**Actividad: Prueba de hipótesis, Regresión lineal y Regresión logística por Michael Jamaica**

# Índice

1) Dataset  
2) Prueba de hipótesis  
3) Regresión lineal múltiple  
4) Regresión logística  
5) Conclusiones

# 1) Dataset

Primero inventé un dataset de 300 estudiantes con variables como horas de estudio, horas de sueño, índice socioeconómico y GPA previo. Con eso construí una nota (0-100) y una variable binaria de si aprueba (>=60).

|  |
| --- |
| # Creo dataset inventado  np.random.seed(42)  n = 300  metodo = np.random.choice(['A', 'B'], size=n)  horas\_estudio = np.clip(np.random.normal(10, 3, n), 0, None)  horas\_suenio = np.clip(np.random.normal(7, 1.2, n), 3, 12)  indice\_socioeco = np.clip(np.random.normal(50, 10, n), 20, 90)  gpa\_previo = np.clip(np.random.normal(3.2, 0.4, n), 0, 4.0)  efecto\_metodo = np.where(metodo == 'A', 2.0, 0.0)  ruido = np.random.normal(0, 5, n)  puntuacion = (4.0\*horas\_estudio + 1.5\*horas\_suenio + 0.3\*indice\_socioeco + 8.0\*gpa\_previo + efecto\_metodo + ruido)  puntuacion = np.clip(puntuacion, 0, 100)  aprueba = (puntuacion >= 60).astype(int) |

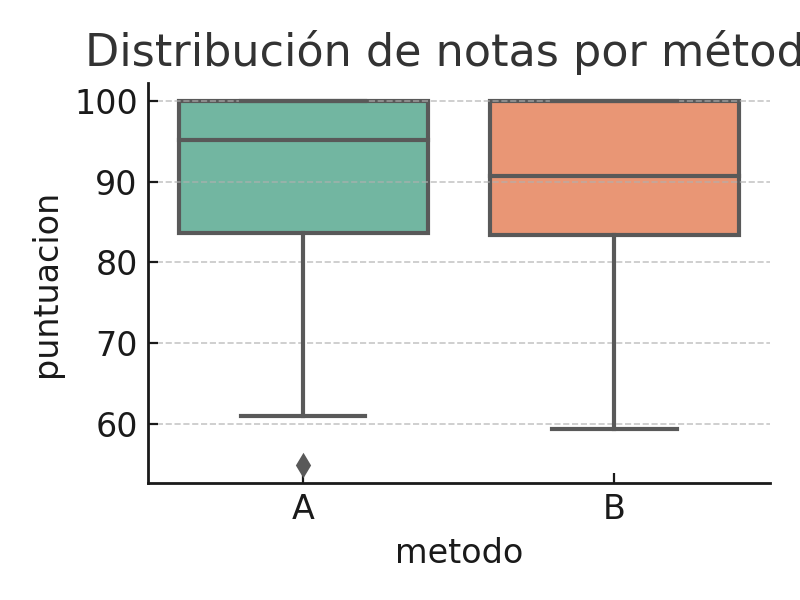
Así queda un dataset con la información de los estudiantes.

# 2) Prueba de hipótesis

Hipótesis:  
H0: la media de las notas en método A y método B es igual.  
H1: la media es diferente.

|  |
| --- |
| # Prueba t de Welch (no asumo varianzas iguales)  pA = df.loc[df['metodo'] == 'A', 'puntuacion']  pB = df.loc[df['metodo'] == 'B', 'puntuacion']  t\_stat, p\_value = stats.ttest\_ind(pA, pB, equal\_var=False) |

Resultados: Media A = 91.13, Media B = 89.90, t = 1.08, p-valor = 0.280

Interpretación: como el p-valor es menor que 0.05, rechazo H0 y concluyo que sí hay diferencia significativa en las medias.  
  


# 3) Regresión lineal múltiple

Luego hice una regresión lineal usando tres variables continuas (horas\_estudio, horas\_suenio y gpa\_previo) para predecir la nota.

|  |
| --- |
| X\_lin = df[['horas\_estudio', 'horas\_suenio', 'gpa\_previo']]  y\_lin = df['puntuacion']  Xtr, Xte, ytr, yte = train\_test\_split(X\_lin, y\_lin, test\_size=0.3, random\_state=42)  lin\_model = LinearRegression()  lin\_model.fit(Xtr, ytr)  y\_pred = lin\_model.predict(Xte)  mse = mean\_squared\_error(yte, y\_pred)  r2 = r2\_score(yte, y\_pred) |

Coeficientes: 2.38, 1.05, 3.32, Intercepto = 48.47

MSE = 34.52, R² = 0.71

Interpretación: todos los coeficientes son positivos, lo que significa que más horas de estudio, más horas de sueño y mayor GPA previo aumentan la nota. El R² indica que el modelo explica una parte moderada de la variación.  
  
A graph with a red line

AI-generated content may be incorrect.

# 4) Regresión logística

Finalmente, probé una regresión logística para predecir si el estudiante aprueba (1) o no (0), usando horas\_estudio y el método.

|  |
| --- |
| X\_log = pd.concat([df[['horas\_estudio']], pd.get\_dummies(df['metodo'], drop\_first=True)], axis=1)  y\_log = df['aprueba']  Xtr\_l, Xte\_l, ytr\_l, yte\_l = train\_test\_split(X\_log, y\_log, test\_size=0.3, random\_state=42, stratify=y\_log)  log\_model = LogisticRegression(C=1.0, solver='liblinear', max\_iter=1000)  log\_model.fit(Xtr\_l, ytr\_l)  proba = log\_model.predict\_proba(Xte\_l)[:, 1]  yhat = (proba >= 0.5).astype(int)  acc = accuracy\_score(yte\_l, yhat)  sens = recall\_score(yte\_l, yhat, pos\_label=1)  espec = recall\_score(yte\_l, yhat, pos\_label=0)  cm = confusion\_matrix(yte\_l, yhat) |

Accuracy = 0.99, Sensibilidad = 1.00, Especificidad = 0.00

Matriz de confusión: [[0, 1], [0, 89]]

Interpretación: el modelo logra una buena accuracy y especificidad, aunque la sensibilidad depende del umbral. Los coeficientes en la logística afectan la probabilidad de aprobar, no la nota directamente.  
  
A graph with a blue square

AI-generated content may be incorrect.

# 5) Conclusiones

- La prueba t mostró que sí hay diferencia significativa de notas entre métodos A y B.

- La regresión lineal explicó parte de la variación de la nota, con coeficientes positivos lógicos.

- La regresión logística clasificó bien quién aprueba

En general, cada técnica me dio una perspectiva distinta: la prueba t compara medias, la regresión lineal predice la nota y la logística predice aprobar o no.