

# Objetivo Aprobar

¿Cómo podemos identificar a los alumnos que tienen riesgo de desaprobación un curso antes de que lo hagan?

Contamos con información de un conjunto de alumnos.

La información refiere a características de cursadas.

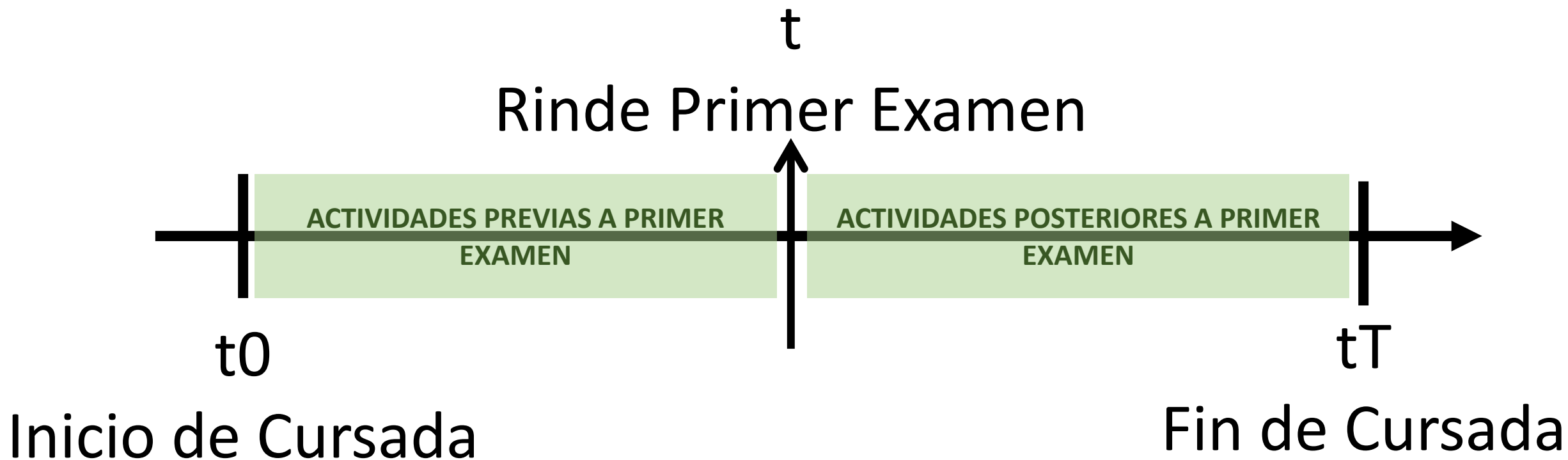
Tenemos información sobre qué exámenes rindió, con qué nota, a qué altura del curso.

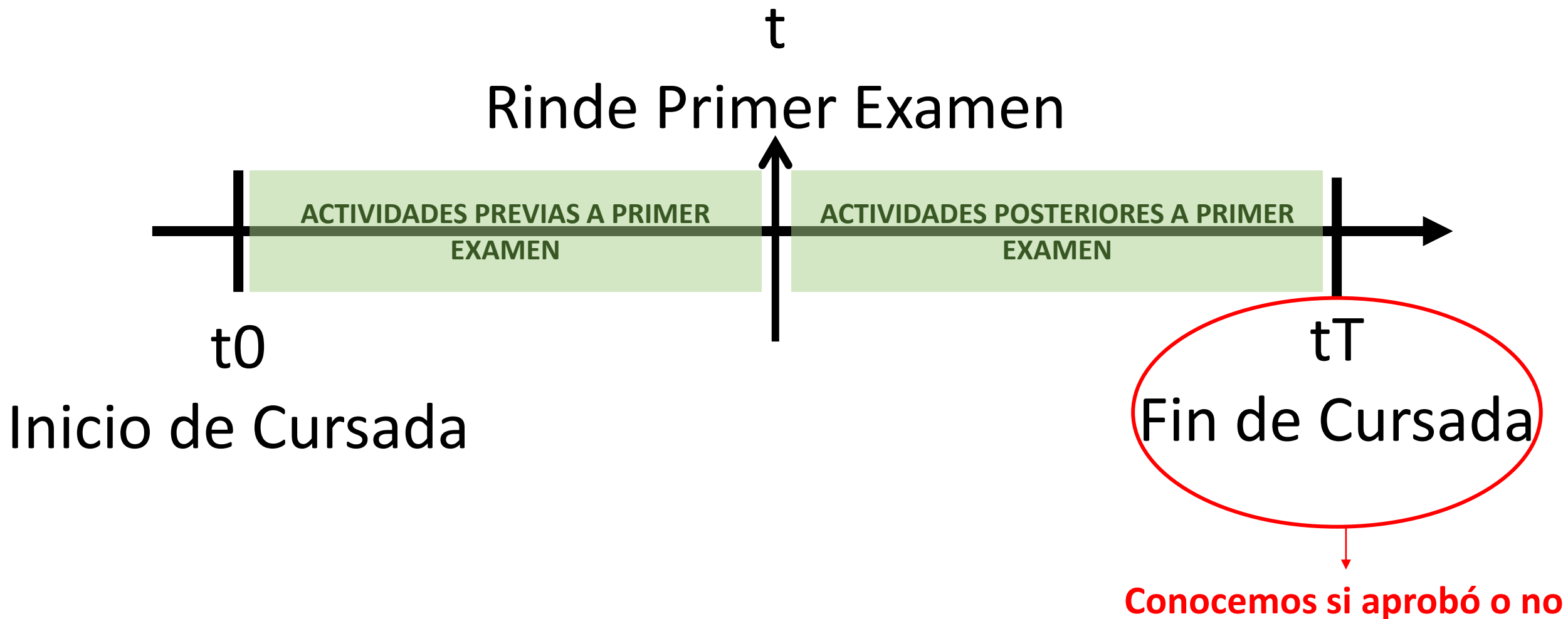
También, sobre qué actividades (TP, Integradoras, etc) desarrolló, con su nota y momento de entrega.

