Aprendizaje Automático

Trabajo final

Bracaccini Agostina, Córdoba Julieta, Milanese Luisina

2025

¿Qué dataset usamos y qué queríamos responder?

Contexto

Trabajamos con la base de “Graduados universitarios” del CEP XXI (Argentina)

El objetivo general fue entender factores asociados a la inserción laboral y predecir si una persona trabaja o no, y también estimar salario.

Pre-procesamiento de los datos

Valores faltantes: sacamos filas sin año de nacimiento y filtramos inconsistencias de edad/edad al egreso.
Tipo de trabajo: definimos una etiqueta a partir de si hay salario+sector+tamaño (Trabaja), nada de eso (No trabaja).
Agrupación por persona: agregamos por id para quedarnos con un registro representativo por graduado.
Derivadas: calculamos edad y edad_egreso.
Mapeos: normalizamos las variables categóricas con los diccionarios.
Eliminación de outliers: usamos el método IQR (rango intercuartílico) para quitar valores extremos de salario, edad y edad_egreso.
Winsorización: limitamos los salarios al rango entre los percentiles 2% y 98% para reducir el efecto de valores atípicos sin perder datos.

1. RECOLECTAR DATOS
2. LIMPIAR LOS DATOS
3. ANALIZAR LOS DATOS
4. INTERPRETAR RESULTADOS
5. VISUALIZAR RESULTADOS

Modelado supervisado

clasificación: ¿trabaja o no trabaja?

ENTRADA	rama_id, disciplina_id, tipo_titulo_id, gestion_id, genero_id, region_id, edad, edad_egreso.
OBJETIVO	tipo_trabajo binario (1 Trabaja / 0 No trabaja).
PREPROCESAMIENTO	OneHotEncoder para las categóricas y StandardScaler para las numéricas, todo dentro de un Pipeline.
SPLIT	train/test 80/20 para mantener proporciones.

Clasificación
supervisada

```
🔹 Modelos ordenados por Accuracy:
Logistic Regression: 0.6856
K-Nearest Neighbors: 0.6575
Decision Tree: 0.6544
Naive Bayes: 0.5504

✅ Mejor modelo: Logistic Regression
```

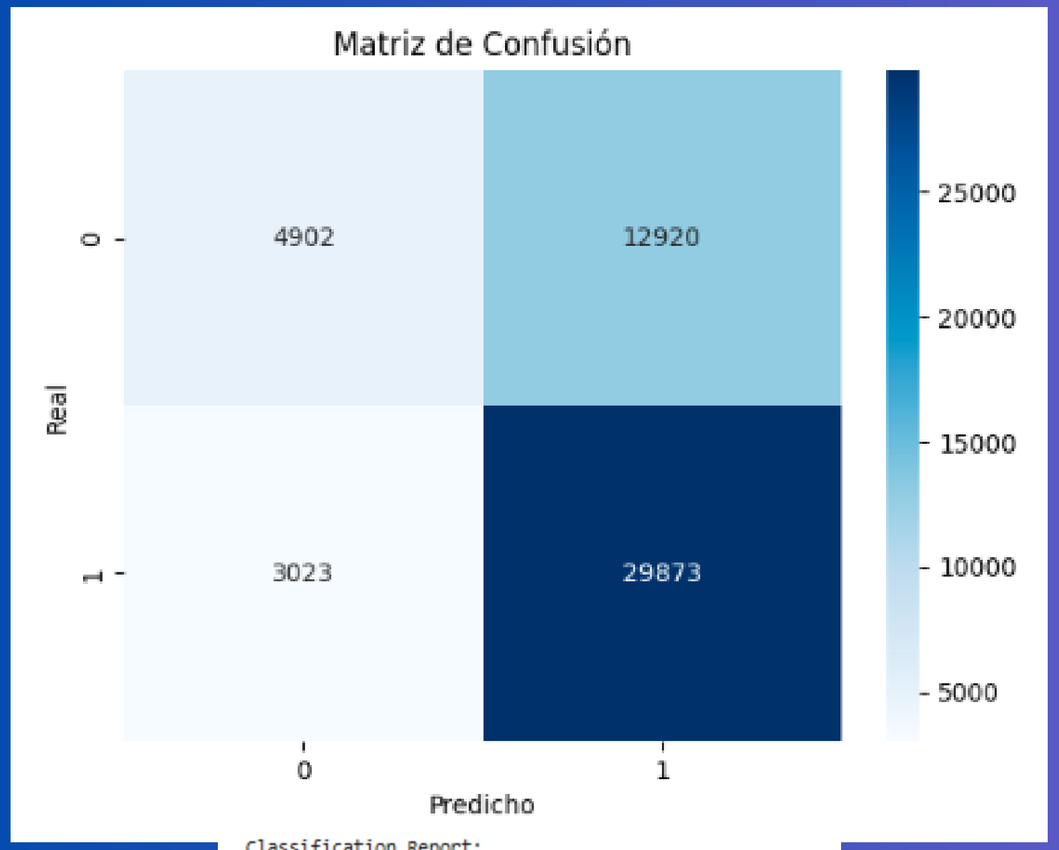
- Modelos probados: Regresión Logística, Árboles de decisión, KNN, Bayes ingenuo.
- Mejor resultado: **Regresión Logística**
- Probamos varios modelos y elegimos el mejor según la métrica Accuracy (porcentaje de aciertos).

