**¿Por qué importa A/B Testing?**

A/B Testing es una herramienta fundamental para tomar decisiones basadas en evidencia. En lugar de asumir que un cambio en un modelo, producto o estrategia será mejor, lo probamos directamente con datos. Esto evita implementar cambios que podrían no aportar mejoras o incluso empeorar el rendimiento.

En el contexto de modelos predictivos, el A/B testing permite comparar modelos y verificar si realmente hay una mejora estadísticamente significativa en métricas como accuracy, F1, RMSE o AUC. Sin A/B testing, podríamos estar implementando modelos que solo parecen mejores por casualidad o por overfitting.

**2. ¿Qué tipos de A/B Testing hay?**

Existen diferentes tipos de A/B testing, dependiendo de qué se quiere probar y cómo se diseña el experimento.

* El **A/B testing clásico** compara dos variantes (modelo A contra modelo B). Se mide una métrica objetivo en ambos grupos y se prueba si la diferencia es significativa.
* El **A/B/n** extiende esto a más de dos variantes. Por ejemplo, podrías comparar tres modelos distintos para ver cuál es mejor.
* El **multivariante** no compara variantes completas, sino combinaciones de elementos. Por ejemplo, se prueban distintas combinaciones de parámetros o técnicas para entender no solo cuál funciona mejor, sino **por qué**.
* El **testing secuencial** permite analizar los resultados a medida que llegan los datos. Esto puede acelerar decisiones si una variante se vuelve claramente superior desde el inicio. Aquí entra el enfoque SPRT (Sequential Probability Ratio Test), que permite detener un experimento si hay suficiente evidencia.
* La **inferencia causal** se usa cuando no puedes hacer una asignación completamente aleatoria (como en A/B clásico), pero quieres estimar el efecto de un modelo, tratamiento o política, controlando variables que podrían sesgar el análisis.
* El **multi-armed bandit** reemplaza el A/B tradicional con un enfoque adaptativo. En lugar de asignar tráfico de manera fija, el sistema va ajustando las probabilidades de asignación según qué opción va funcionando mejor, maximizando el beneficio mientras aprende.

**3. ¿Qué enfoques existen para analizar A/B testing?**

Hay dos enfoques principales: el **frecuentista** y el **bayesiano**.

* El enfoque **frecuentista** formula una hipótesis nula ("no hay diferencia entre A y B") y calcula un p-valor: la probabilidad de observar una diferencia como la que se ha medido si la hipótesis nula fuera cierta. Si el p-valor es menor a un umbral (por ejemplo, 0.05), se rechaza la hipótesis nula. Este enfoque es el más tradicional y el más usado.

**Enfoque del Muestreo Frecuentista**

**📌 Principios:**

* La probabilidad es la **frecuencia relativa de un evento** si se repite muchas veces.
* No hay "creencias" previas, solo se trabaja con los datos observados.
* Los parámetros (como la media, la proporción, etc.) son **fijos pero desconocidos**.
* Los datos son **aleatorios**, porque provienen de un muestreo aleatorio.

**🔍 Muestreo frecuentista:**

* Seleccionas una muestra aleatoria de la población.
* Haces inferencia sobre parámetros fijos usando **estadísticas muestrales**.
* Por ejemplo: Si llamas a 10,000 personas de un total de 70,000, asumes que esas 10,000 representan la población completa **si se seleccionaron al azar**.

**✅ Ventajas:**

* Métodos clásicos, bien documentados.
* Uso de pruebas como z-test, t-test, chi-cuadrado.
* Muy útil cuando tienes una muestra grande y aleatoria.

**⚠️ Limitaciones:**

* No permite incorporar **conocimiento previo**.
* No puedes actualizar la inferencia si vas obteniendo más datos (sin rehacer todo el análisis).
* No puedes hacer inferencias sobre individuos, solo sobre parámetros poblacionales.
* El enfoque **bayesiano** parte de una creencia previa y actualiza esa creencia con los datos observados. En lugar de hablar de p-valores, calcula directamente la probabilidad de que una variante sea mejor que otra. Por ejemplo, podrías decir que "hay un 85% de probabilidad de que el modelo A sea mejor que el modelo B".

**Enfoque del Muestreo Bayesiano**

**📌 Principios:**

* La probabilidad refleja **grados de creencia** (subjetivos o informados).
* Se combina el conocimiento previo (**prior**) con la evidencia nueva (**likelihood**) para obtener una **posterior**.
* Los parámetros **también son aleatorios** y se modelan con distribuciones de probabilidad.

**🔍 Muestreo bayesiano:**

* Se parte de una distribución **a priori** sobre el parámetro (por ejemplo, la tasa de contacto histórica).
* A medida que recolectas datos, los vas **acumulando y actualizando** tu inferencia.
* Ejemplo: Si al principio crees que la tasa de contacto es ~16%, y ves que en tus primeras 1,000 llamadas contesta el 20%, entonces tu nueva estimación se ajusta en función de ambas.

**✅ Ventajas:**

* Se puede actualizar continuamente con datos nuevos (ideal para tests secuenciales o adaptativos).
* Mejor interpretación: puedes decir cosas como “hay un 95% de probabilidad de que el modelo B sea mejor que A”.
* Útil con muestras más pequeñas (si tienes un buen prior).
* Permite incorporar experticia, datos históricos u otros modelos.

