Forest Classifier con:

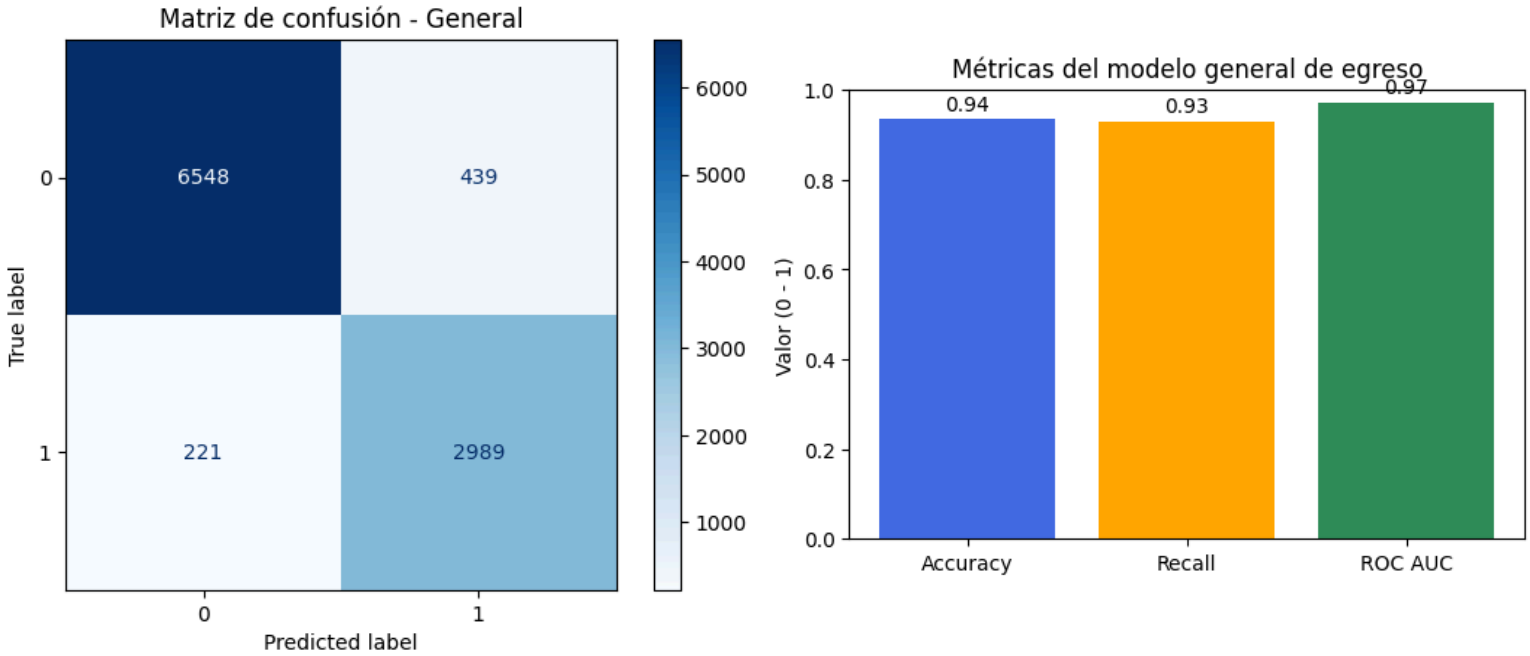
- 300 árboles de decisión
- **class_weight='balanced'** para manejar el leve desbalance de clases
- División 75 % train / 25 % test estratificada

El modelo se integró en un Pipeline con imputación de valores faltantes y codificación categórica (**OneHotEncoder**).

El rendimiento fue muy bueno:

MÉTRICA	VALOR
Accuray	0.94
Recall	0.93
ROC AUC	0.97

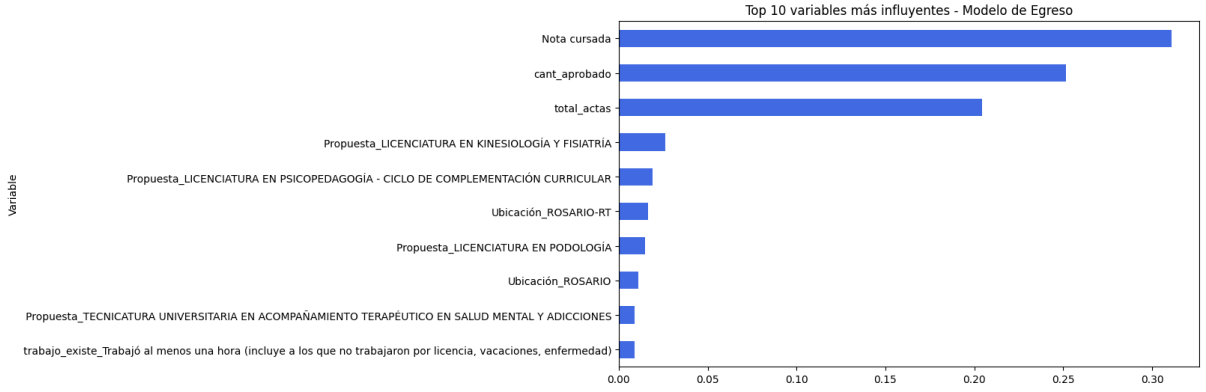
La matriz de confusión mostró una alta tasa de aciertos en ambas clases (egresado/no egresado), mientras que el gráfico de métricas visualizó el equilibrio entre precisión y sensibilidad.



El análisis de importancia del modelo destacó como principales predictores:

1. Nota cursada
2. Cantidad de materias aprobadas
3. Total de actas

Estas variables reflejan directamente el desempeño académico, mientras que factores como la modalidad, ubicación y propuesta contribuyen en menor medida, aportando contexto al rendimiento.



El modelo fue guardado en formato .pkl junto con las importancias de variables en un archivo Excel para futuras interpretaciones o implementaciones institucionales.

6- ANÁLISIS DESCRIPTIVO DE LOS DESERTORES

En esta etapa final se buscó comprender las características de los estudiantes que abandonan sus estudios (estado final = Pasivo), focalizando en aspectos laborales, convivenciales y académicos. Mientras que el bloque anterior se centró en un modelo predictivo de egreso, este análisis descriptivo tuvo el objetivo de profundizar en el perfil de quienes no lograron finalizar la carrera, para detectar posibles patrones de riesgo.

Filtrado: El análisis se realizó sólo sobre carreras de Tecnicatura y Licenciatura, ya que representan los principales trayectos académicos con estructura formal de cursado y egreso dentro de la institución. De este modo se excluyeron trayectos cortos, cursos o posgrados, que no comparten la misma dinámica académica ni las mismas causas de abandono.