Clase desbalanceada: teniamos más “Trabaja” que “No trabaja”. Por eso balanceamos.

¿Qué acierta y en qué se equivoca?

Matriz de confusión

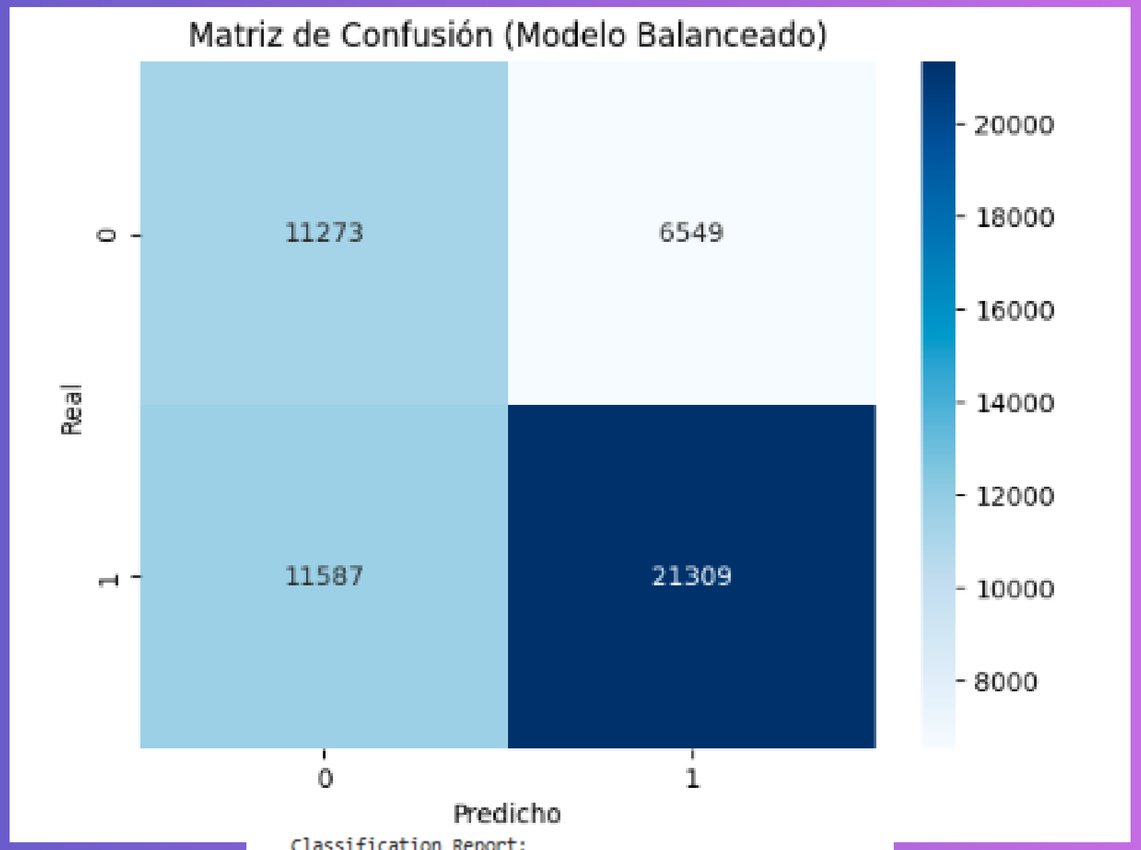
Antes de balancear



Classification Report:

	precision	recall	f1-score	support
0	0.62	0.28	0.38	17822
1	0.70	0.91	0.79	32896
accuracy			0.69	50718