No tenemos información sobre características idiosincráticas del alumno, como sexo, edad, educación previa, residencia, características de padres, condiciones ambientales, etc.

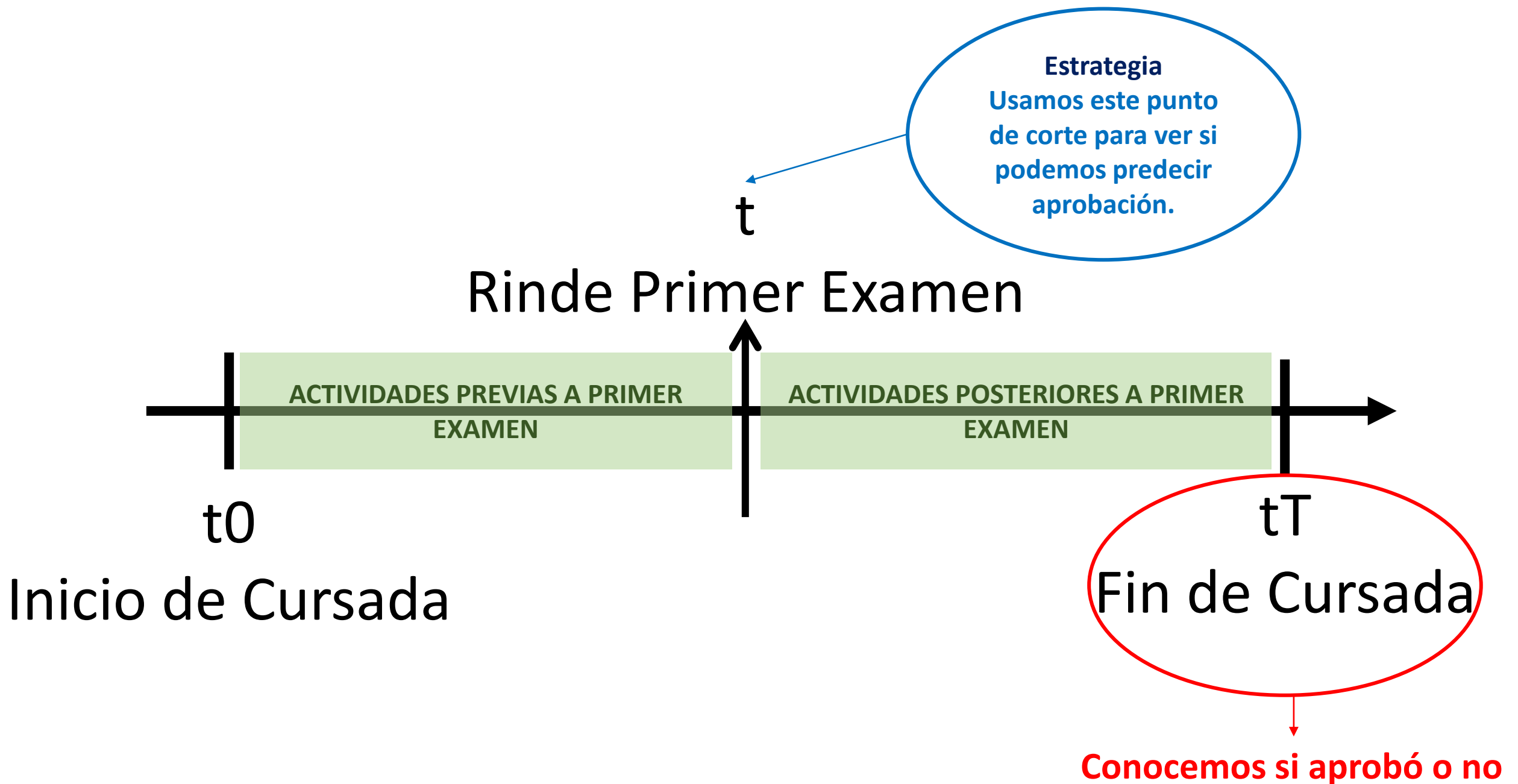
Dado esto, ¿podemos predecir si  
un alumno va a aprobar o no una  
materia?