Ambos enfoques son válidos, pero tienen ventajas distintas. El bayesiano es más intuitivo para tomar decisiones prácticas, mientras que el frecuentista es más conservador y tradicionalmente aceptado.

**SPRT (Prueba de razón de verosimilitud secuencial)** es un método estadístico creado por Abraham Wald que permite tomar decisiones entre dos hipótesis **a medida que los datos se recogen**, sin esperar a que termine todo el experimento. Es muy útil cuando quieres tomar decisiones **rápidas y eficientes**.

**⚖️ ¿Cómo funciona?**

Supón que quieres comparar dos versiones de un modelo (o estrategia) y definir:

* **H₀ (hipótesis nula)**: el modelo A y el modelo B son igual de buenos.
* **H₁ (hipótesis alternativa)**: el modelo B es mejor que A.

A diferencia de un A/B testing clásico (que fija el tamaño de muestra de antemano), SPRT **analiza los datos uno por uno o por lotes** y calcula la **razón de verosimilitud**

Luego se compara con dos umbrales Donde los valores A y B dependen de los niveles de error que toleras

**Por qué es útil?**

* ✅ Permite **detener el experimento antes** si ya tienes suficiente evidencia.
* ✅ Reduce el número de muestras requeridas en muchos casos.
* ✅ Útil en entornos donde cada muestra cuesta dinero o tiempo (como tests online, medicina, modelos ML en producción...).

**🤖 ¿Se puede usar para modelos de ML?**

¡Sí! Puedes usarlo para:

* Evaluar si un modelo nuevo es significativamente mejor (por accuracy, log loss, etc.)
* Decidir si implementar un modelo en producción más rápido
* Comparar tasas de error entre dos clasificadores en línea

**🧠 Comparación con A/B Frecuentista y Bayesiano**

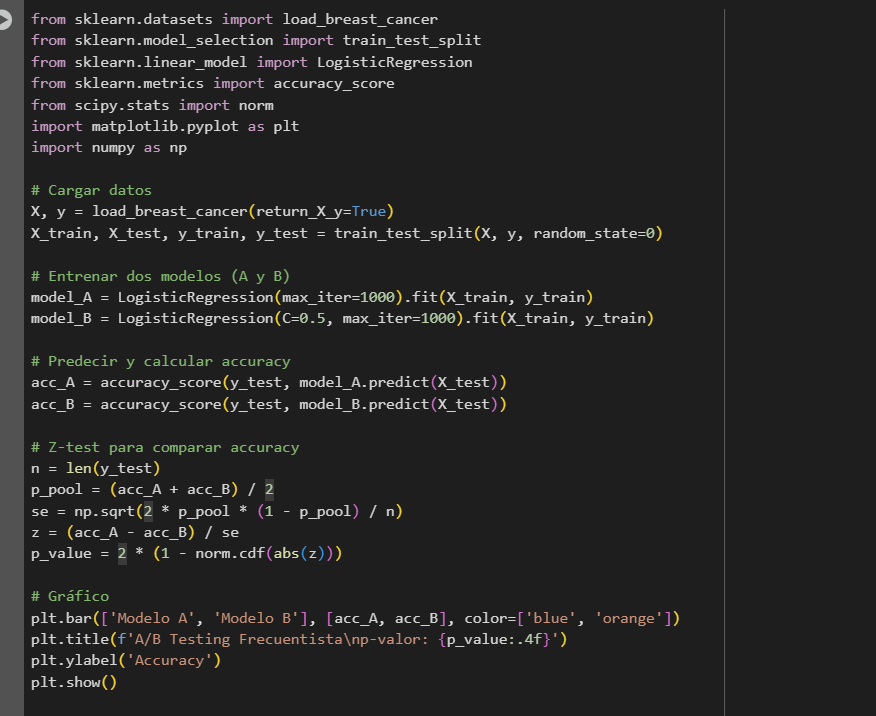
| **Método** | **Toma decisiones…** | **Usa tamaño fijo** | **Basado en…** |
| --- | --- | --- | --- |
| A/B clásico | al final | Sí | p-valor, z-test |
| SPRT | durante el test | No | likelihood ratio |
| Bayesiano | durante el test | No | distribución posterior |