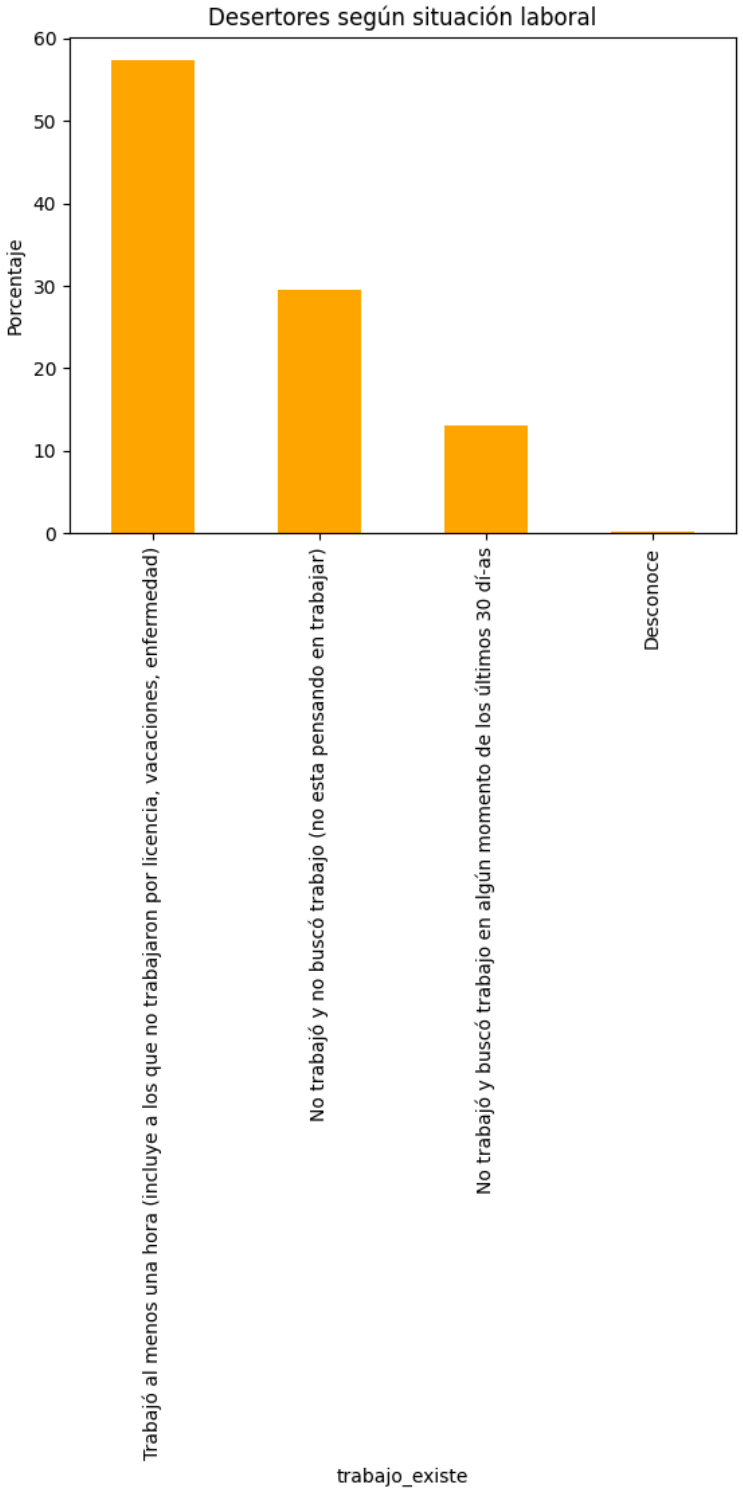
Inicialmente se identificaron 5540 estudiantes con estado **Pasivo**, es decir, aquellos que interrumpieron o abandonaron sus estudios. Sin embargo, no todos los registros representan verdaderos casos de deserción: algunos podrían ser estudiantes que apenas iniciaron la carrera, se reinscribieron, o tuvieron notas faltantes por causas administrativas.

Por eso se aplicó un filtro adicional de calidad, seleccionando únicamente a quienes:

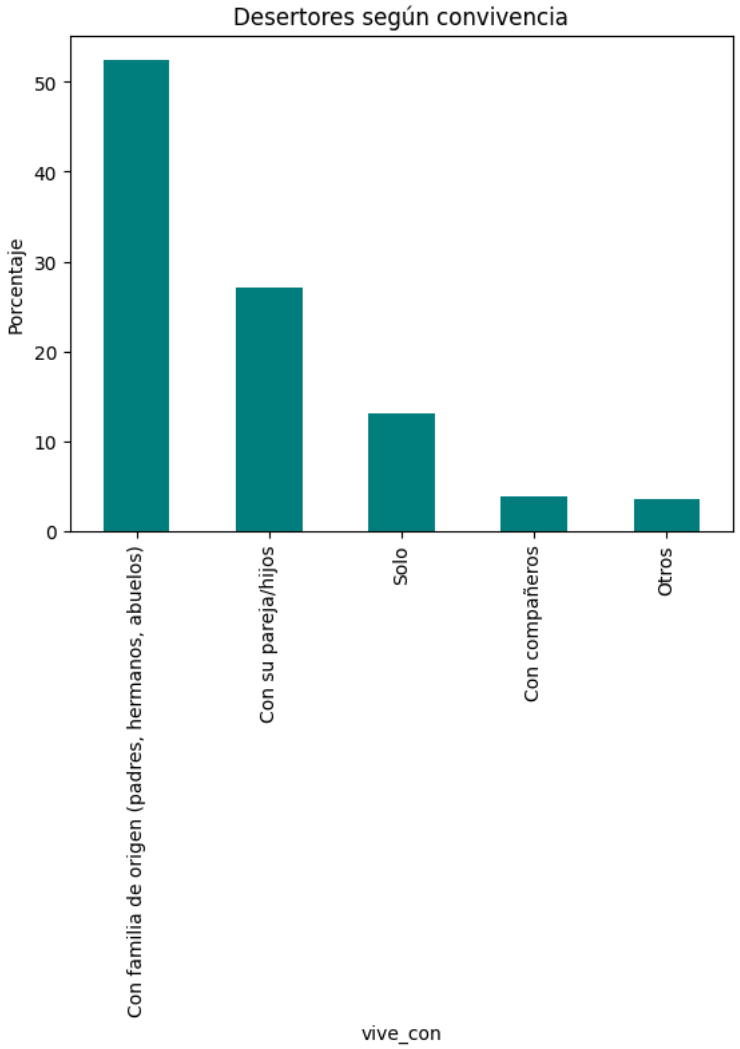
- Tuvieran al menos 3 materias aprobadas, para asegurar un mínimo recorrido académico.
- Poseyeran una nota de cursada válida mayor a 0, descartando errores o vacíos de información.