Hay múltiples posibles  
estrategias o caminos para  
responder esto con los datos  
disponibles: elegimos una.









Procesamos información para  
obtener características de la  
cursada hasta el día del primer  
examen.

## Características (Features):

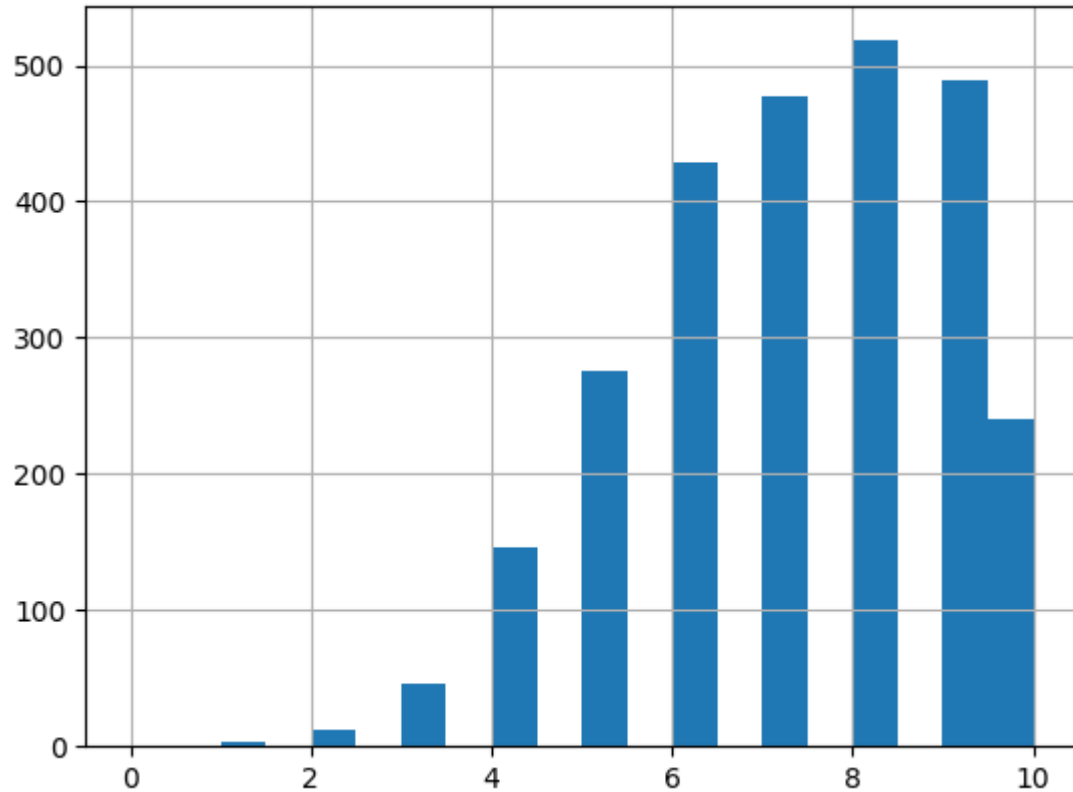
- Nota del primer examen.
- Score promedio de actividades previas a primer examen.
- Tiempo promedio de entrega de actividades.
- Cantidad de días desde inicio hasta el primer examen.

