

Prediciendo y Comprendiendo

Machine Learning en la Exploración de Datos
Socioeconómicos

Por Ivan Vendrell



Definición del problema

El análisis de datos socioeconómicos y demográficos juega un papel crucial en la comprensión de las dinámicas de una población y en la identificación de factores que pueden influir en diversos aspectos de la vida de los individuos.

En este trabajo, se aborda el problema tanto de *predecir un nivel de éxito basado en diversas características socioeconómicas y demográficas* y *modelar la relación entre el estado de empleo y el nivel educativo de una población*.

El objetivo principal de este análisis es *construir un modelo predictivo capaz de clasificar individuos en diferentes niveles de éxito e identificar y cuantificar la influencia de diferentes factores en el porcentaje de educación general de una población*, utilizando técnicas de Machine Learning para extraer información valiosa del dataset disponible.




Descripción del Dataset

El dataset utilizado en este trabajo es una fusión de 3 datasets procedentes de la página de estadística Europea <https://ec.europa.eu/eurostat/web/main/home> y contiene información sobre diversas características socioeconómicas y demográficas.

Los datasets originales constan de:

Data set original	Filas	Columnas
estat_edat_lfse_03_en.csv	131064	23
estat_edat_lfs_9912_encsv	894200	25
estat_edat_lfs_9904_en.csv	1144732	25



Aunque los datos parecen abrumadores hay que tener en cuenta que muchas de las columnas eran redundantes, empiezan en el 1983 en algunos casos y que consta la información de todos los países europeos. Tras una primera limpieza (y renombramiento de columnas) nos encontramos:

Variables Numéricas: 'percentage_education_general' (porcentaje de población con educación general), 'percentage_by_employment_status' (porcentaje de población por estado de empleo), 'percentage_education_by_birth' (porcentaje de población con educación por lugar de nacimiento); estas tres corresponden al dato distintivo de cada dataset.

Variables Categóricas: 'country' (país de origen), 'sex' (sexo), 'age_group' (grupo de edad), 'education_level' (nivel educativo alcanzado), 'year' (año de la encuesta o dato).

Durante la exploración inicial, se identificaron valores faltantes en algunas columnas, alta correlación entre 'percentage_education_general' y 'percentage_by_employment_status', variedad de categorías en 'education_level' que requirieron limpieza y posible agrupación. Se tomaron decisiones de limpieza como la eliminación de filas con valores faltantes en la variable objetivo de regresión ('percentage_education_general'), la agrupación de ciertos niveles educativos en la columna 'education_level' para preparar el dataset para el análisis de Machine Learning.



Contexto

Tras la limpieza inicial y una vez fusionados los tres datasets iniciales, el dataset resultante consta de 2525570 líneas y 13 columnas .

El análisis de los factores que determinan el nivel de éxito socioeconómico y el nivel educativo general de una población es crucial para abordar problemas de desigualdad, movilidad social y desarrollo. Los modelos de Machine Learning, tanto de clasificación (para predecir el nivel de éxito) como de regresión (para predecir el porcentaje de educación general), ofrecen herramientas valiosas para obtener información accionable que puede informar políticas públicas más efectivas, identificar poblaciones vulnerables, optimizar la asignación de recursos y comprender las dinámicas socioeconómicas y educativas subyacentes. Este trabajo se alinea con el creciente interés en aplicar el aprendizaje automático para generar *insights* significativos en las ciencias sociales y la planificación del desarrollo.



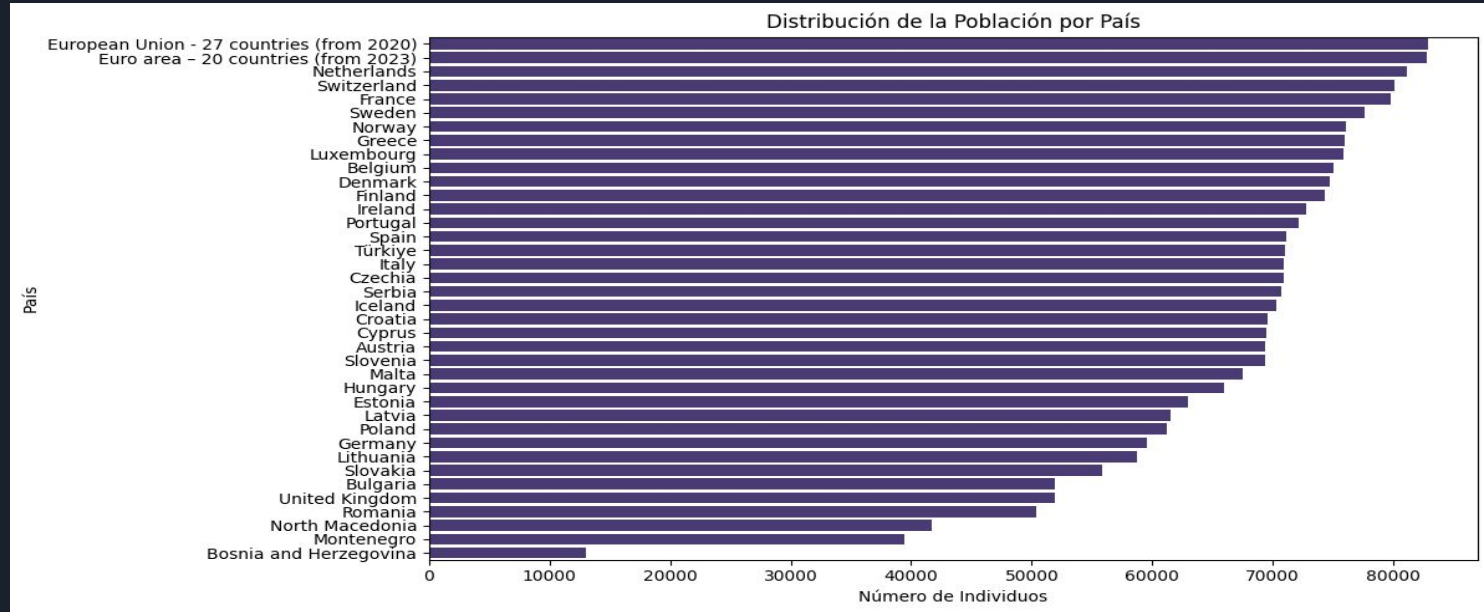
EDA - Análisis Exploratorio de Datos

Detectando los Cimientos del Modelo: La calidad de nuestros modelos depende directamente de nuestra comprensión inicial de los datos.

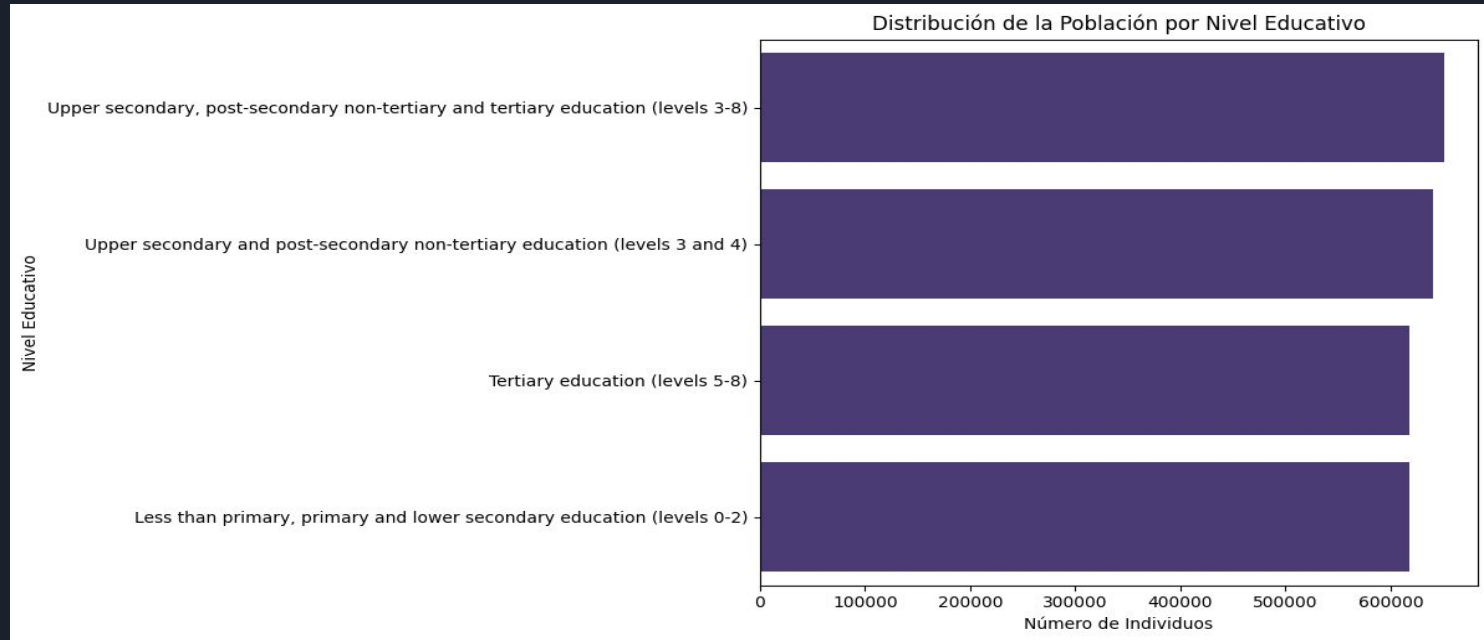
Evitando Errores Costosos: La exploración visual nos ayuda a identificar problemas como valores atípicos o desequilibrios que podrían sesgar nuestros modelos.