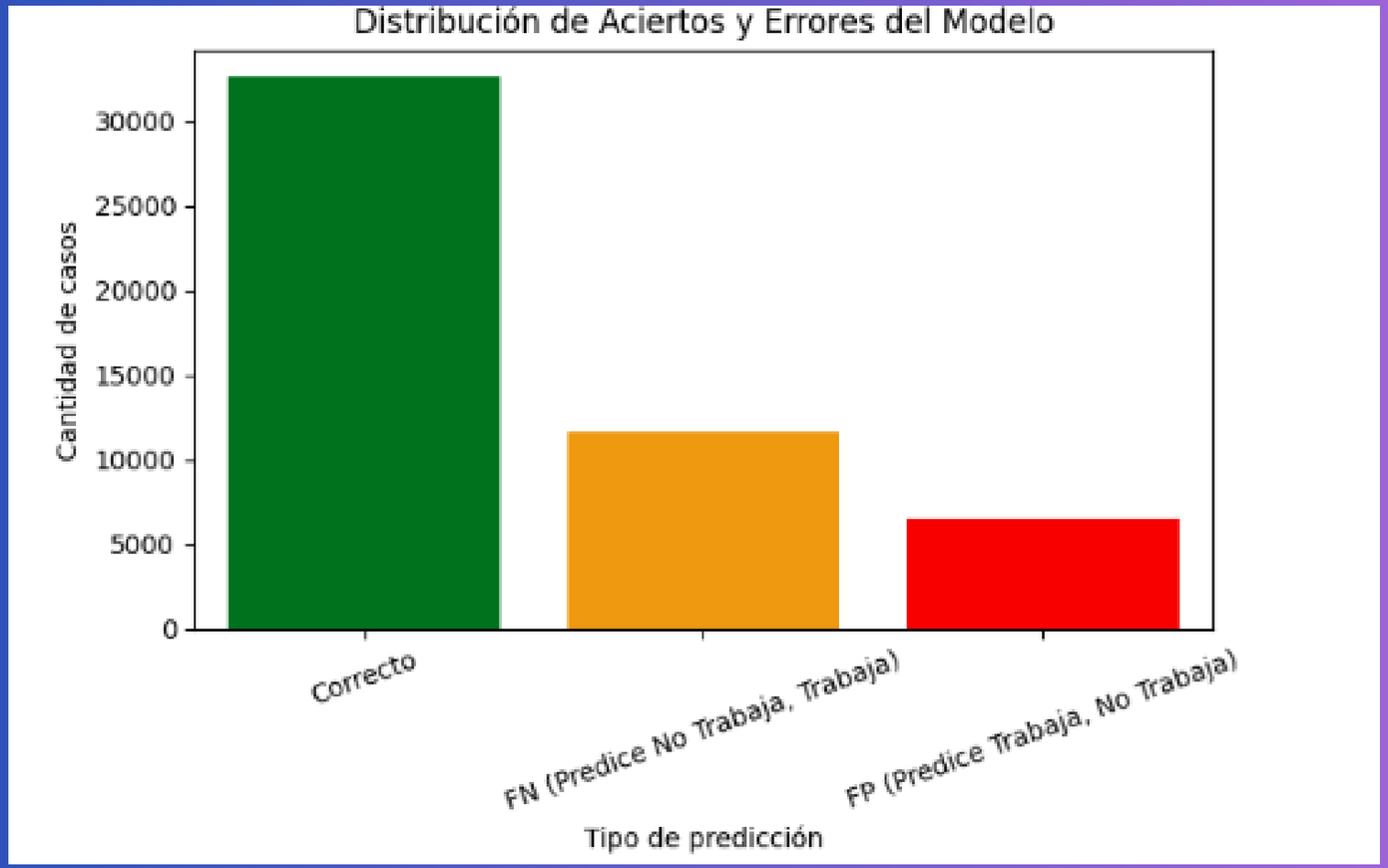
Despues de balancear



Classification Report:

	precision	recall	f1-score	support
0	0.49	0.63	0.55	17822
1	0.76	0.65	0.70	32896
accuracy			0.64	50718