Target:

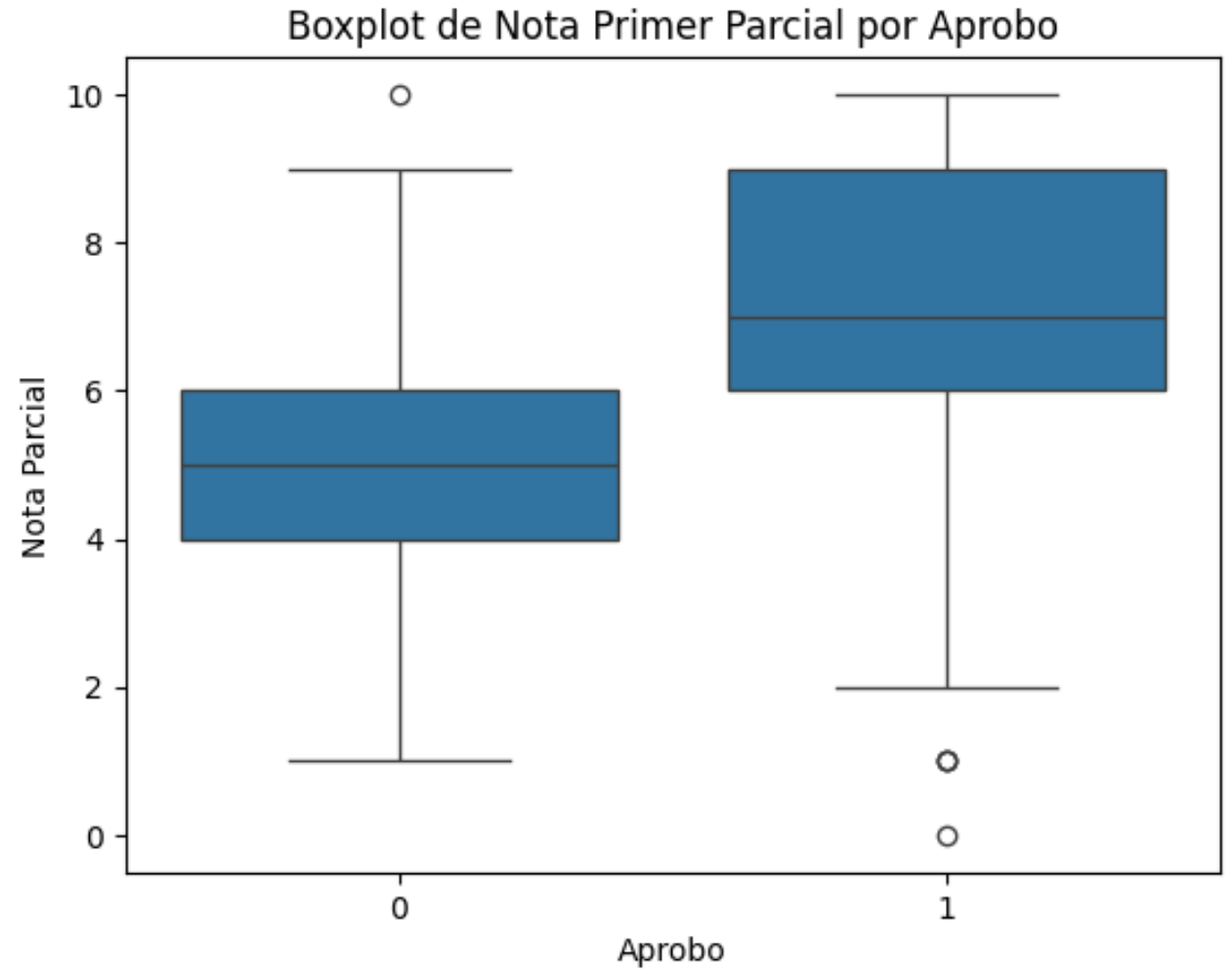
¿Desaprobó la materia?