Maximizando el Potencial Predictivo: Al comprender las relaciones visualmente, podemos seleccionar las características más relevantes y diseñar modelos más precisos.

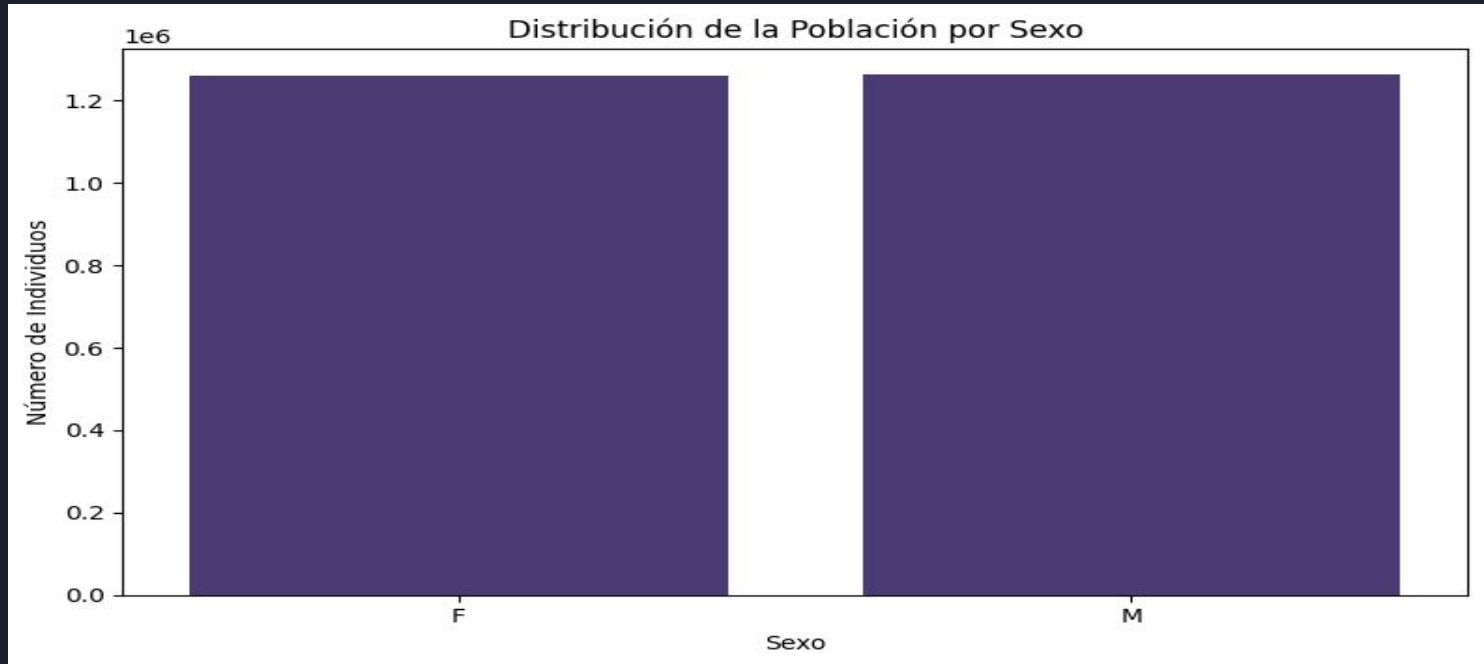
Comprender la representación de cada país es esencial para evaluar la posible influencia de la nacionalidad en las variables objetivo (nivel de éxito, porcentaje de educación general) y para considerar si la muestra es representativa de una población más amplia.



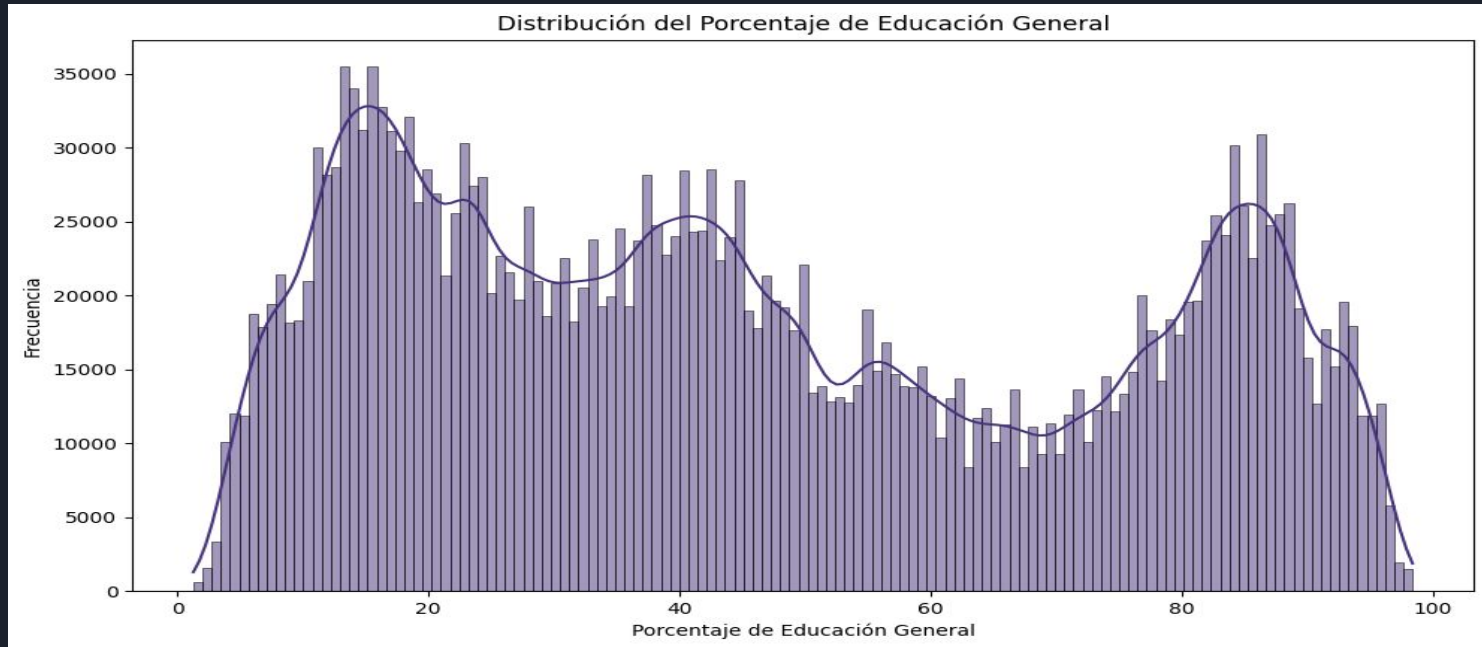
Dado que el nivel educativo es una variable clave tanto para la definición del "nivel de éxito" como para la predicción del "porcentaje de educación general", conocer la composición educativa de la muestra es fundamental para interpretar los resultados del modelado en relación con las diferentes categorías educativas.