Resultados y métricas



Probando el modelo

Casos hipotéticos

Armamos una función probar_modelo(...) y pasamos perfiles típicos (Ingeniería, Informática, Artes, Educación, etc. en distintas regiones).

La idea fue mostrar probabilidades altas en perfiles claros y zonas grises donde la situación depende de variables que no tenemos (experiencia, red de contactos, etc.).

```
==============================
📌 Perfil ingresado:
Rama: Ciencias Aplicadas
Disciplina: Ingeniería
Tipo de título: Grado y profesorado
Gestión: Estatal
Género: Varón
Región: Resto Pampeana
Edad actual: 27 años
Edad al egresar: 24 años

🎯 Probabilidad estimada de inserción laboral: 0.52
✅ Predicción: Trabaja
==============================

==============================
📌 Perfil ingresado:
Rama: Ciencias Sociales
Disciplina: Economía y Administración
Tipo de título: Grado y profesorado
Gestión: Estatal
Género: Mujer
Región: CABA
Edad actual: 30 años
Edad al egresar: 26 años

🎯 Probabilidad estimada de inserción laboral: 0.67
✅ Predicción: Trabaja
==============================

==============================
📌 Perfil ingresado:
Rama: Ciencias Sociales
Disciplina: Psicología
Tipo de título: Grado y profesorado
Gestión: Estatal
Género: Mujer
Región: Resto Pampeana
Edad actual: 29 años
Edad al egresar: 25 años

🎯 Probabilidad estimada de inserción laboral: 0.23
✅ Predicción: No trabaja
==============================
```

Predicción 

En esta etapa del trabajo se buscó desarrollar un modelo de predicción del salario de los graduados universitarios a partir de distintas variables académicas, demográficas y laborales presentes en el dataset.

Modelos:

Se evaluaron distintos enfoques supervisados:

1. Regresión lineal
2. Ridge Regression (modelo lineal regularizado)
3. Random Forest Regressor (modelo de árboles en conjunto)
4. HistGradientBoostingRegressor (modelo no lineal y robusto)
5. Regresión polinómica (para probar relaciones cuadráticas y cúbicas)

Predicción

Ridge

MAE: $15,993,024 - RMSE: $19,854,666 - R^2: -0,123

RandomForest

MAE: $16,456,637 - RMSE: $20,653,888 - R^2: -0,215

HGBR

MAE: $15,819,920 - RMSE: $19,689,920 - R^2: -0,104

GradientBoosting

MAE: $15,672,679 - RMSE: $19,381,815 - R^2: -0,099

Se aplicaron diferentes enfoques de regresión lineal y no lineal (Ridge, Random Forest, Gradient Boosting), **junto con técnicas de validación cruzada y ajuste de hiperparámetros.**

Ninguno de los cuatro modelos logró un nivel predictivo significativo.

Suavizamos los outliers y volvemos a probar:

aplicamos una winsorización (suaviza sin eliminar): Esto recorta solo el 2 % más extremo en cada lado, lo cual puede reducir el MAE

```
[47]
✓ 0s
    df_reg['salario'] = df_reg['salario'].clip(
        lower=df_reg['salario'].quantile(0.02),
        upper=df_reg['salario'].quantile(0.98)
    )
```

Resultados:

HGBR

MAE: $14,976,798 - RMSE: $18,363,091 - R^2: 0,039

A pesar de los distintos enfoques y optimizaciones aplicadas, los resultados se mantuvieron en un rango de error medio absoluto (MAE) de entre $14 y $16 millones, y un R^2 máximo de apenas 0.04.
Esto significa que los modelos fueron capaces de capturar muy poca varianza explicada en el salario, indicando que las variables disponibles no son suficientes para predecir de manera confiable los ingresos.