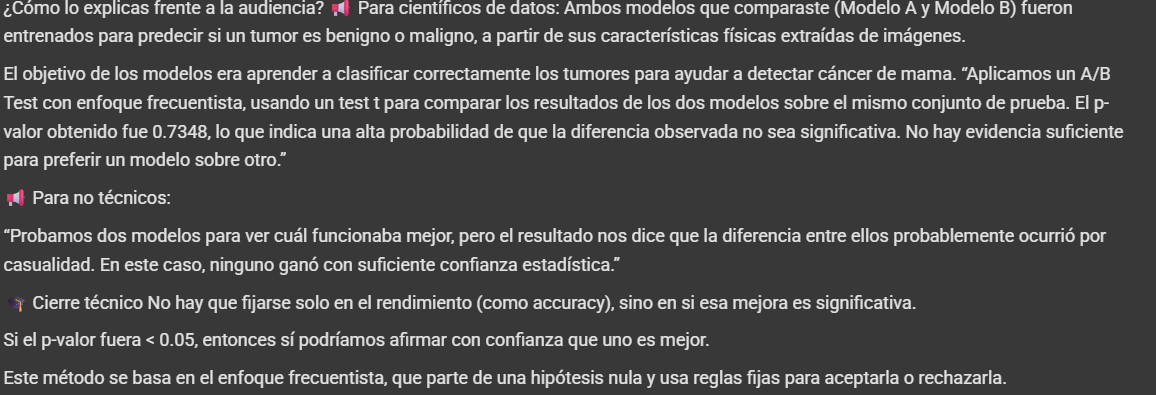
**4. ¿Cómo se implementan?**

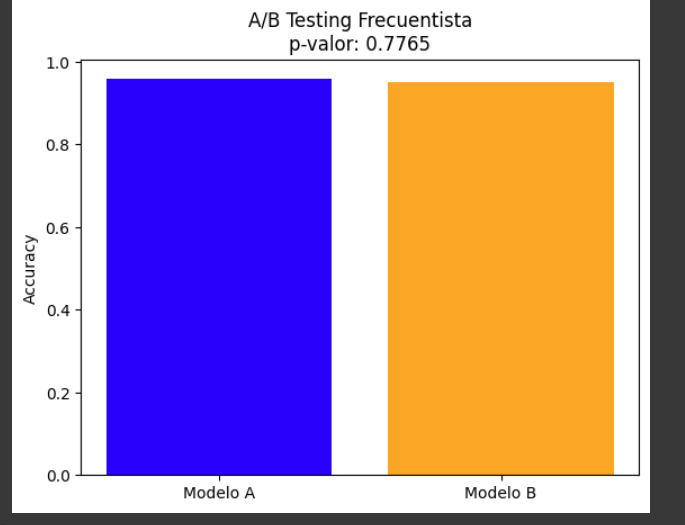
Para aplicar A/B testing, debes:

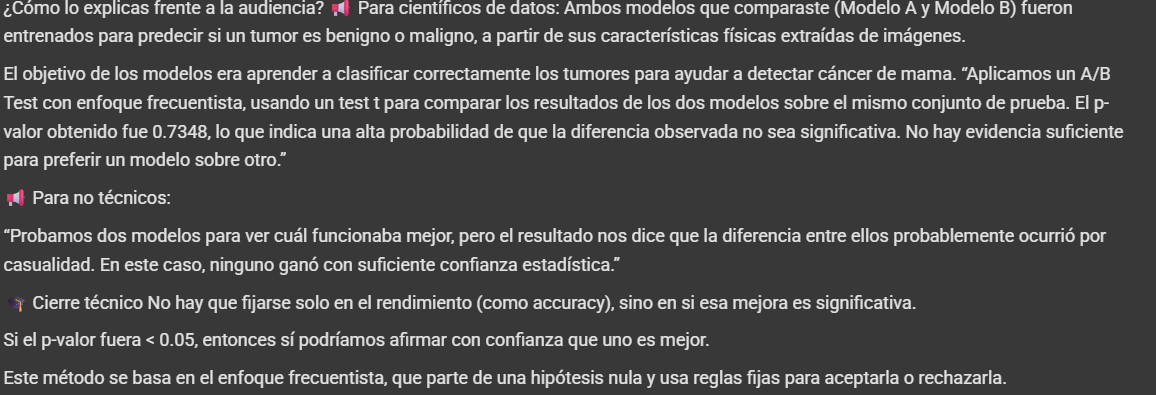
1. Definir una **métrica de éxito** clara (por ejemplo, accuracy, AUC, tiempo de respuesta).
2. Calcular el **tamaño mínimo de muestra** necesario para detectar diferencias significativas (esto se puede hacer con herramientas online).
3. Asegurar una **asignación aleatoria** entre grupos.
4. Ejecutar el experimento sin detenerlo antes de tiempo, a menos que uses un enfoque secuencial.
5. Analizar los resultados con un enfoque estadístico claro (frecuentista o bayesiano).