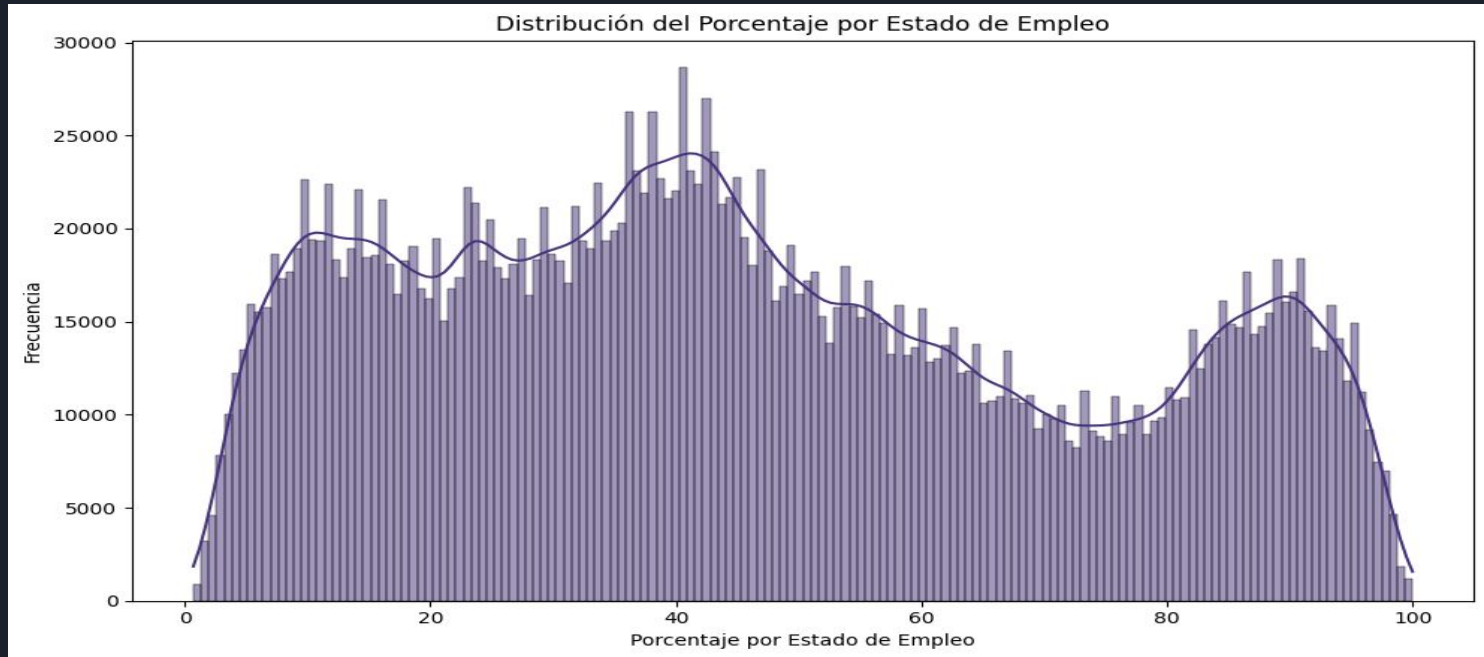
Es importante analizar si existe un equilibrio de género en el dataset, ya que el sexo podría ser un factor relevante en las dinámicas socioeconómicas y educativas que estamos investigando.



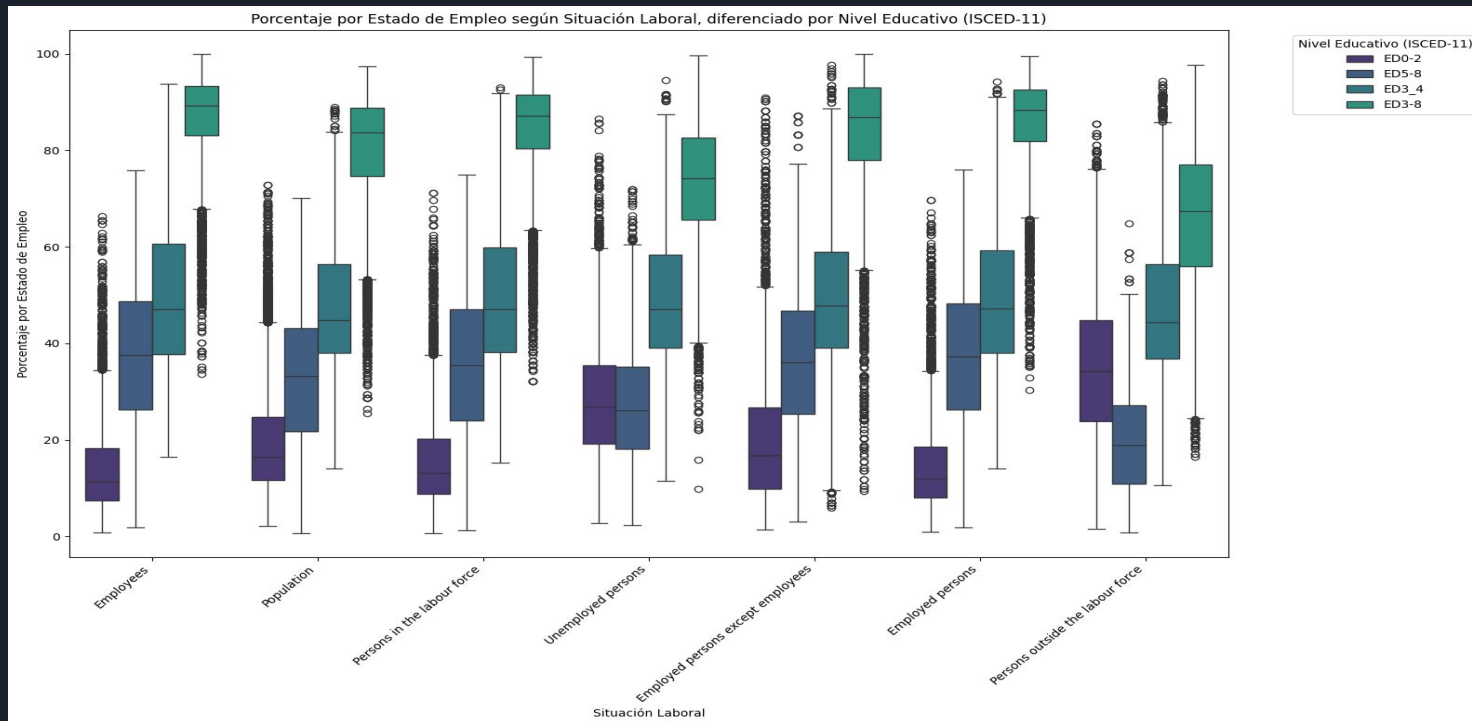
El histograma muestra una distribución con **varios picos**, indicando grupos distintos en el 'Porcentaje de Educación General'. Esto sugiere que la población no tiene un nivel educativo general único predominante. Los picos podrían **reflejar diferentes generaciones, sistemas educativos o factores socioeconómicos**. Esta distribución compleja implica que la relación con otras variables podría ser no lineal, haciendo que modelos como **Random Forest** sean potencialmente más adecuados para la predicción.



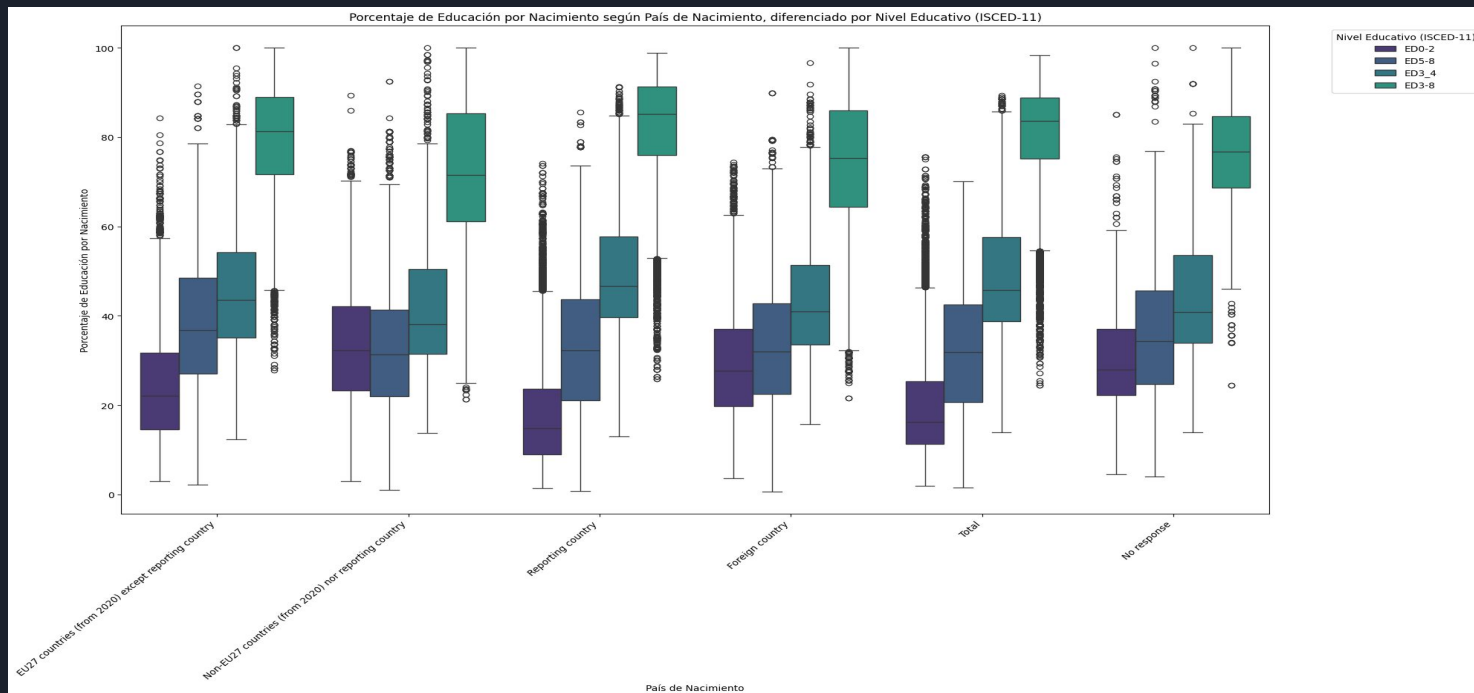
El histograma muestra múltiples picos en el 'Porcentaje por Estado de Empleo', indicando grupos distintos con diferentes niveles de empleo. Estos grupos podrían estar influenciados por edad, educación, país o economía. Dado que esta variable predice el nivel educativo, su distribución compleja sugiere una relación no uniforme en la población, lo que requiere un análisis detallado.