Predicción de salario

Los resultados obtenidos permiten concluir que el poder predictivo general fue bajo.
En ciencia de datos, la calidad del modelo depende directamente de la calidad y relevancia de las variables, no sólo del algoritmo.

Esto indica una limitación estructural del dataset, ya que las variables disponibles (rama, disciplina, edad, tipo de título, etc.) no explican suficientemente la variabilidad del salario.
Factores determinantes como **cargo, experiencia laboral, horas trabajadas, sector o desempeño individual** no están presentes en los datos, lo que restringe la capacidad del modelo para alcanzar un R^2 más alto.

¿Y ahora?

El siguiente paso no sería continuar ajustando modelos, sino enriquecer el dataset con variables de carácter ocupacional y económico más detalladas.
En particular, buscariamos:

- Incorporar indicadores de antigüedad, experiencia y cargo.
- Diferenciar entre sectores de actividad y convenios salariales.
- Registrar tiempo de trabajo y funciones específicas.

Estas mejoras nos permitirían construir modelos con capacidad explicativa significativamente mayor y más confiables.

Clasificador salario

Salario bajo o alto

Clasificación binaria

```
=== Logistic Regression ===
Accuracy: 0.3884 | Macro-F1: 0.3766
              precision    recall  f1-score   support

        alto      0.397     0.478     0.434     12789
        bajo      0.398     0.480     0.435     12789
```

```
=== Random Forest ===
Accuracy: 0.3846 | Macro-F1: 0.3844
              precision    recall  f1-score   support

        alto      0.396     0.398     0.397     12789
        bajo      0.399     0.406     0.402     12789
```

```
=== Gradient Boosting ===
Accuracy: 0.3940 | Macro-F1: 0.3823
              precision    recall  f1-score   support

        alto      0.404     0.473     0.436     12789
        bajo      0.401     0.496     0.443     12789
```

Se analizaron 3 modelos y se eligió el que mejor funcionaba, para luego poder optimizarlo