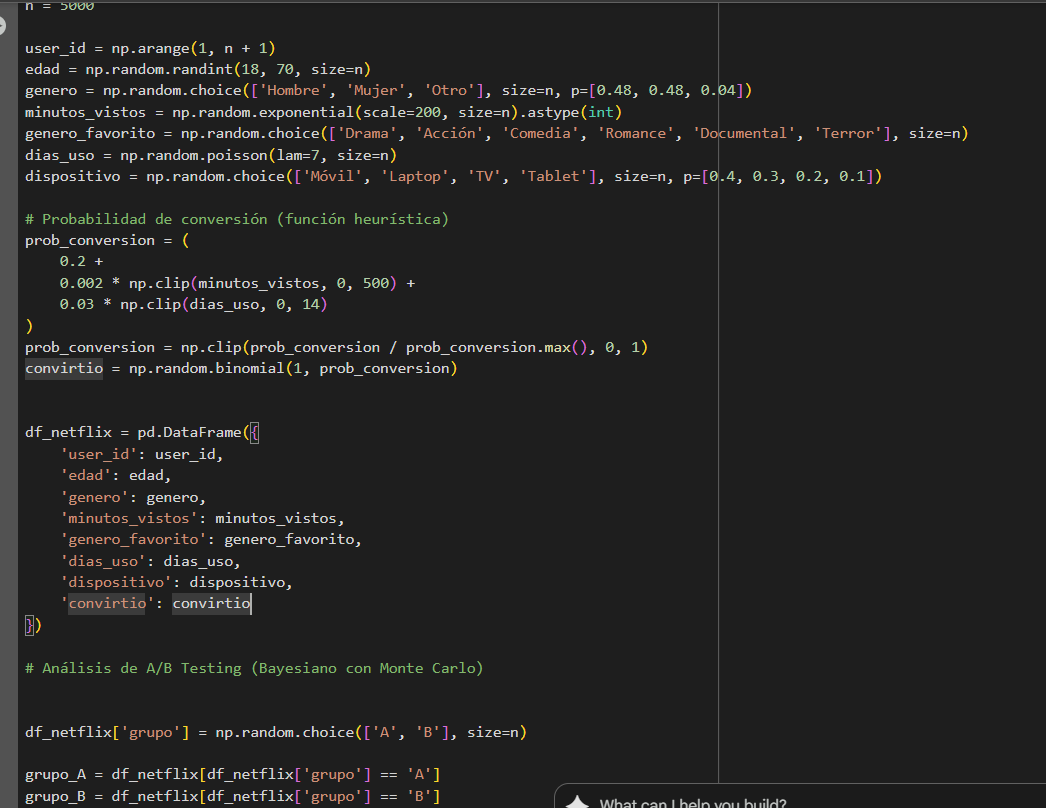
Para modelos de machine learning, el A/B testing puede aplicarse en un entorno simulado o en producción. En simulación, se puede usar un conjunto de validación o test para dividir entre modelos A y B, y comparar métricas. En producción, puedes redirigir tráfico de usuarios reales o peticiones de predicción a los distintos modelos.

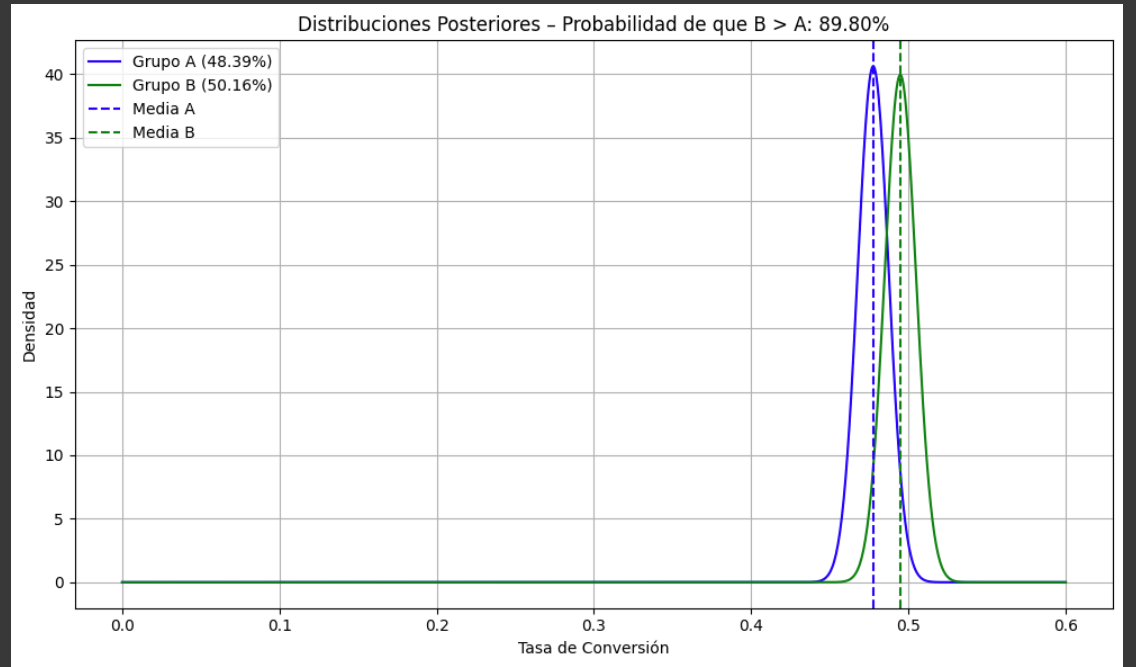












Realizamos un experimento A/B con 5,000 usuarios, asignados aleatoriamente a dos versiones de una experiencia de prueba gratuita en un servicio estilo Netflix:

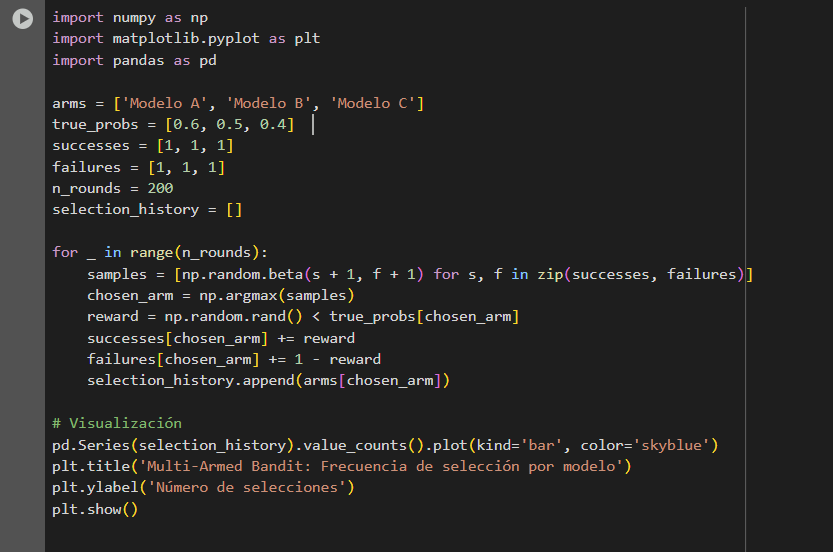
* **Grupo A**: Versión actual de la prueba gratuita.
* **Grupo B**: Versión alternativa (por ejemplo, con recomendaciones personalizadas o UI diferente).

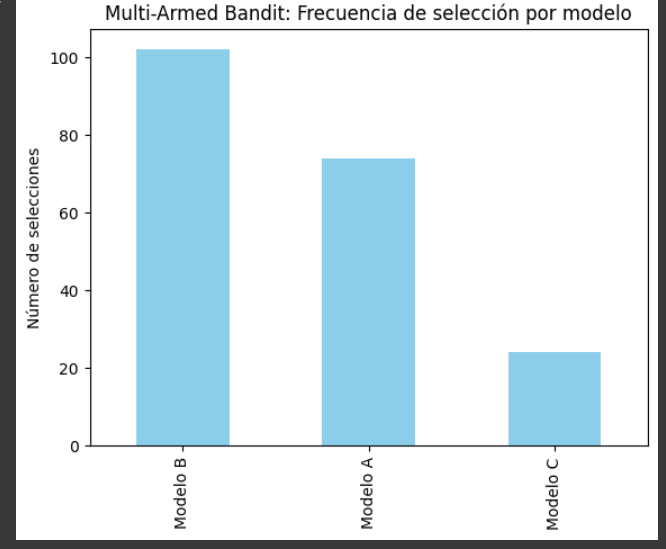
El objetivo es determinar **cuál grupo logra una mayor conversión a suscripción** al final del periodo de prueba.

Este gráfico representa la **incertidumbre y distribución posterior** de la **tasa de conversión** para cada grupo, utilizando un enfoque **bayesiano con simulación Monte Carlo**:

* **Curva azul (Grupo A)**:
  + Tasa estimada de conversión: **48.39%**
  + Distribución posterior de la tasa de conversión bajo la evidencia observada.
* **Curva verde (Grupo B)**:
  + Tasa estimada de conversión: **50.16%**
  + Distribución posterior correspondiente al grupo B.
* **Líneas punteadas** indican la **media** de cada distribución.
* **Grupo B tiene una tasa de conversión más alta**, aunque la diferencia no es extrema.
* **La probabilidad de que el Grupo B tenga una mayor tasa de conversión que el Grupo A es del 89.80%.**
* Este resultado fue obtenido usando **10,000 simulaciones Monte Carlo** de las distribuciones posteriores. Si bien no alcanza el umbral típico del 95% de certeza, **la evidencia favorece a la variante B**

**Multi banded**





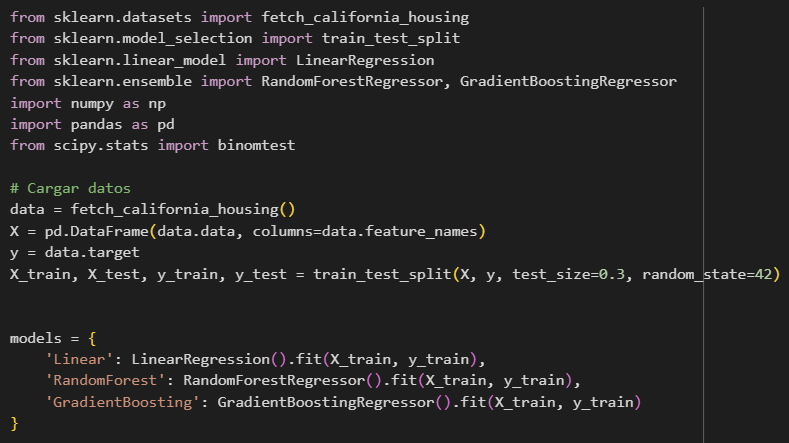
**¿Qué hace el sistema? bayesiano**

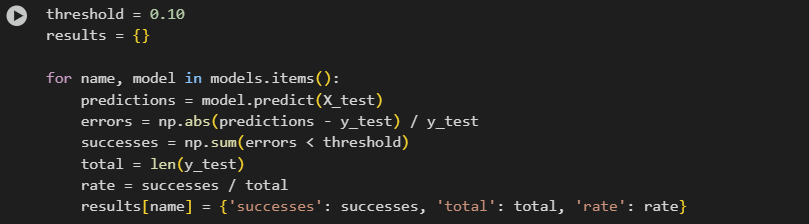
* Compara tres formas distintas de predecir el precio de una casa.
* Después de cada predicción, evalúa si fue suficientemente buena (error menor al 10%).
* Usa un método inteligente llamado **Thompson Sampling** para decidir cuál modelo probar en cada caso