Este criterio permitió concentrar el análisis en 2246 desertores reales, es decir, estudiantes que efectivamente avanzaron parte del trayecto y luego abandonaron.

Situación laboral: El análisis laboral mostró que más de la mitad de los desertores (≈58 %) trabajaba al menos algunas horas por semana, mientras que cerca del 30 % no trabajaba ni buscaba empleo. Esto sugiere que la actividad laboral no siempre es la causa directa del abandono, aunque puede incidir indirectamente al limitar el tiempo disponible para estudiar o asistir a clases.



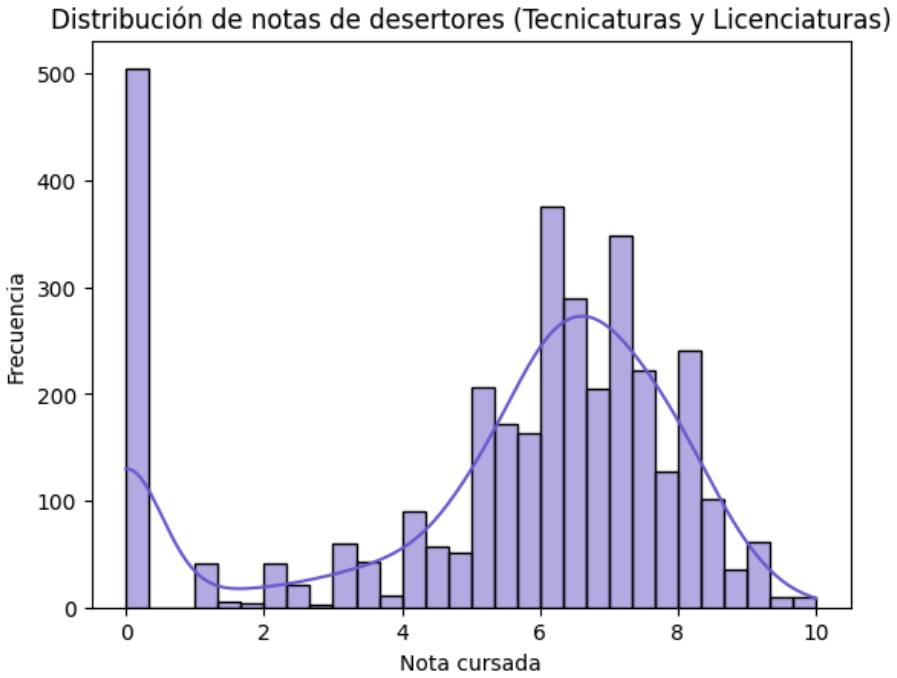
Condiciones de convivencia: La mayoría de los desertores (52 %) vive con su familia de origen (padres, hermanos, abuelos), seguida por un 27 % que convive con su pareja e hijos. Una menor proporción (13 %) vive sola, y solo un 4 % con compañeros. Estos datos reflejan que la dependencia familiar o la convivencia con responsabilidades propias (pareja/hijos) pueden representar contextos de presión o falta de tiempo, influyendo en la decisión de abandonar.



Desempeño académico promedio: Entre los desertores filtrados, los indicadores académicos fueron los siguientes:

INDICADOR	PROMEDIO
Nota Cursada	6.83
Materias Aprobadas	11.85
Materias Reprobadas	2.39
Materias Ausentes	2.43