# Análisis Descriptivo

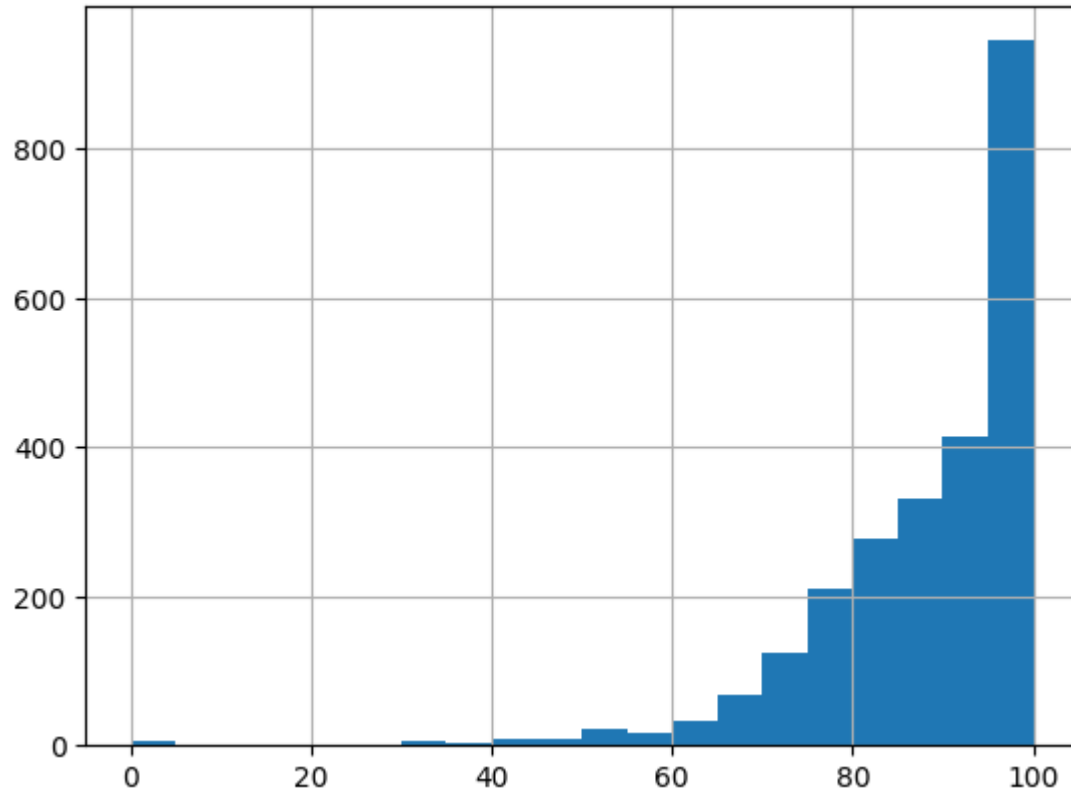
# Nota de Primer Parcial



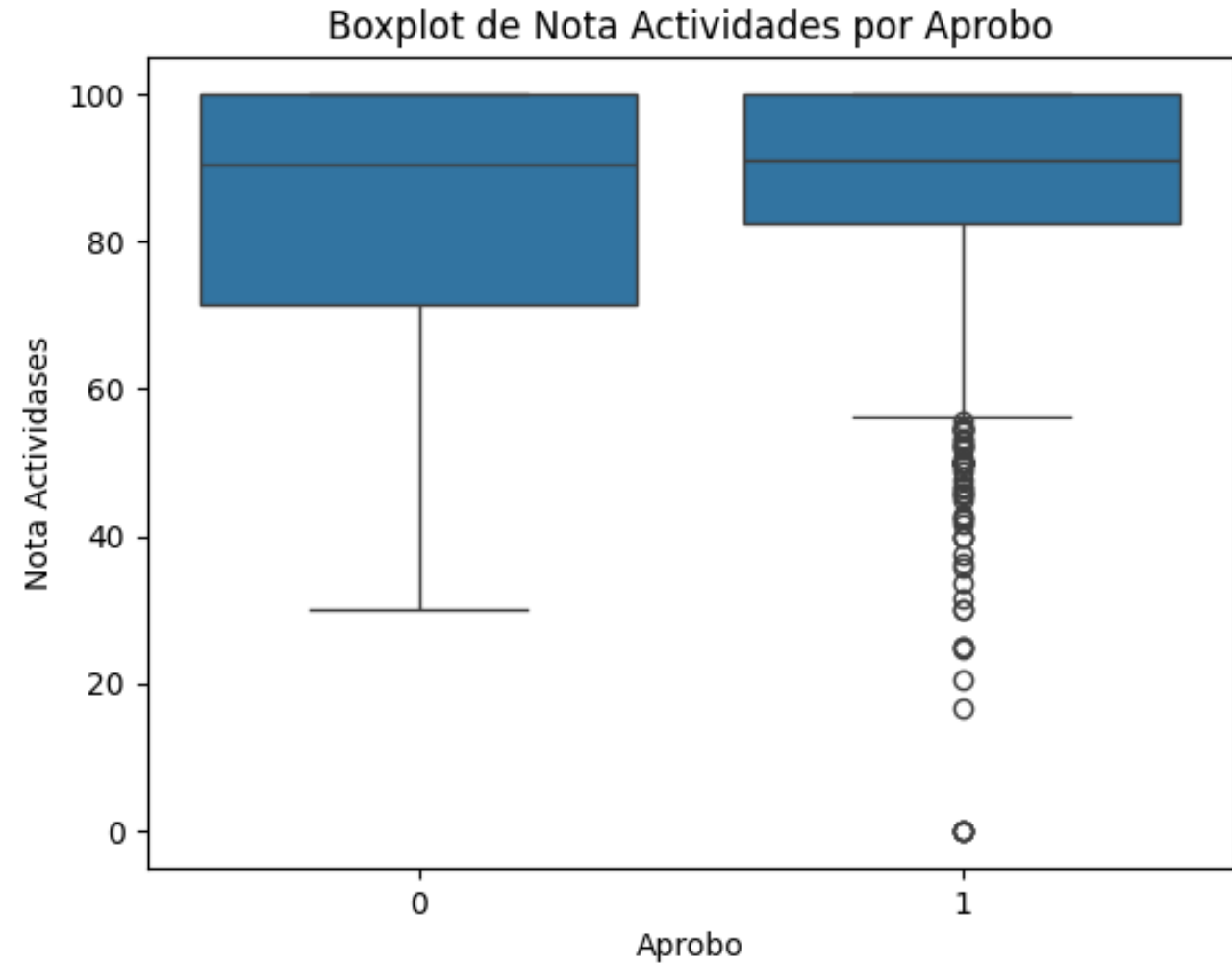
El promedio es 7.2 y la mediana es 6, distribución con cola a la izquierda.