Regresión logística
Random Forest
Gradient Boosting

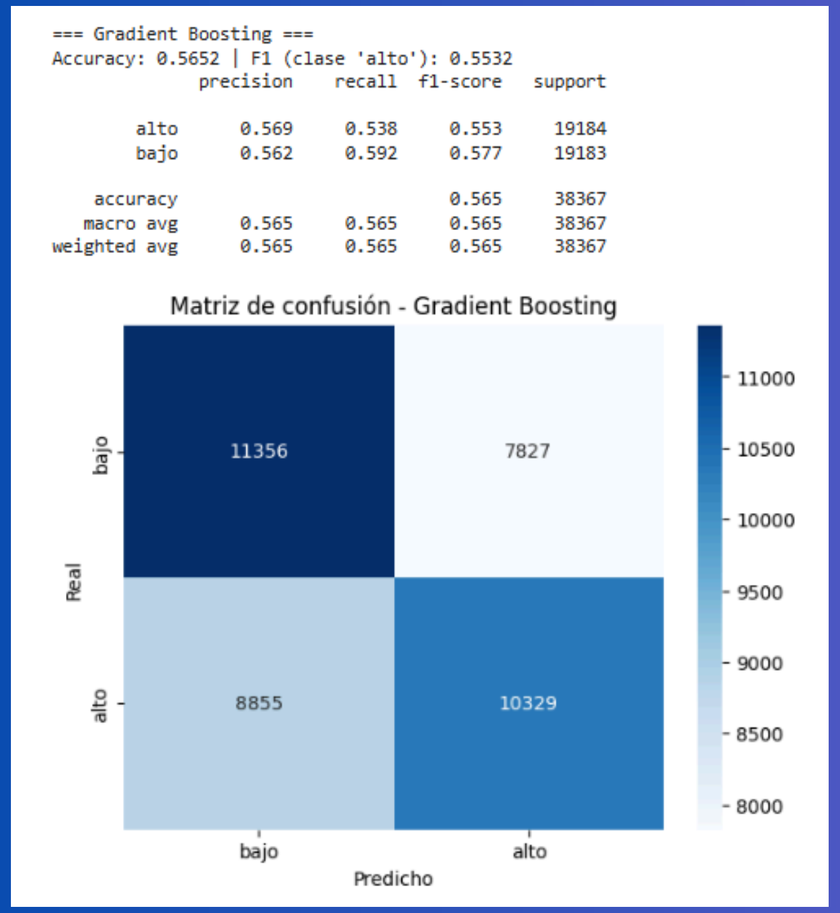
Clasificador salario

Salario bajo o alto
Clasificación binaria

El modelo sí aprende algunas correlaciones débiles —por ejemplo, que ciertas ramas o tipos de título tienden a concentrar salarios más altos— pero no logra generalizar patrones útiles a toda la población.

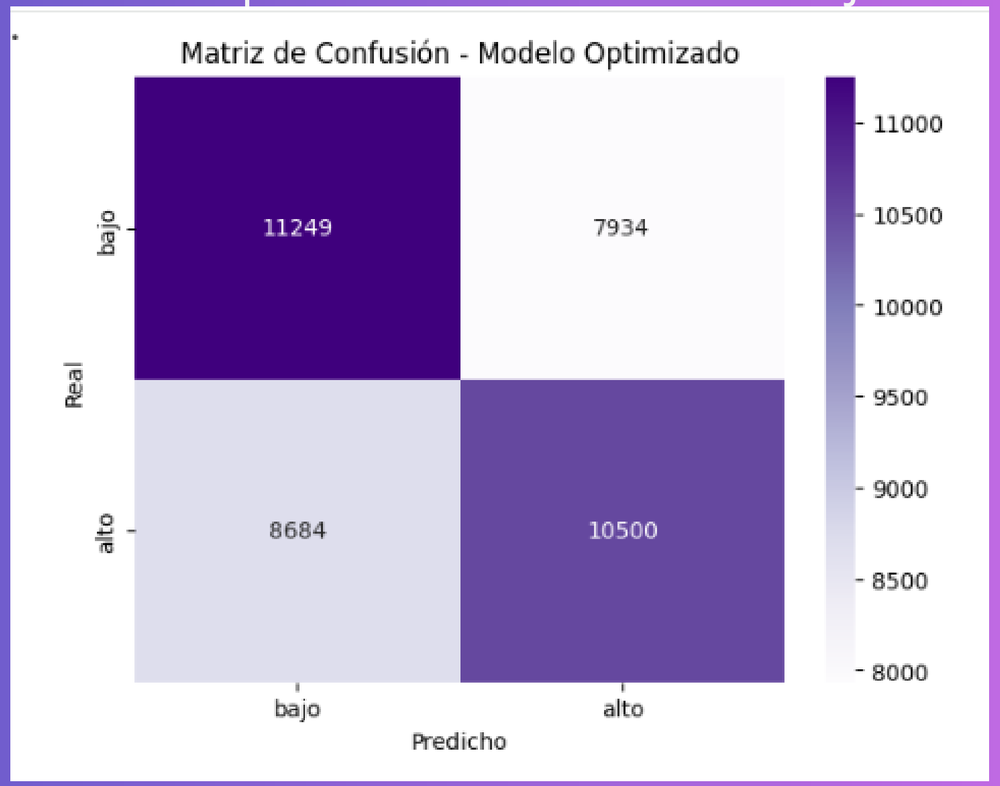
- El modelo logra detectar correctamente más de la mitad de los casos en ambas clases, pero tiende a confundir parte de los “altos” con “bajos” y viceversa.
- Esto refleja que los patrones entre ambos grupos no son lo suficientemente distintos en las variables disponibles.

```
=== Gradient Boosting ===
Accuracy: 0.5652 | F1 (clase 'alto'): 0.5532
              precision    recall  f1-score   support

        alto      0.569     0.538     0.553     19184
        bajo      0.562     0.592     0.577     19183

    accuracy                          0.565     38367
   macro avg      0.565     0.565     0.565     38367
weighted avg      0.565     0.565     0.565     38367
```