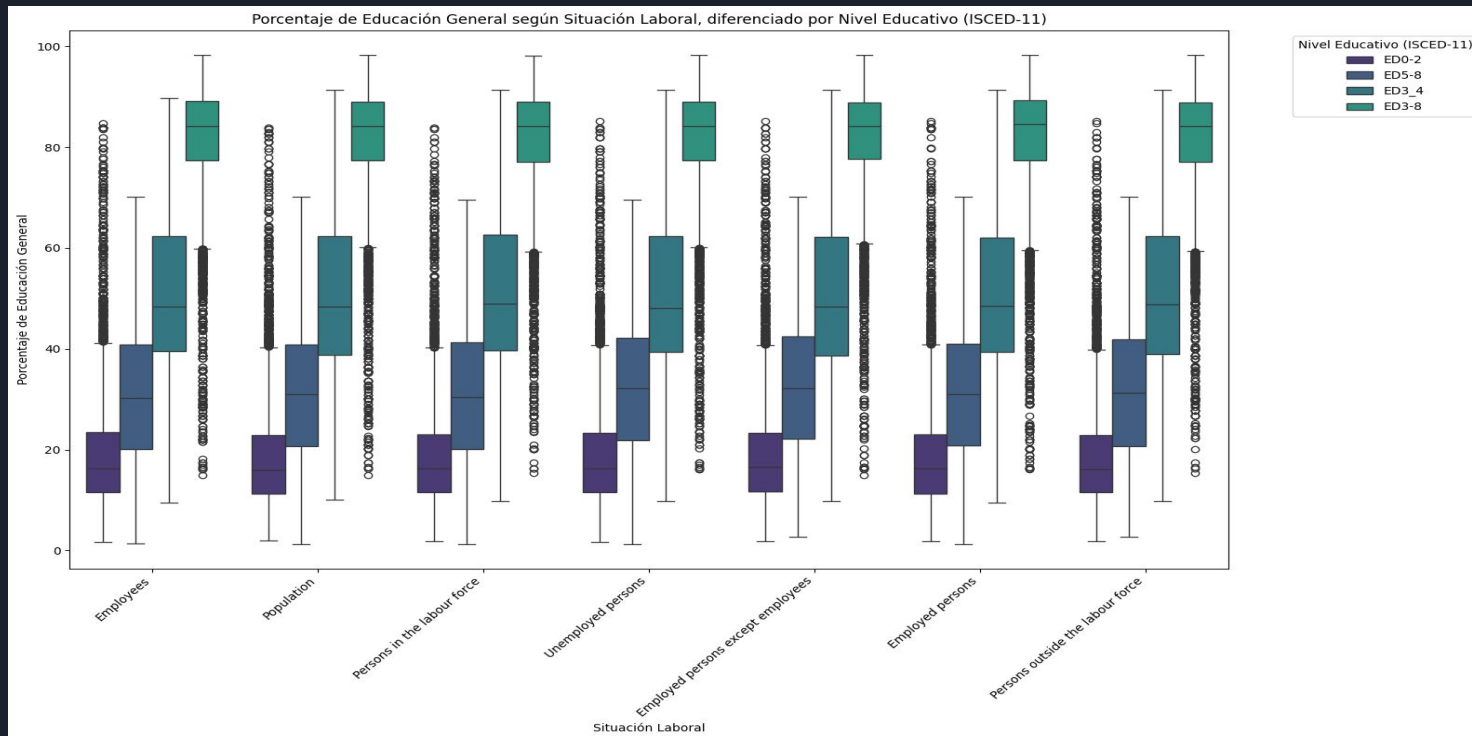
Se observan variaciones significativas entre situaciones laborales y niveles educativos. El nivel educativo influye en el porcentaje de empleo dentro de cada situación laboral. Los valores atípicos señalan casos particulares a considerar.



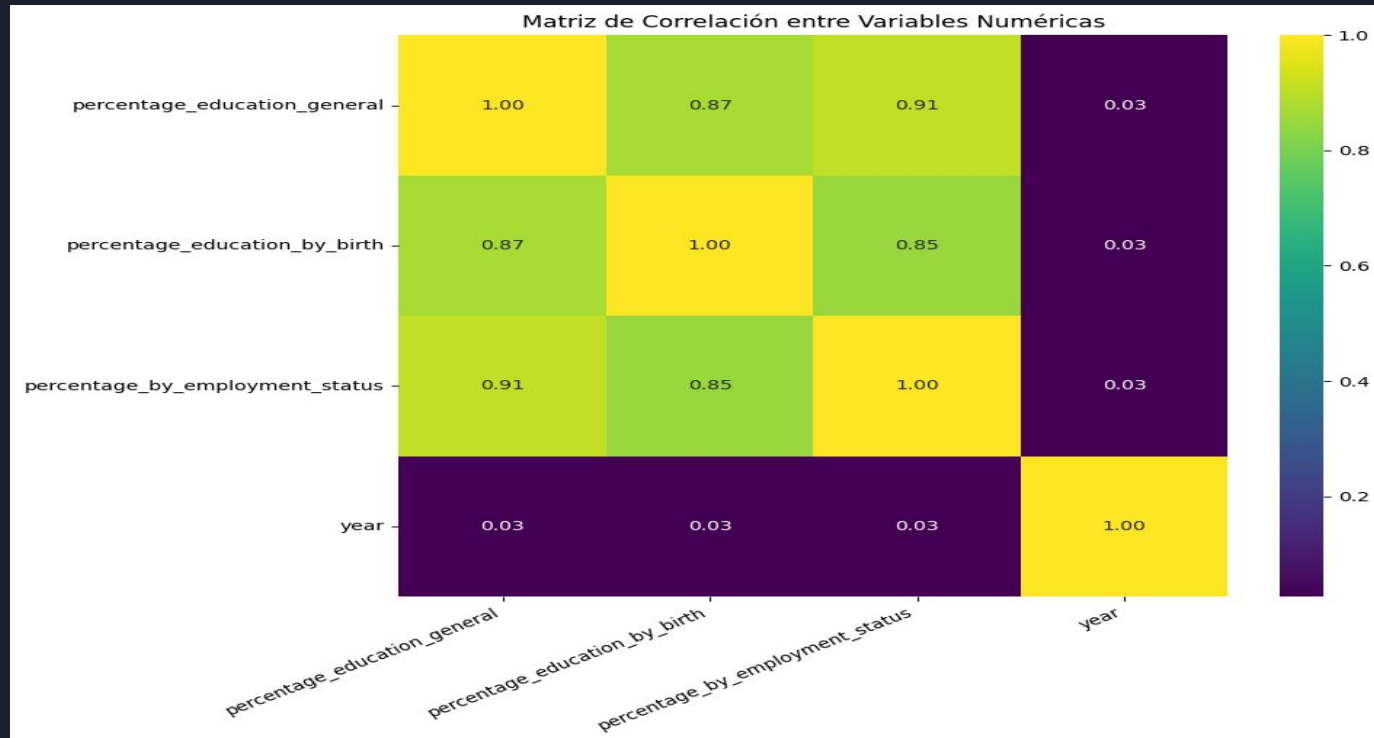
Se aprecian diferencias significativas entre países de nacimiento y la distribución de niveles educativos dentro de cada uno. Los outliers señalan casos atípicos en la relación entre origen y procedencia educativa.