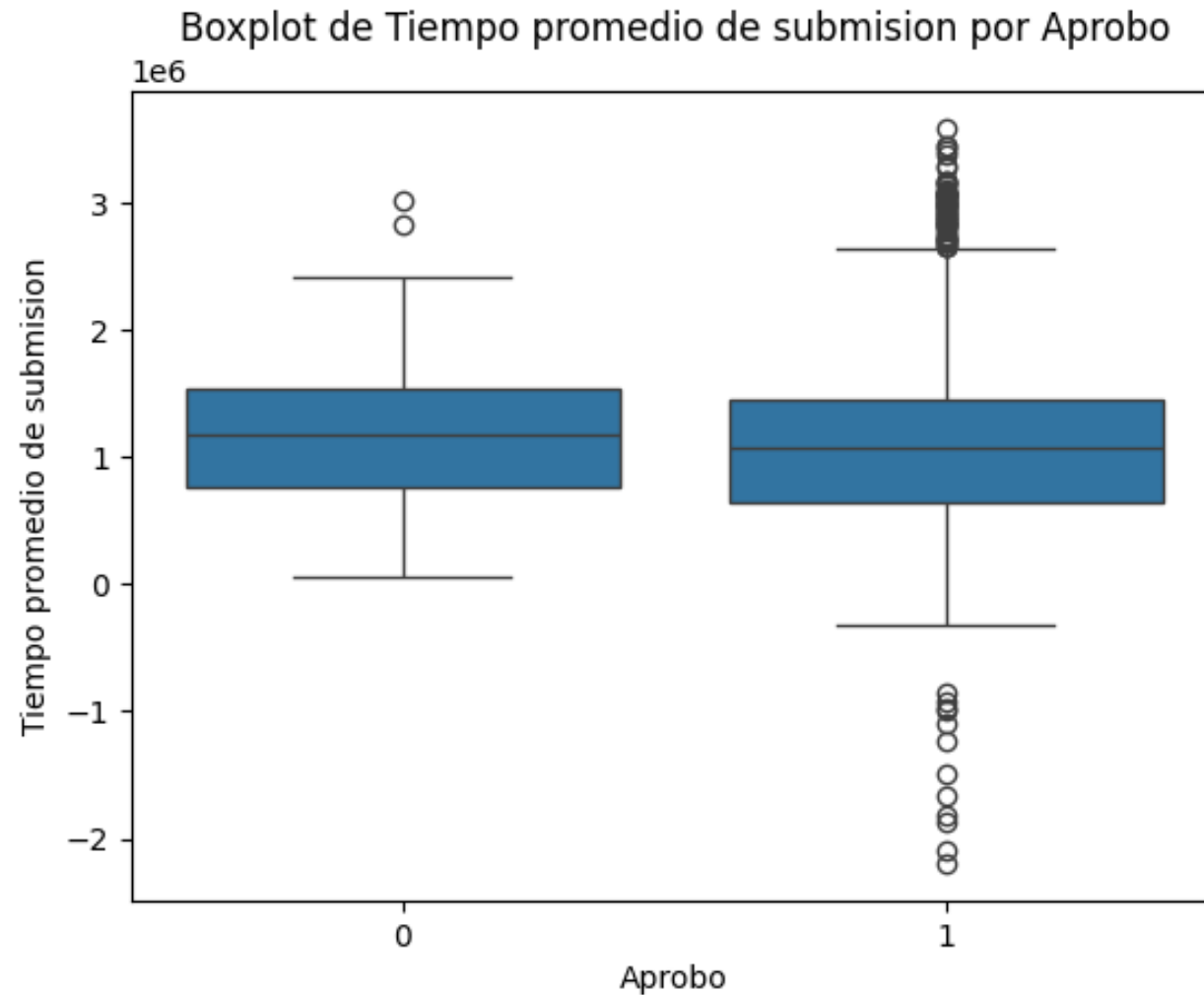
# Score de Actividades



El promedio es 88.3 y la mediana es 91, distribución con cola a la izquierda.