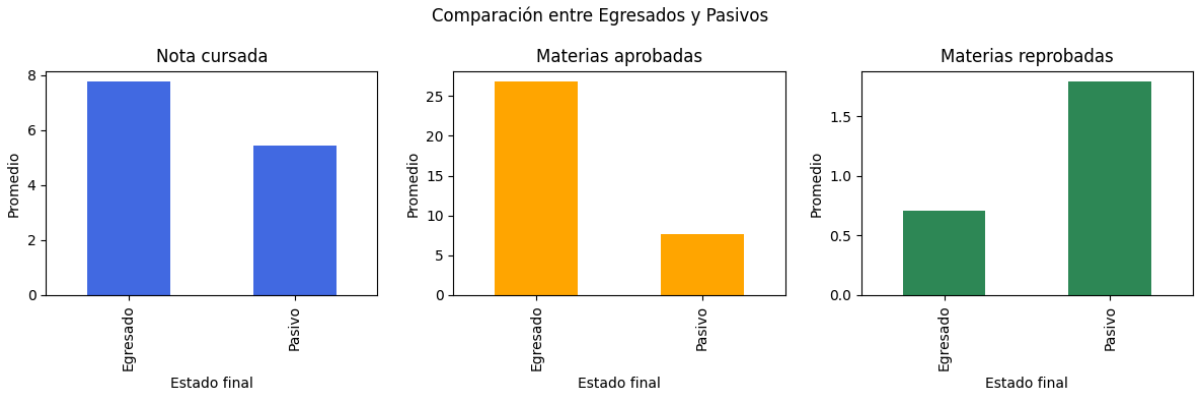
La distribución de notas presenta una forma bimodal: un grupo de desertores con notas muy bajas (abandono temprano) y otro con notas medias (abandono avanzado), evidenciando que existen dos perfiles distintos de deserción.



Comparación entre egresados y pasivos: La comparación entre egresados y desertores muestra diferencias significativas en todas las variables académicas:

VARIABLE	EGRESADOS	PASIVOS
Nota Cursada	7.75	5.42
Materias Aprobadas	26.79	7.66
Materias Reprobadas	0.71	1.69

Los estudiantes que egresan no solo alcanzan promedios más altos, sino que también tienen más del triple de materias aprobadas y menor proporción de reprobadas, confirmando que el rendimiento académico es el principal predictor del éxito o abandono.



Este análisis permitió perfilar con mayor precisión a los estudiantes en riesgo de deserción:

- Académicamente, presentan bajo rendimiento y progresan lentamente.
- Socialmente, suelen convivir con familia o pareja, lo que puede limitar la dedicación exclusiva al estudio.
- Laboralmente, una parte importante trabaja durante la cursada, afectando la disponibilidad de tiempo.

El filtrado aplicado fue fundamental para aislar los casos representativos de abandono real, evitando distorsiones por datos administrativos o alumnos que no alcanzaron un recorrido académico significativo. En conjunto con el modelo predictivo de la parte anterior, esta sección aporta una visión complementaria y explicativa sobre los factores que influyen en la permanencia y el egreso en la educación superior.

7- MODELO PREDICTIVO DE DESERCIÓN

En esta última etapa se desarrolló un modelo predictivo orientado a estimar la probabilidad de abandono o deserción de los estudiantes, identificando los factores que más contribuyen a esta situación. El modelo se complementa con el de egreso presentado anteriormente, proporcionando una visión integral del comportamiento académico de la población estudiantil.

Se creó una variable binaria denominada **es_pasivo**, donde:

- 1 representa a los estudiantes con estado Pasivo (abandono o interrupción de cursado).
- 0 representa a los estudiantes Activos o Egresados.

A partir del dataset completo se seleccionaron variables tanto académicas como contextuales:

- *Académicas:* **Nota cursada**, **total_actas**, **cant_aprobado**, **cant_reprobado**, **cant_ausente**.
- *Laborales:* **trabajo_existe**, **trabajo_hora_sem**, **trabajo_carrera**.
- *Sociales y demográficas:* **sexo**, **Modalidad**, **Ubicación**, **Propuesta**, **tipo_vivienda**, **vive_con**, **Tiene Beca?**.

Estas variables fueron elegidas por su relevancia en el análisis exploratorio previo y por reflejar dimensiones clave del estudiante: rendimiento, carga laboral y entorno social.