Los niveles educativos más altos generalmente se asocian con mayores porcentajes de educación general, pero la distribución varía según la situación laboral. Los outliers indican casos atípicos en esta relación.



Matriz de correlación de Pearson que muestra fuertes correlaciones positivas entre el porcentaje de educación general, el porcentaje de educación por lugar de nacimiento y el porcentaje por estado de empleo. La variable 'year' muestra una correlación casi nula con las demás variables numéricas. La alta correlación entre las variables socioeconómicas justifica su uso en el modelado, pero también advierte sobre la posible multicolinealidad.





Conclusiones Generales del Análisis Exploratorio de Datos (EDA)

Distribuciones Multimodales: Hemos observado distribuciones complejas y multimodales en variables clave como el 'Porcentaje de Educación General' y el 'Porcentaje por Estado de Empleo', lo que sugiere la presencia de distintos grupos o patrones dentro de la población. Esto implica que las relaciones con otras variables podrían no ser lineales.

Fuerte Interrelación Socioeconómica: Existe una fuerte correlación positiva entre el 'Porcentaje de Educación General', el 'Porcentaje de Educación por Lugar de Nacimiento' y el 'Porcentaje por Estado de Empleo', lo que indica que estos aspectos del éxito socioeconómico están estrechamente vinculados.

Influencia de Variables Categóricas: El análisis a través de box plots reveló cómo variables categóricas como la 'Situación Laboral', el 'País de Nacimiento' y el 'Nivel Educativo (ISCED-11)' influyen significativamente en la distribución de las variables numéricas de interés.

Presencia de Outliers: Se identificaron outliers en varias distribuciones, lo que sugiere la existencia de grupos con características atípicas que podrían requerir una consideración especial en el análisis o el modelado.

Estabilidad Temporal Limitada: La variable 'year' mostró una correlación muy baja con las otras variables numéricas, sugiriendo que las tendencias principales observadas podrían ser relativamente estables durante el periodo cubierto por los datos.



Justificación de modelado

Modelos No Lineales: Dada la multimodalidad de las distribuciones y las posibles interacciones complejas entre las variables categóricas y numéricas, los modelos que puedan capturar relaciones no lineales (como los basados en árboles: Random Forest, Gradient Boosting) son candidatos prometedores para la tarea de predicción. El rendimiento del modelo Random Forest en la tarea de regresión inicial apoya esta dirección.

Consideración de la Multicolinealidad: La fuerte correlación entre las variables predictoras ('percentage_education_by_birth' y 'percentage_by_employment_status') y la variable objetivo ('percentage_education_general') justifica su uso, pero también advierte sobre la posible multicolinealidad. Esto podría influir en la estabilidad e interpretabilidad de modelos lineales como la regresión lineal, sugiriendo que modelos más robustos a la multicolinealidad o técnicas de selección de características podrían ser necesarios.

Importancia de las Variables Categóricas: La influencia significativa de las variables categóricas observada en los box plots sugiere que estas deben ser incorporadas de manera efectiva en los modelos. Técnicas de codificación de variables categóricas (one-hot encoding, label encoding) serán necesarias para su inclusión en algoritmos que solo manejan entradas numéricas.


Manejo de Outliers (Decisión para la Fase de Modelado): La decisión sobre cómo manejar los outliers (eliminación, transformación, modelos robustos) se tomará durante la fase de modelado, evaluando su impacto en el rendimiento y la generalización de los modelos.



Selección e ingeniería de características

Nuestra selección de características se basará en la evidencia del EDA:

- **Variables Numéricas Correlacionadas:** Incluiremos 'percentage_education_general', 'percentage_education_by_birth' y 'percentage_by_employment_status' como candidatas a predictoras, conscientes de la posible multicolinealidad que podría requerir técnicas de reducción de dimensionalidad o selección de características. Su fuerte correlación con la variable objetivo (ya sea directamente en regresión o potencialmente relacionada con las clases en clasificación) sugiere un alto poder predictivo.
- **Variables Categóricas Significativas:** 'labour_force', 'country_birth' e 'iscled11' serán incorporadas debido a su clara influencia en las variables numéricas, lo que indica que contienen información importante para discriminar entre los niveles de éxito o para predecir el porcentaje de educación general.
- **'year':** Inicialmente, podríamos incluir 'year' y evaluar su importancia a través de modelos o técnicas de selección de características más avanzadas, dada su baja correlación lineal directa.



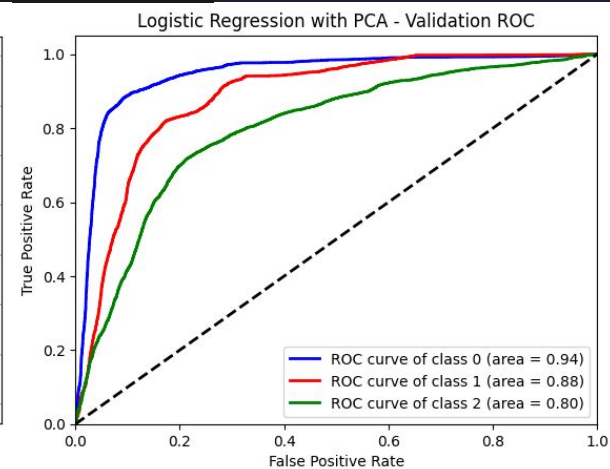
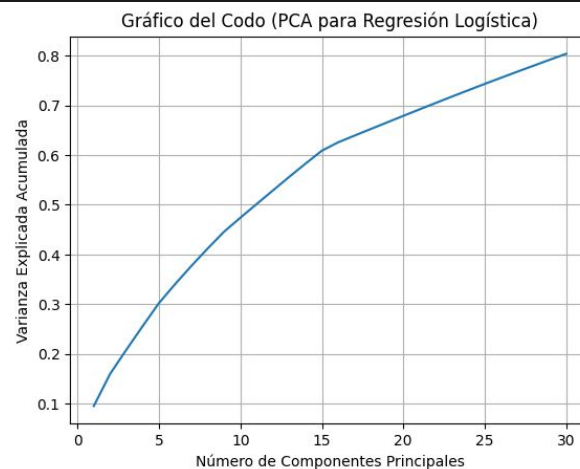
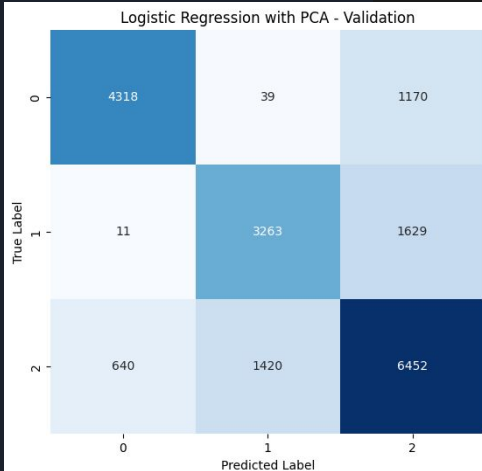
Entrenamiento y Evaluación de Modelos de Clasificación

- **Regresión logística**
- **Random Forest Classifier**
- **PCA / NO-PCA**

Best parameters for Logistic Regression with PCA: {'classifier__C': 10}

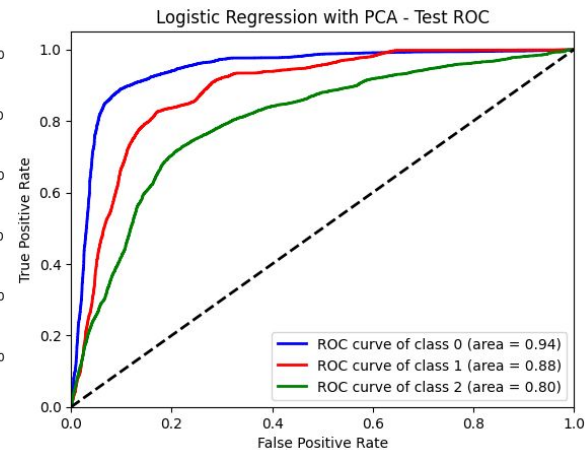
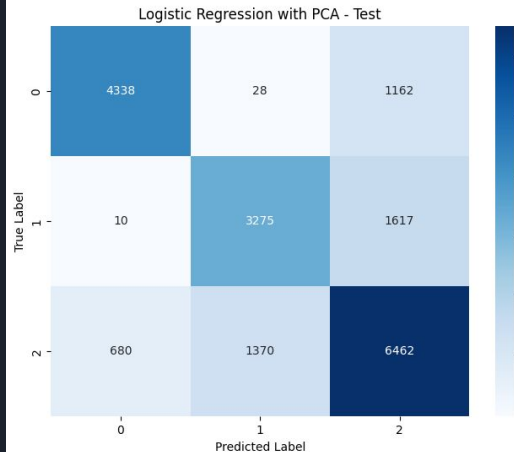
--- Logistic Regression with PCA - Validation Set Evaluation ---

	precision	recall	f1-score	support
Alto	0.87	0.78	0.82	5527
Bajo	0.69	0.67	0.68	4903
Medio	0.70	0.76	0.73	8512
accuracy			0.74	18942
macro avg	0.75	0.73	0.74	18942
weighted avg	0.75	0.74	0.74	18942