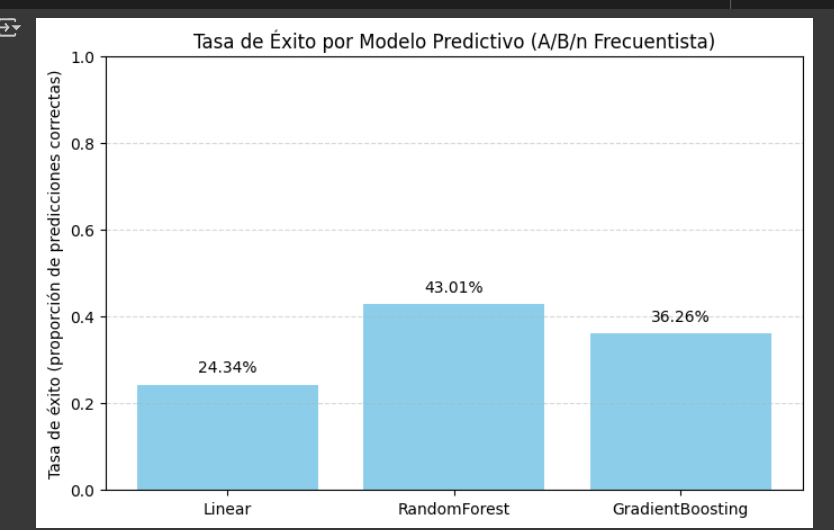
 ¿Por qué? Porque fue el que **más veces acertó** dentro del margen aceptable.

 Pero el sistema **no se casa con un solo modelo**: sigue probando los otros dos, por si en algún momento empiezan a funcionar mejor.