Se utilizó nuevamente un **Random Forest Classifier**, con los mismos criterios metodológicos que en el modelo de egreso:

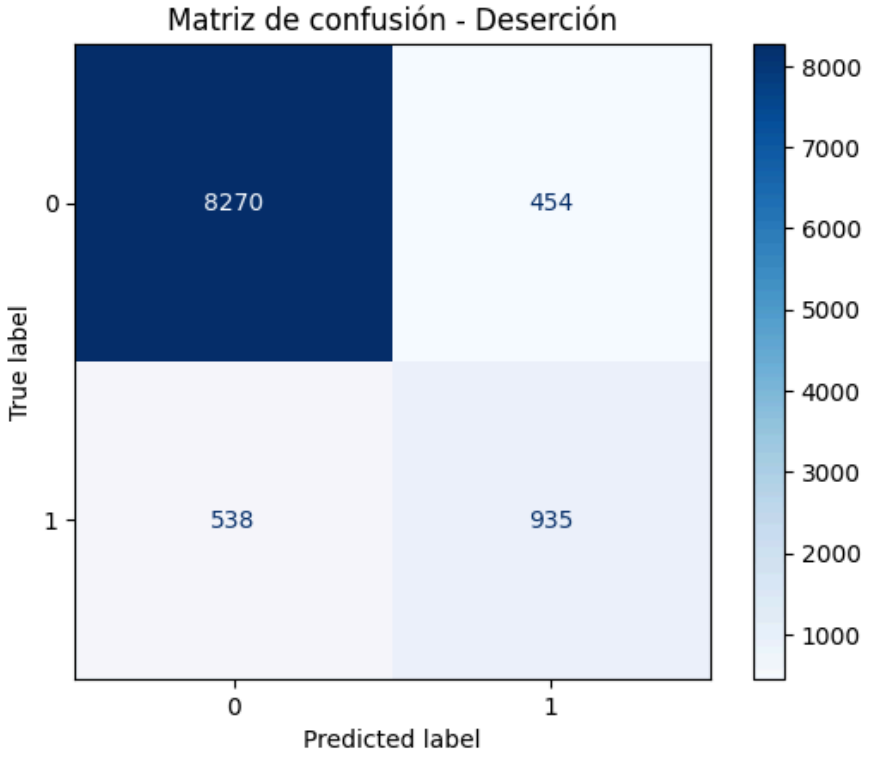
- 300 árboles de decisión.
- Balanceo de clases automático (**class_weight='balanced'**).
- División estratificada 75 % / 25 % (train/test).
- Preprocesamiento unificado con imputación y codificación categórica (OneHotEncoder).

El entrenamiento se realizó sobre 10.197 observaciones, distribuidas entre desertores y no desertores.

Resultados: El modelo logró un rendimiento sólido, con un buen equilibrio entre sensibilidad (recall) y precisión:

MÉTRICA	VALOR
Accuracy	0.90
Recall (clase pasivo)	0.63
ROC AUC	0.92

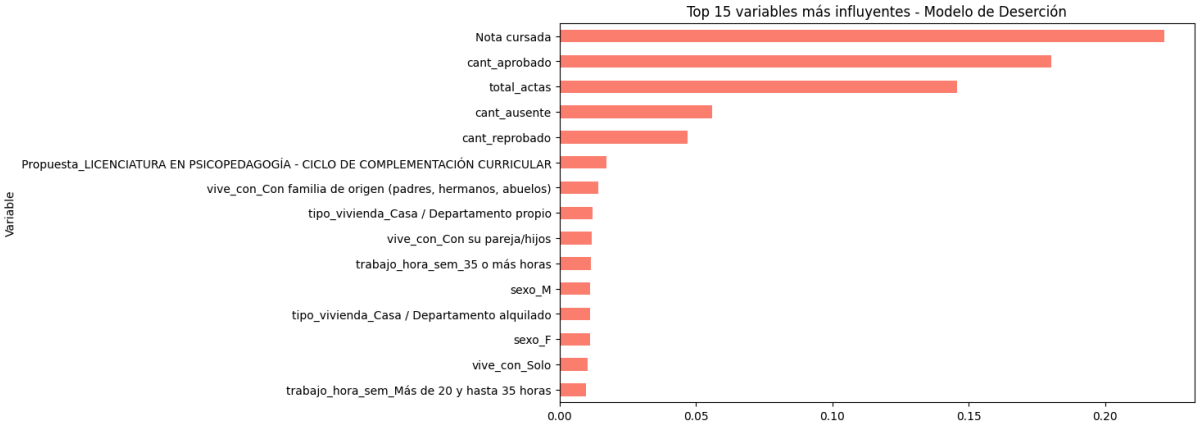
La matriz de confusión muestra una correcta clasificación de la mayoría de los casos, aunque con una ligera tendencia a subestimar los abandonos —lo cual es esperable dada la menor proporción de desertores en el conjunto de datos.