--- Logistic Regression with PCA - Test Set Evaluation ---

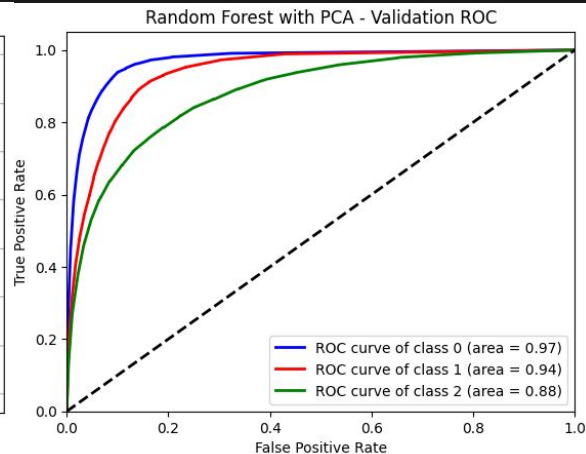
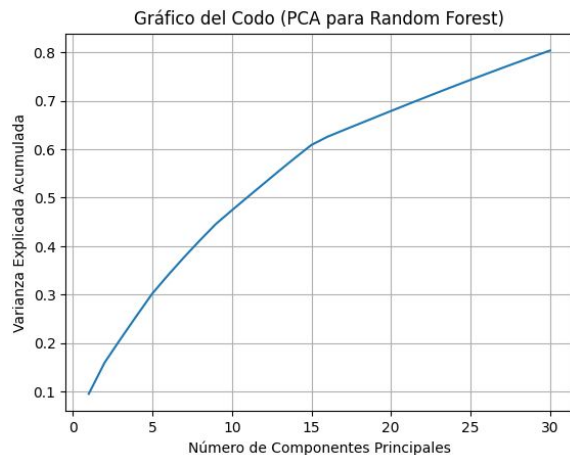
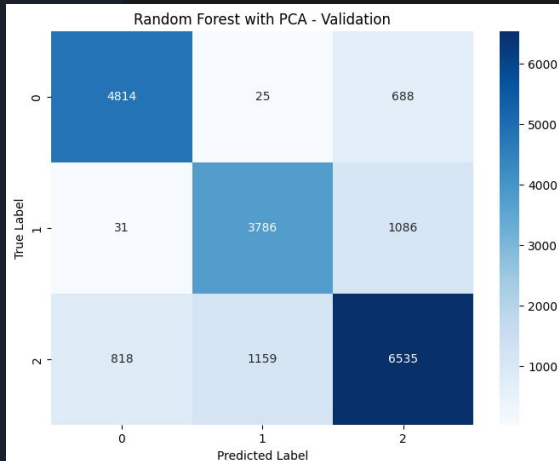
	precision	recall	f1-score	support
Alto	0.86	0.78	0.82	5528
Bajo	0.70	0.67	0.68	4902
Medio	0.70	0.76	0.73	8512
accuracy			0.74	18942
macro avg	0.75	0.74	0.74	18942
weighted avg	0.75	0.74	0.74	18942



Best parameters for Random Forest with PCA: {'classifier_max_depth': None, 'classifier_n_estimators': 20}

--- Random Forest with PCA - Validation Set Evaluation ---

	precision	recall	f1-score	support
Alto	0.85	0.87	0.86	5527
Bajo	0.76	0.77	0.77	4903
Medio	0.79	0.77	0.78	8512
accuracy			0.80	18942
macro avg	0.80	0.80	0.80	18942
weighted avg	0.80	0.80	0.80	18942



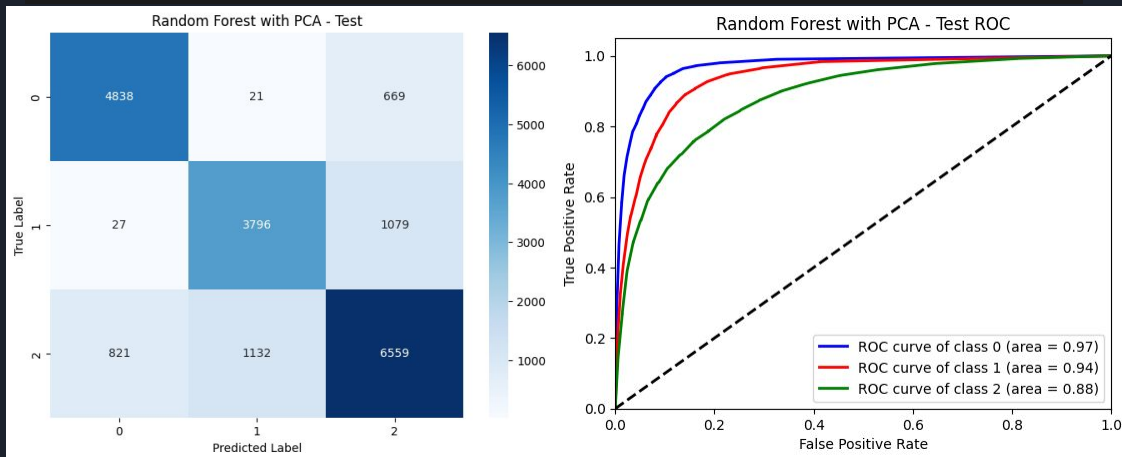
```

--- Random Forest with PCA - Test Set Evaluation ---
              precision    recall  f1-score   support

     Alto      0.85      0.88      0.86     5528
     Bajo      0.77      0.77      0.77     4902
     Medio      0.79      0.77      0.78     8512

 accuracy      0.80
macro avg      0.80      0.81      0.80     18942
weighted avg   0.80      0.80      0.80     18942

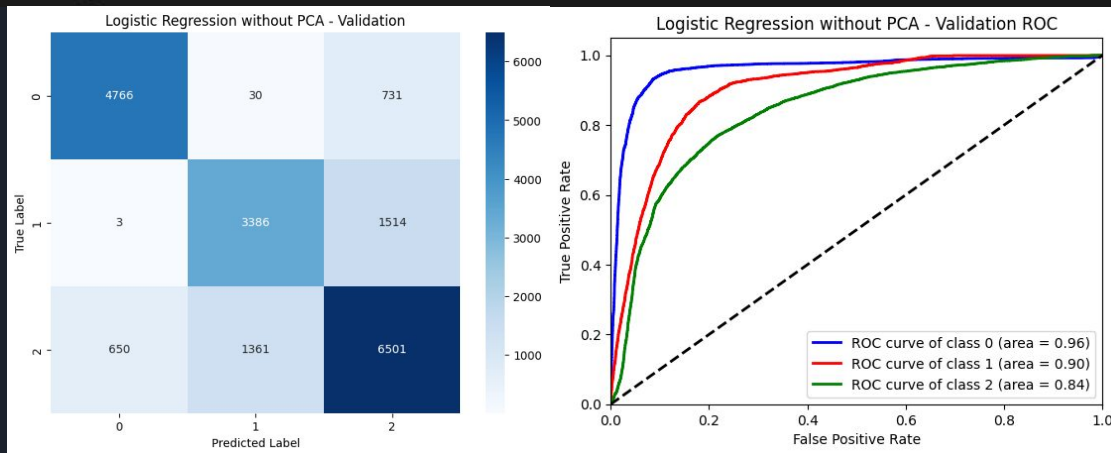
```