ABN



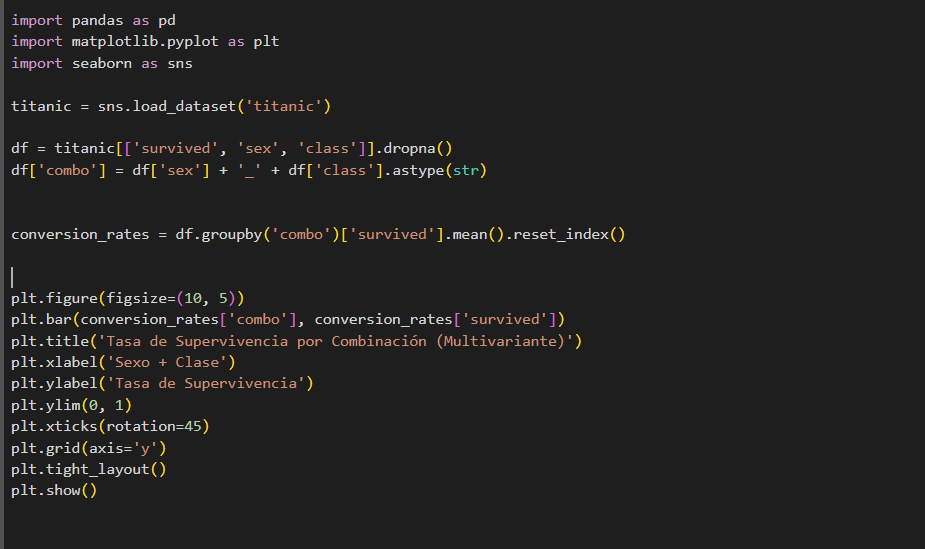


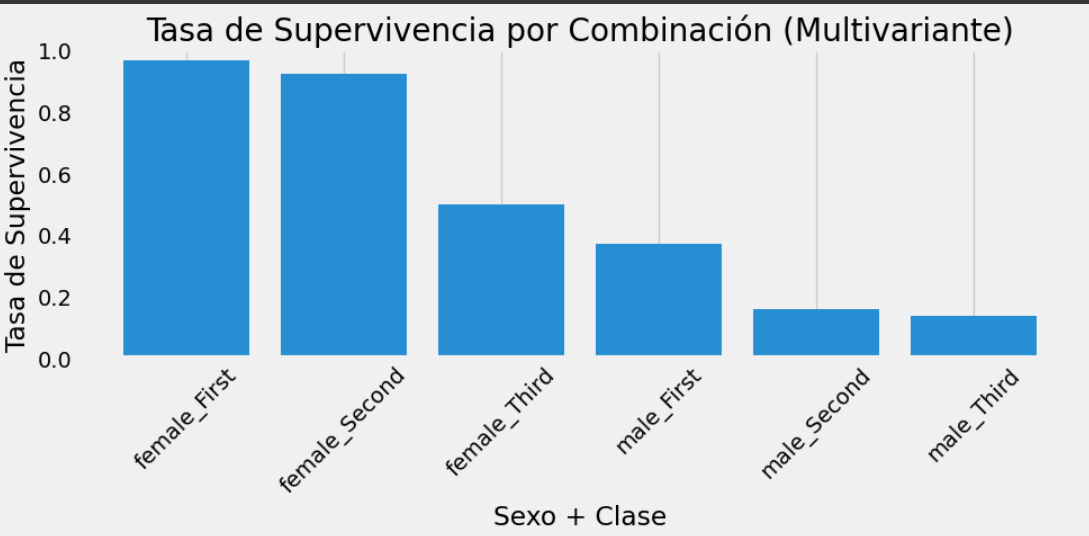


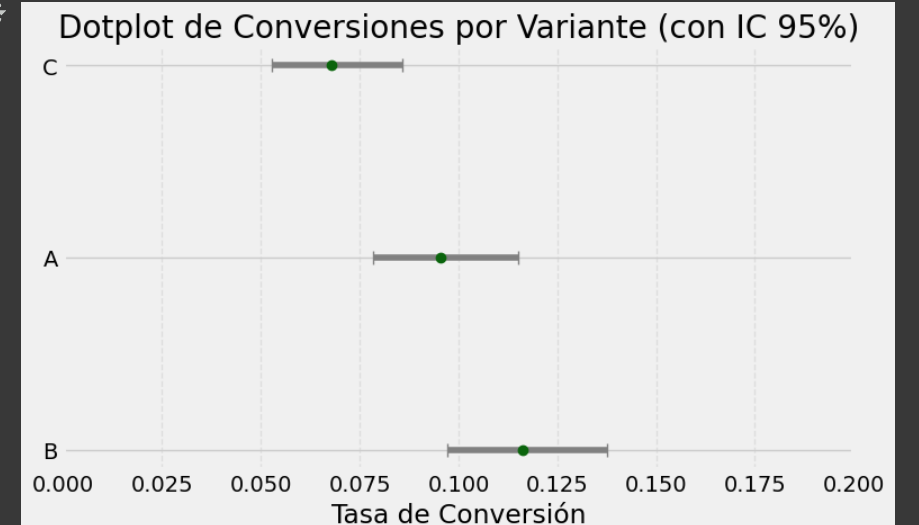
El A/B/n test mide el **rendimiento absoluto** de cada modelo en términos de una métrica binaria (por ejemplo, éxito vs. fracaso). Cada modelo se evalúa sobre una muestra independiente o estratificada, y se calcula:

* **Tasa de éxito**: proporción de casos en que el modelo acertó.
* **Intervalo de confianza**: rango donde esperamos que esté la tasa real.
* **Significancia estadística**: si las diferencias entre modelos son suficientemente grandes como para no atribuirlas al azar

MULTIVARIANTES







* **Tres variantes** (A, B, C) de un A/B/n test.
* Cada punto representa la **tasa de conversión estimada** para esa variante.
* Las líneas horizontales son **intervalos de confianza del 95%**, calculados con métodos binomiales clásicos.

: No hay evidencia clara de que una variante supere significativamente a las demás. Los intervalos se solapan, y si el p-value del chi2\_contingency es > 0.05, no se rechaza la hipótesis nula de igualdad de proporciones.

**ué muestra esta gráfica?**

* **Eje Y (vertical)**: Cada grupo multivariante, definido por la combinación de sexo y clase. Ejemplos:
  + male\_Third: hombres en tercera clase
  + female\_First: mujeres en primera clase
* **Eje X (horizontal)**: La **tasa de supervivencia** para cada grupo, entre 0 y 1.
* **Punto azul**: Es la **estimación puntual** de la tasa de supervivencia (proporción de personas que sobrevivieron en ese grupo).
* **Líneas grises horizontales**: Representan el **intervalo de confianza del 95%** para esa tasa. Cuanto más cortas, más precisa es la estimación.

**🔍 ¿Qué insights ofrece?**

* **female\_First** tiene una tasa de supervivencia cercana a 1.0 con un intervalo muy estrecho → alta supervivencia y alta certeza.
* **male\_Third** tiene la tasa más baja y un intervalo más amplio → baja supervivencia y más incertidumbre.
* Puedes ver claramente cómo el **sexo** y la **clase** interactúan: ser mujer y estar en primera clase es altamente protector, mientras que ser hombre en tercera clase es lo opuesto.

 El enfoque **bandit** es ideal para sistemas en producción que deben **tomar decisiones en tiempo real** y adaptarse.

 El enfoque **frecuentista** es mejor para **evaluaciones comparativas estáticas**, como informes, benchmarks o validaciones previas al despliegue.