Variables: El análisis de importancia de variables evidenció que los principales predictores de deserción son:

1. Nota cursada.
2. Cantidad de materias aprobadas.
3. Total de actas.
4. Modalidad y ubicación de cursado.
5. Condiciones laborales y convivencia.

Esto confirma nuevamente que el desempeño académico sigue siendo el factor central, pero además se observa el peso de variables contextuales, como el entorno familiar y la compatibilidad entre trabajo y estudio, que afectan la continuidad académica.



El modelo predictivo de deserción alcanzó un nivel de desempeño alto (**ROC AUC = 0.924**), lo que indica una capacidad muy buena para diferenciar entre estudiantes que abandonan y los que permanecen activos.

Estos resultados permiten generar herramientas de alerta temprana, donde el sistema institucional podría identificar estudiantes en riesgo y actuar preventivamente mediante estrategias de acompañamiento y tutoría.

Finalmente, el modelo fue guardado en formato .pkl junto con el archivo de importancias de variables (**Importancias_Variables_Desercion.xlsx**), permitiendo su reutilización y evaluación futura por parte de la institución.

8- CONCLUSIÓN FINAL

El presente trabajo tuvo como propósito analizar y modelar el comportamiento académico de los estudiantes de la institución, buscando comprender los factores que influyen tanto en el egreso como en la deserción universitaria.

A partir de la base de datos institucional **“Base_Pasantes.xlsx”**, se realizó un proceso completo de limpieza, integración y normalización de información, uniendo hojas con registros académicos y datos personales de los alumnos. Este paso inicial permitió consolidar una única fuente de información confiable sobre la cual desarrollar el análisis.

Posteriormente, se crearon nuevas variables (**es_egresado**, **es_pasivo**, **estado_final**) que permitieron clasificar a los estudiantes según su situación académica. Esto sentó las bases para construir modelos predictivos robustos y realizar comparaciones entre grupos.

En la fase exploratoria, se observaron patrones claros:

- La mayoría de los alumnos permanecen activos o egresan, mientras que un 14 % abandona o queda pasivo.
- La Licenciatura en Kinesiología y Fisiatría concentra una gran parte de los registros, lo cual marca un fuerte desbalance en la distribución por carrera.
- Las variables académicas más representativas fueron la nota de cursada, el total de actas y la cantidad de materias aprobadas.

Se desarrollaron dos modelos principales, ambos basados en **Random Forest Classifier**:

1. Modelo de egreso (**es_egresado**):
   Alcanzó un rendimiento muy alto (Accuracy = 0.94, ROC AUC = 0.97), demostrando una gran capacidad para distinguir entre estudiantes egresados y no egresados. Este modelo reafirmó que el desempeño académico es el principal predictor del éxito.
2. Modelo de deserción (es_pasivo):
   Obtuvo un Accuracy de 0.90 y un ROC AUC de 0.92, reflejando un buen equilibrio entre precisión y sensibilidad. Cobraron relevancia variables contextuales como **Modalidad de cursado**, **Ubicación**, **Convivencia** y **Situación laboral**.

El estudio descriptivo sobre los desertores permitió identificar perfiles específicos:

- Más del 50 % vive con su familia de origen y trabaja parcialmente.
- Presentan notas promedio más bajas (≈6.8) y pocas materias aprobadas (≈12).
- En comparación, los egresados tienen notas superiores y más del triple de materias aprobadas.

Este análisis refuerza la idea de que la deserción no se explica por una sola causa, sino por una combinación de factores académicos, sociales y laborales.

El trabajo permitió construir una visión integral del desempeño estudiantil, combinando análisis exploratorio, modelado predictivo y evaluación interpretativa.

Se logró determinar que:

- El rendimiento académico es el principal indicador tanto de egreso como de deserción.
- Las condiciones personales y laborales actúan como factores moduladores.
- Los modelos desarrollados pueden ser utilizados por la institución como herramientas de apoyo a la gestión educativa, ayudando a detectar estudiantes en riesgo y a diseñar estrategias de acompañamiento personalizadas.

En síntesis, el trabajo no solo aportó resultados estadísticos, sino también insumos prácticos para la toma de decisiones institucionales, demostrando cómo la ciencia de datos puede aplicarse eficazmente en el ámbito educativo.