Best parameters for Logistic Regression without PCA: {'classifier__C': 10}

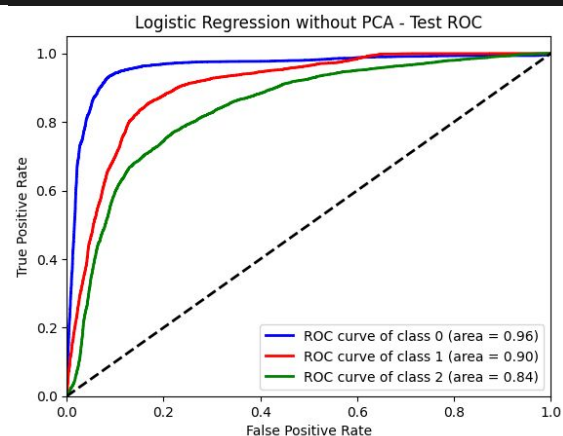
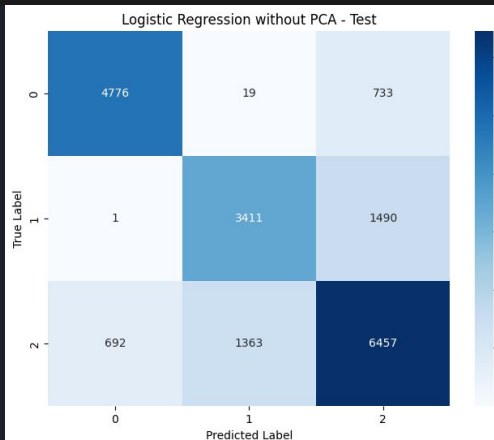
--- Logistic Regression without PCA - Validation Set Evaluation ---

	precision	recall	f1-score	support
Alto	0.88	0.86	0.87	5527
Bajo	0.71	0.69	0.70	4903
Medio	0.74	0.76	0.75	8512
accuracy			0.77	18942
macro avg	0.78	0.77	0.77	18942
weighted avg	0.77	0.77	0.77	18942



--- Logistic Regression without PCA - Test Set Evaluation ---

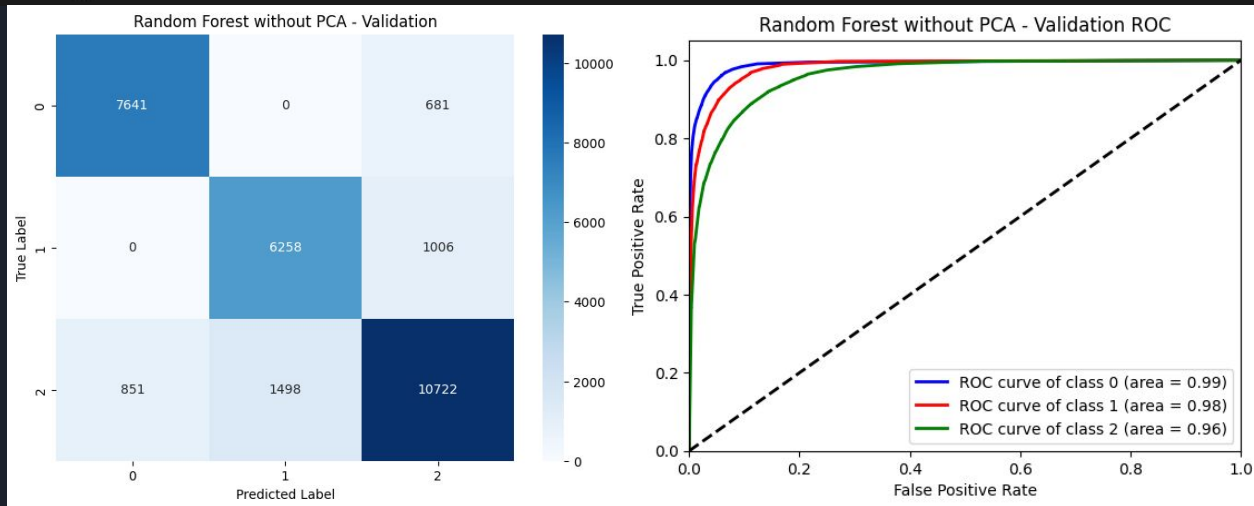
	precision	recall	f1-score	support
Alto	0.87	0.86	0.87	5528
Bajo	0.71	0.70	0.70	4902
Medio	0.74	0.76	0.75	8512
accuracy			0.77	18942
macro avg	0.78	0.77	0.77	18942
weighted avg	0.77	0.77	0.77	18942



Best parameters for Random Forest without PCA: {'classifier_max_depth': None, 'classifier_n_estimators': 20}

--- Random Forest without PCA - Validation Set Evaluation ---

	precision	recall	f1-score	support
Alto	0.90	0.92	0.91	8322
Bajo	0.81	0.86	0.83	7264
Medio	0.86	0.82	0.84	13071
accuracy			0.86	28657
macro avg	0.86	0.87	0.86	28657
weighted avg	0.86	0.86	0.86	28657

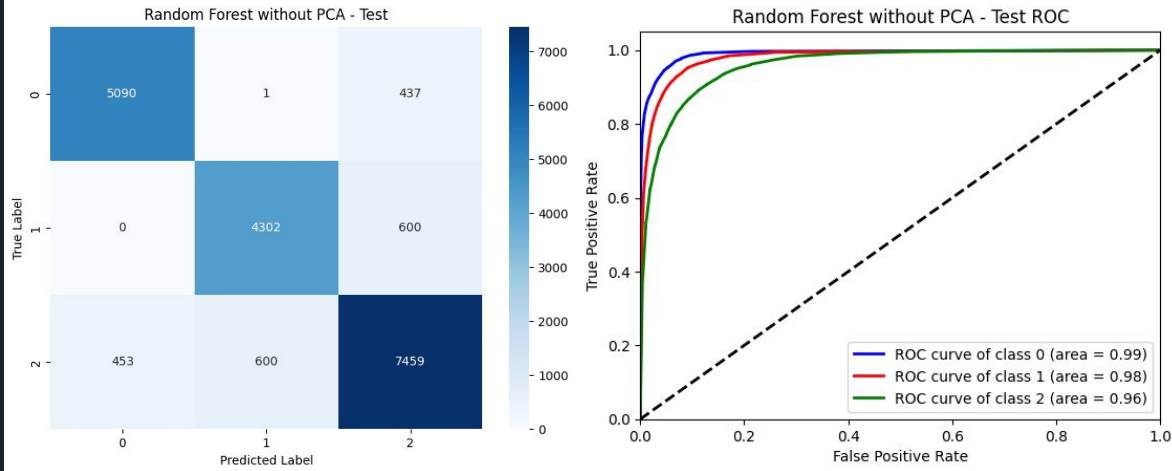


--- Random Forest without PCA - Test Set Evaluation ---

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

Alto	0.92	0.92	0.92	5528
Bajo	0.88	0.88	0.88	4902
Medio	0.88	0.88	0.88	8512

accuracy			0.89	18942
macro avg	0.89	0.89	0.89	18942
weighted avg	0.89	0.89	0.89	18942



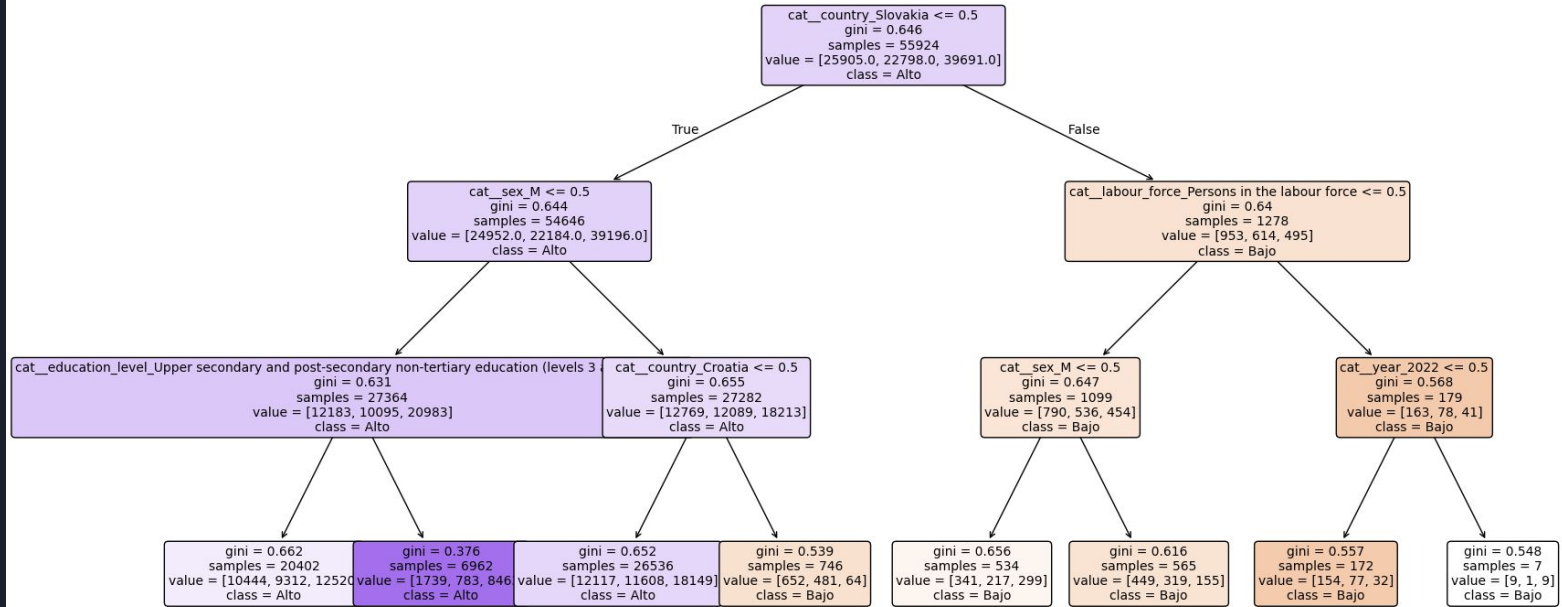
Importancia de características

```
--- Feature Importance - Random Forest without PCA ---
                                     feature  importance
47  cat__education_level_Upper secondary, post-sec...  0.194171
44  cat__education_level_Less than primary, primar...  0.109445
45  cat__education_level_Tertiary education (level...  0.079234
46  cat__education_level_Upper secondary and post-...  0.055863
40                                     cat__age_group_15-24  0.028904
..                                     ...             ...
20                                     cat__country_Lithuania  0.003304
34                                     cat__country_Sweden  0.003276
8                                     cat__country_Estonia  0.002977
2                                     cat__country_Bosnia and Herzegovina  0.002976
43                                     cat__age_group_55-64  0.001421

[76 rows x 2 columns]
```