**Dataset:** [IHDP (Infant Health and Development Program)] — (disponible en causalml, simulado aquí)  
**Ejemplo:** estimar el efecto causal de un "tratamiento" en regresión (por ejemplo, una intervención médica)

**Qué es:** no solo saber si “A funciona mejor que B”, sino entender si **A *causa* un efecto.**

**Ejemplo:**  
En un estudio de bebés, algunos recibieron un tratamiento especial. Queremos saber si eso **realmente ayudó a mejorar su desarrollo**, no solo si hay diferencia entre los grupos.

para **simular qué habría pasado si todos hubieran recibido o no el tratamiento.**

🔍 *Es como preguntarse: ¿el tratamiento ayudó, o esos niños habrían mejorado igual sin él?*

**Nivel de Eficiencia**

* **⚡⚡⚡⚡ (Muy alta):** Bandits (reducen desperdicio de muestras), adaptan su comportamiento.
* **⚡⚡⚡ (Alta):** Bayesian A/B y secuencial (pueden detenerse temprano, menos muestras).
* **⚡⚡ (Moderada):** A/B clásico, multivariante.
* **⚡ (Baja):** A/B/n (muchas comparaciones = más muestras necesarias).

herramientas

**tatsig (para A/B Testing)**

Statsig es una plataforma que facilita la creación, ejecución y análisis de experimentos A/B de forma automatizada y confiable.

🔹 ¿Qué hace?

* Automatiza el diseño experimental (grupos, métricas, duración, etc.).
* Calcula automáticamente p-valores, intervalos de confianza y significancia estadística.
* Ofrece métodos avanzados como tests bayesianos, segmentación automática y análisis de impacto.
* Permite tomar decisiones basadas en datos sin necesidad de escribir todo el código estadístico.

🧠 En resumen:  
👉 Te ayuda a hacer A/B testing más rápido, con menos errores y con visualizaciones claras para stakeholders.

**⚙️ MLflow (para ML + A/B Testing)**

MLflow es una herramienta de tracking de experimentos y gestión del ciclo de vida de modelos de machine learning.

🔹 ¿Qué hace?

* Guarda cada experimento de entrenamiento: código, parámetros, métricas, versiones del modelo.
* Permite comparar resultados de múltiples modelos (como si hicieras un A/B de modelos).
* Se integra con pipelines para desplegar modelos y llevar un seguimiento de su rendimiento en producción.

🧠 En resumen:  
👉 Es útil si haces A/B testing entre modelos de ML y quieres saber cuál rinde mejor y por qué, todo registrado automáticamente.

1. ¿Qué diferencia hay entre Statsig y MLflow?

* Respuesta: Statsig está diseñado para experimentación A/B (sobre producto, UI, etc.), mientras que MLflow es para rastrear experimentos y despliegues de modelos ML. Uno está orientado a decisiones de producto, el otro a modelos de predicción.

1. ¿Necesito saber estadística avanzada para usar Statsig?

* Respuesta: No necesariamente. Statsig automatiza los cálculos más complejos, como p-valores, intervalos y segmentaciones. Pero sí ayuda entender los fundamentos para interpretar bien los resultados.

 Respuesta detallada por tipo de test:

1. 🧪 A/B clásico (frecuentista)
   * ✔️ Puedes añadir nueva población, pero:
     + Debes mantener la aleatorización (random assignment).
     + No debes mirar resultados mientras agregas usuarios (evitar “p-hacking”).
   * ⚠️ Si usas tamaño de muestra fijo, agregar más usuarios después de mirar resultados puede invalidar el test (inflas la tasa de error tipo I).
2. 📈 A/B secuencial (como SPRT o test bayesiano)
   * ✔️ Diseñados para eso: permiten observar resultados a medida que se agregan nuevos datos.
   * Ideal para entornos dinámicos donde el tráfico llega por oleadas.
   * El SPRT, por ejemplo, evalúa continuamente si detener o continuar según evidencia acumulada.
3. 🤖 Multi-armed bandit
   * ✔️ Muy flexible.
   * Está hecho para adaptarse a nueva información y redistribuir tráfico.
   * Puedes añadir usuarios constantemente, y el sistema ajusta las probabilidades de asignación con el tiempo.
4. 🧠 Bayesiano
   * ✔️ Permite incorporación natural de nuevos datos.
   * Puedes actualizar la creencia posterior sin afectar la validez del experimento.
   * Ideal si tu tráfico cambia de forma orgánica o necesitas decisiones continuas.

—

Sobre problemas comunes

1. ¿Qué pasa si interrumpo un test antes de tiempo?

* Respuesta: Puedes obtener un resultado falso (positivo o negativo) por no tener suficiente evidencia estadística. Es por eso que se definen criterios de detención desde el inicio o se usa un enfoque secuencial controlado.

1. ¿Qué tamaño de muestra necesito?

* Respuesta: Depende de la métrica, la variación esperada y el efecto mínimo detectable. Hay calculadoras online (como en Statsig) que ayudan a estimarlo automáticamente.

1. ¿Cómo mido el éxito de un A/B test?

* Respuesta: Debes tener una métrica primaria (como conversión, revenue, precisión del modelo) y definir umbrales. El éxito es cuando hay diferencia significativa a favor de la variante y cumple con los criterios de negocio.

📊 Sobre comunicación de resultados

1. ¿Cómo explicar estos resultados a personas no técnicas?

* Respuesta: Usa visualizaciones simples, evita jerga estadística y enfócate en: cuál fue la hipótesis, qué se probó, qué resultado se obtuvo, y cómo impacta en el negocio o producto.

1. ¿Por qué confiar en un test si solo cambió un pequeño porcentaje?

* Respuesta: Porque incluso pequeños cambios pueden tener impacto significativo cuando se escalan a millones de usuarios. Lo importante es que la diferencia sea estadísticamente confiable.