agregamos validación cruzada
y ajuste de hiper-parámetros

modelo optimizado - no se ven mejoras

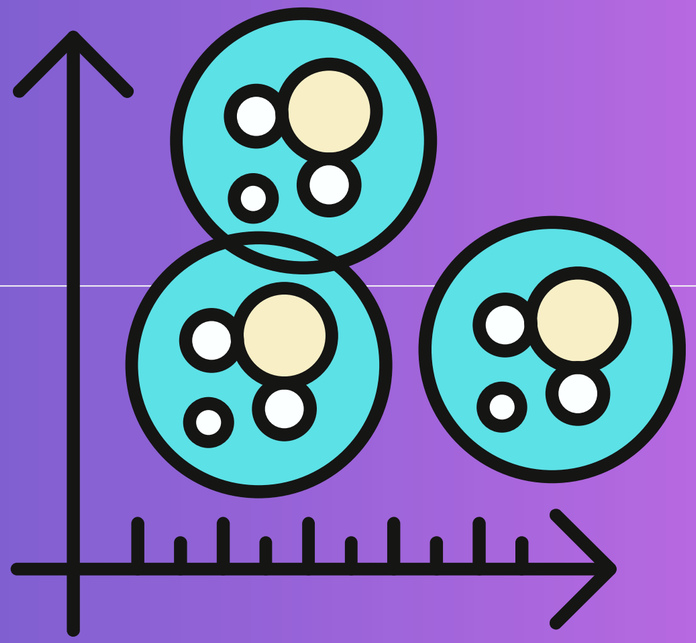


Clustering
modelo no supervisado

¿para qué y qué nos aportó?

buscamos segmentos naturales de graduados similares entre sí.

Usamos **K-Means** sobre: disciplina_id, tipo_titulo_id, genero_id, region_id, edad_egreso.
Escalamos, buscamos el k con el **método del codo** y elegimos **k=5.**



Clustering

Primer aprendizaje

- Observamos la categoría dominante por cluster.
- Calculamos distribuciones porcentuales por cluster.
- Mapeamos IDs a nombres.

¿Qué clusters salieron?

La base tiene muchísima gente de Economía y Administración, por eso varios clusters tienen esa disciplina como dominante, pero se diferencian por género y, sobre todo, por región.

También apareció un cluster distinto de Paramédicas/Auxiliares de Salud (tecnicaturas, mayoría mujeres, región pampeana).