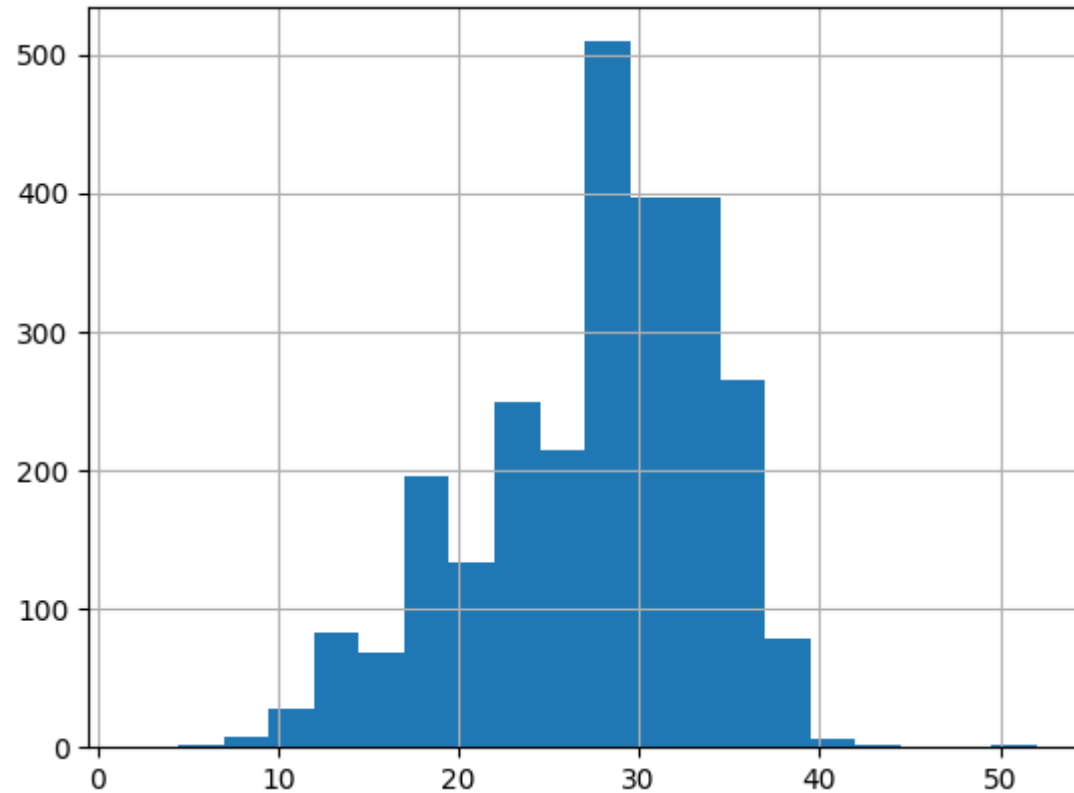


# Tiempo hasta entrega de actividades

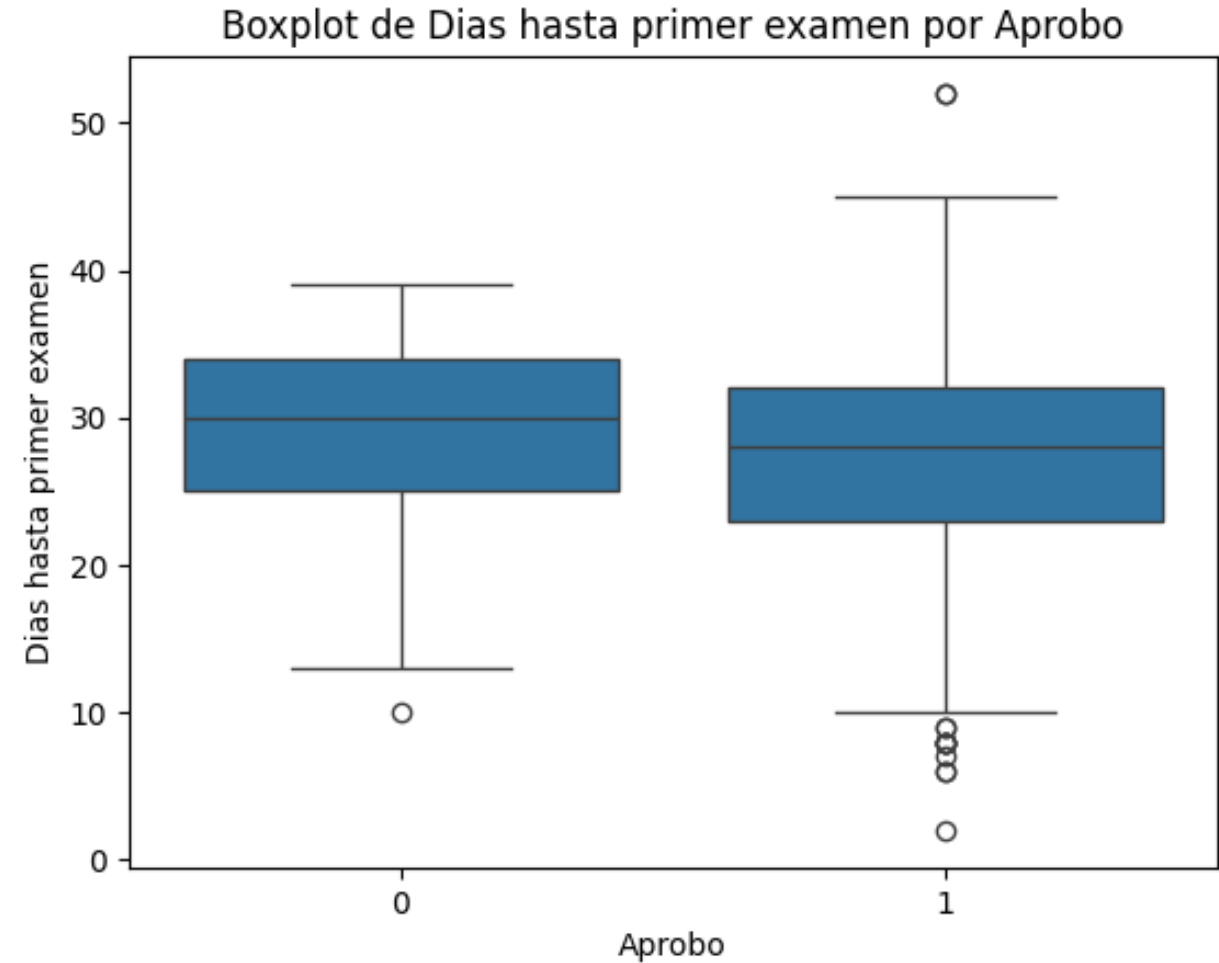




# Días hasta primer examen



El promedio es 27 y la mediana es 28, la distribución es aproximadamente simétrica.



# Aprendizaje Automático

Target:

¿Desaprobó la materia?

Se hizo un estudio preliminar  
para ver qué algoritmos  
tendían a predecir mejor

# Top 10

De cada 100 cursadas que desaprobaron en la realidad, ¿cuántas es capaz el modelo de predecir?

	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	recall_score	Time Taken
Model						
BernoulliNB	0.70	0.77	0.77	0.80	0.85	0.01
PassiveAggressiveClassifier	0.80	0.75	0.75	0.87	0.70	0.01
Perceptron	0.75	0.75	0.75	0.84	0.75	0.01
NearestCentroid	0.76	0.73	0.73	0.84	0.70	0.01
QuadraticDiscriminantAnalysis	0.84	0.70	0.70	0.89	0.55	0.01
CalibratedClassifierCV	0.85	0.68	0.68	0.90	0.50	0.03
LogisticRegression	0.85	0.68	0.68	0.90	0.50	0.01
LinearSVC	0.85	0.68	0.68	0.90	0.50	0.07