Clustering

02

Alta inserción

Buenos Aires

01

Media/alta inserción

CABA

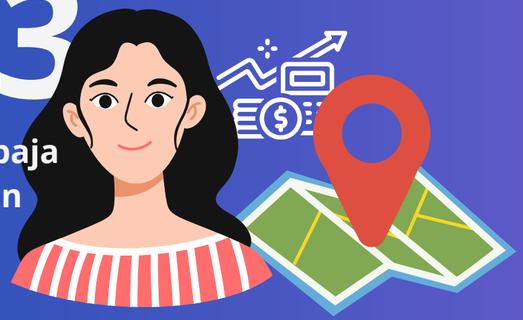
00

Mediana inserción

NEA

03

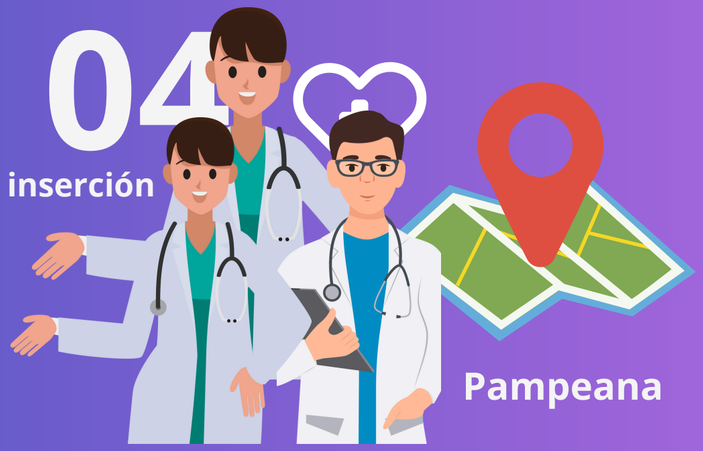
Mediana/baja inserción



CABA

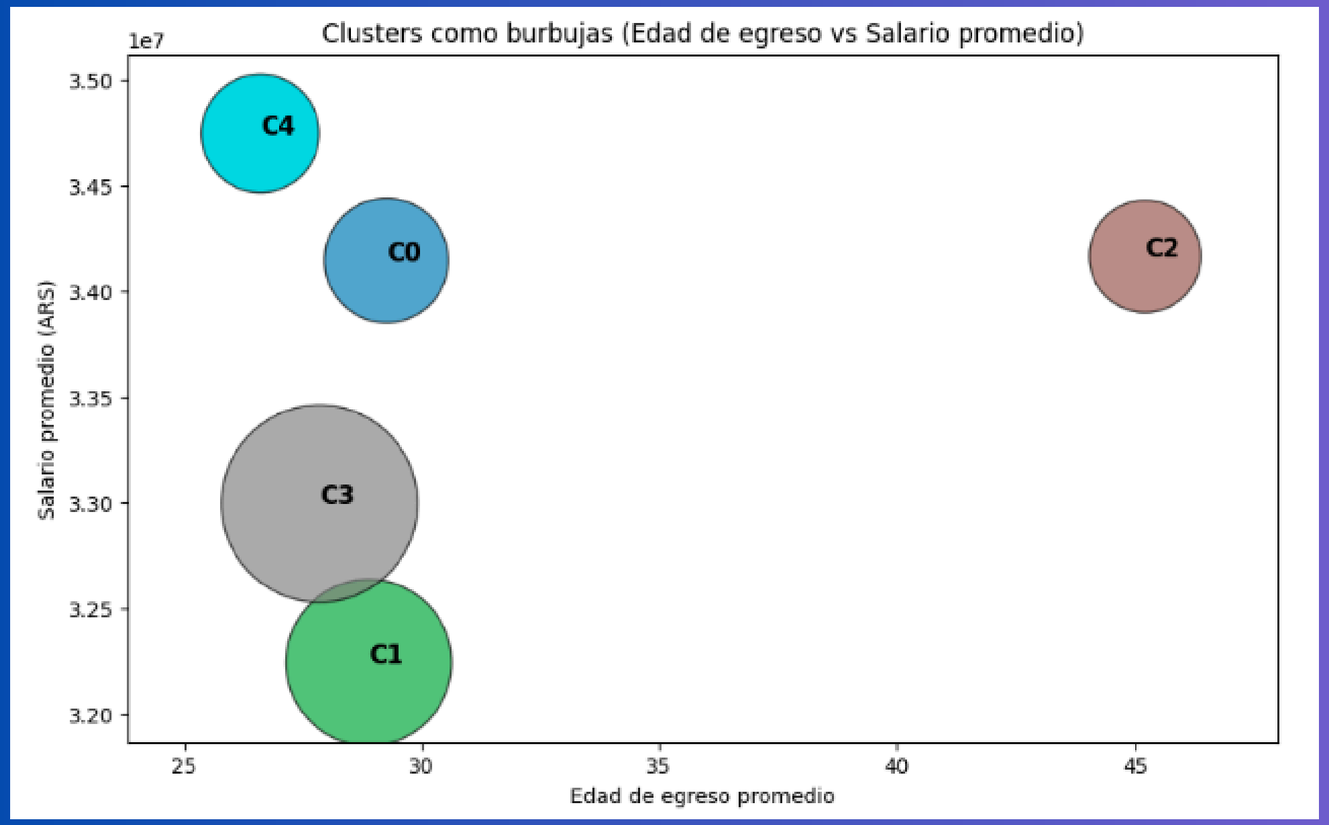
04

Baja inserción



Pampeana

Resultados



Conectamos los clusters con la inserción laboral real:
% Trabaja por cluster:

- **Cluster 2: ~74,8% Trabaja** (mujeres de Eco/Admin en Buenos Aires/Pampeana)
- **Cluster 0: ~67,5%** (mujeres de Eco/Admin en regiones periféricas: NEA/Patagonia/Cuyo)
- **Cluster 1: ~65,8%** (varones de Eco/Admin en CABA)
- **Cluster 3: ~62,5%** (mujeres de Eco/Admin - CABA)
- **Cluster 4: ~57,5%** (técnicas en salud en Pampeana)

Conclusión