RidgeClassifierCV	0.84	0.68	0.68	0.89	0.50	0.01
RidgeClassifier	0.84	0.68	0.68	0.89	0.50	0.01

Tomamos el mejor y sumamos  
algunos de los usados en  
general por la ciencia de datos

## Bernoulli Naive Bayes

	precision	recall	f1-score	support
0	0.99	0.69	0.82	772
1	0.07	0.85	0.12	20
accuracy			0.70	792
macro avg	0.53	0.77	0.47	792
weighted avg	0.97	0.70	0.80	792

## Extreme Gradient Boosting

	precision	recall	f1-score	support
0	0.98	0.97	0.97	772
1	0.08	0.10	0.09	20
accuracy			0.95	792
macro avg	0.53	0.53	0.53	792
weighted avg	0.95	0.95	0.95	792

## Red Neuronal

	precision	recall	f1-score	support
0	0.98	0.85	0.91	772
1	0.08	0.50	0.13	20
accuracy			0.84	792
macro avg	0.53	0.67	0.52	792
weighted avg	0.96	0.84	0.89	792

# Anexo:

# Importancia de las Features



# Usamos Regresión Logística para obtener el orden de importancia

1. nota\_parcial (1.57)
2. tiempo\_hasta\_submision (0.13)
3. Score (0.11)
4. dias\_hasta\_primer\_examen (0.02)

# ¡Gracias!