Visualización de un árbol de decisión

Primer Árbol de Decisión en Random Forest



Análisis de la Distribución del Nivel de Éxito por Variables Categóricas Clave

```
--- Distribución del Nivel de Éxito por Sexo ---
nivel_exito      Alto      Bajo      Medio
sex
F                0.281720  0.238126  0.480154
M                0.301887  0.279441  0.418672

--- Distribución del Nivel de Éxito por Grupo de Edad ---
nivel_exito      Alto      Bajo      Medio
age_group
15-24            0.352581  0.308645  0.338774
15-64            0.246545  0.261098  0.492357
25-54            0.278811  0.244657  0.476533
55-64            0.000000  0.000000  1.000000

--- Distribución del Nivel de Éxito por Nivel Educativo ---
nivel_exito      Alto      Bajo \
education_level
Less than primary, primary and lower secondary ... 0.008514 0.666439
Tertiary education (levels 5-8)                   0.019508 0.329304
Upper secondary and post-secondary non-tertiary... 0.237735 0.060569
Upper secondary, post-secondary non-tertiary an... 0.872123 0.001969

nivel_exito      Medio
education_level
Less than primary, primary and lower secondary ... 0.325047
Tertiary education (levels 5-8)                   0.651188
Upper secondary and post-secondary non-tertiary... 0.701695
Upper secondary, post-secondary non-tertiary an... 0.125908
```




Entrenamiento y Evaluación de Modelos de Clasificación. Conclusiones.

El Random Forest sin PCA es el modelo más efectivo para clasificar el nivel de éxito. Su alta accuracy tanto en la validación como en el conjunto de prueba lo convierte en la mejor opción.

La reducción de dimensionalidad mediante PCA no mejoró el rendimiento y, de hecho, pareció perjudicar ligeramente a ambos modelos. Esto podría indicar que las características originales, en su totalidad, contienen información importante para la clasificación.

Los niveles educativos y el grupo de edad son predictores importantes del nivel de éxito, según lo identificado por el Random Forest sin PCA.

El sexo parece ser una variable con alta varianza en la primera componente principal cuando se aplica PCA, aunque su relación directa con la predicción del nivel de éxito es menos clara en los modelos sin PCA.



Entrenamiento y Evaluación de Modelos de Regresión

- **Regresión lineal**
- **Random Forest Regressor**

Tamaño del DataFrame original: 2525570

Tamaño del DataFrame muestreado (1%): 25256

Best parameters for Linear Regression: {'regressor__fit_intercept': True}

--- Linear Regression - Test Set Evaluation ---

Mean Squared Error (MSE): 100.47

Mean Absolute Error (MAE): 7.23

R-squared (R2): 0.86

Best parameters for Random Forest Regressor: {'regressor__max_depth': None, 'regressor__n_estimators': 20}

--- Random Forest Regressor - Test Set Evaluation ---

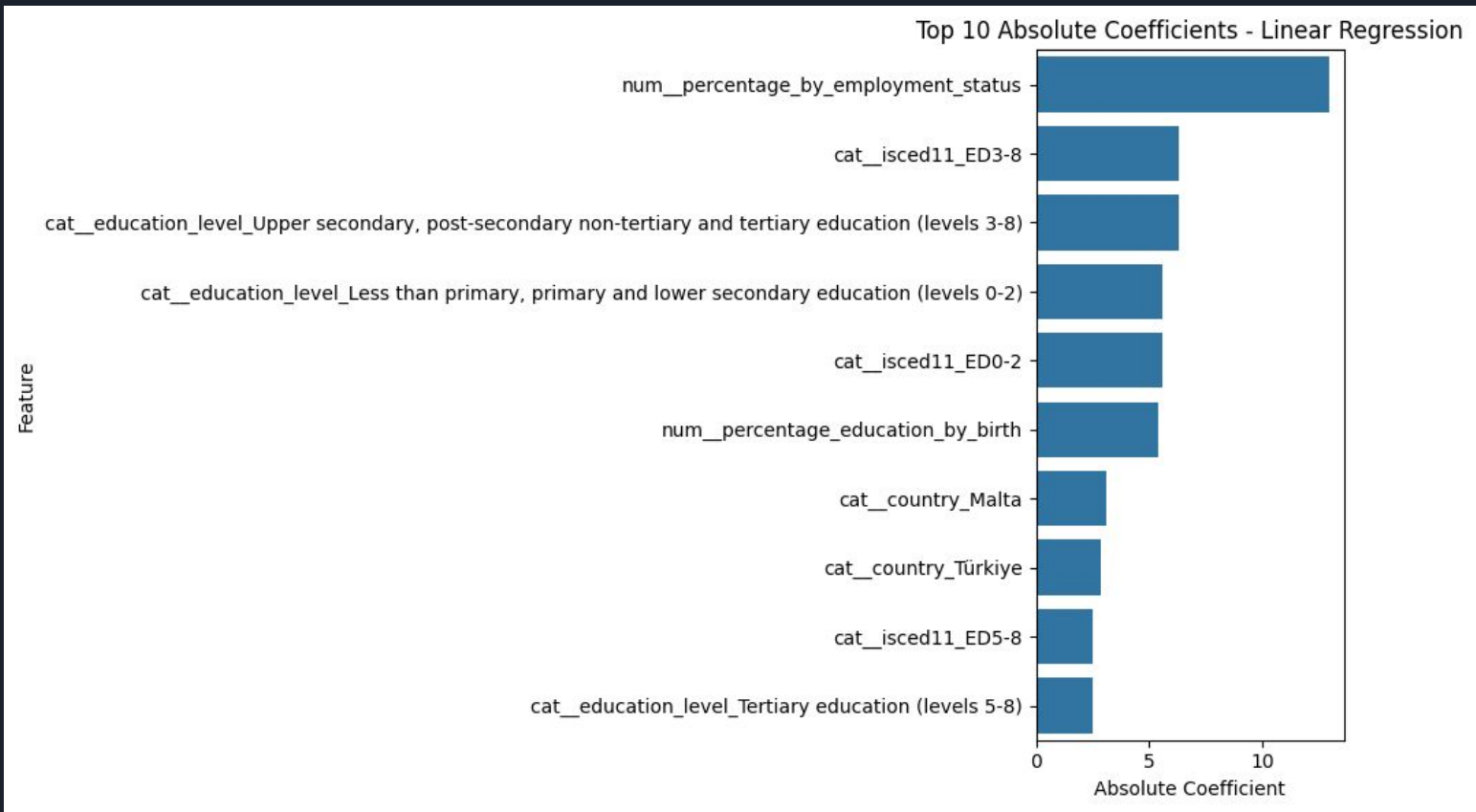
Mean Squared Error (MSE): 24.60

Mean Absolute Error (MAE): 3.29

R-squared (R2): 0.97

--- Feature Coefficients (Absolute Value) - Linear Regression ---

	feature	importance
2	num__percentage_by_employment_status	12.965728
56	cat__isced11_ED3-8	6.331764
62	cat__education_level_Upper secondary, post-sec...	6.331764
59	cat__education_level_Less than primary, primar...	5.596642
55	cat__isced11_ED0-2	5.596642
..
54	cat__Age class_From 55 to 64 years	0.047328
66	cat__age_group_55-64	0.047328
32	cat__country_Romania	0.027615
78	cat__labour_force_Population	0.017830
20	cat__country_Ireland	0.006108



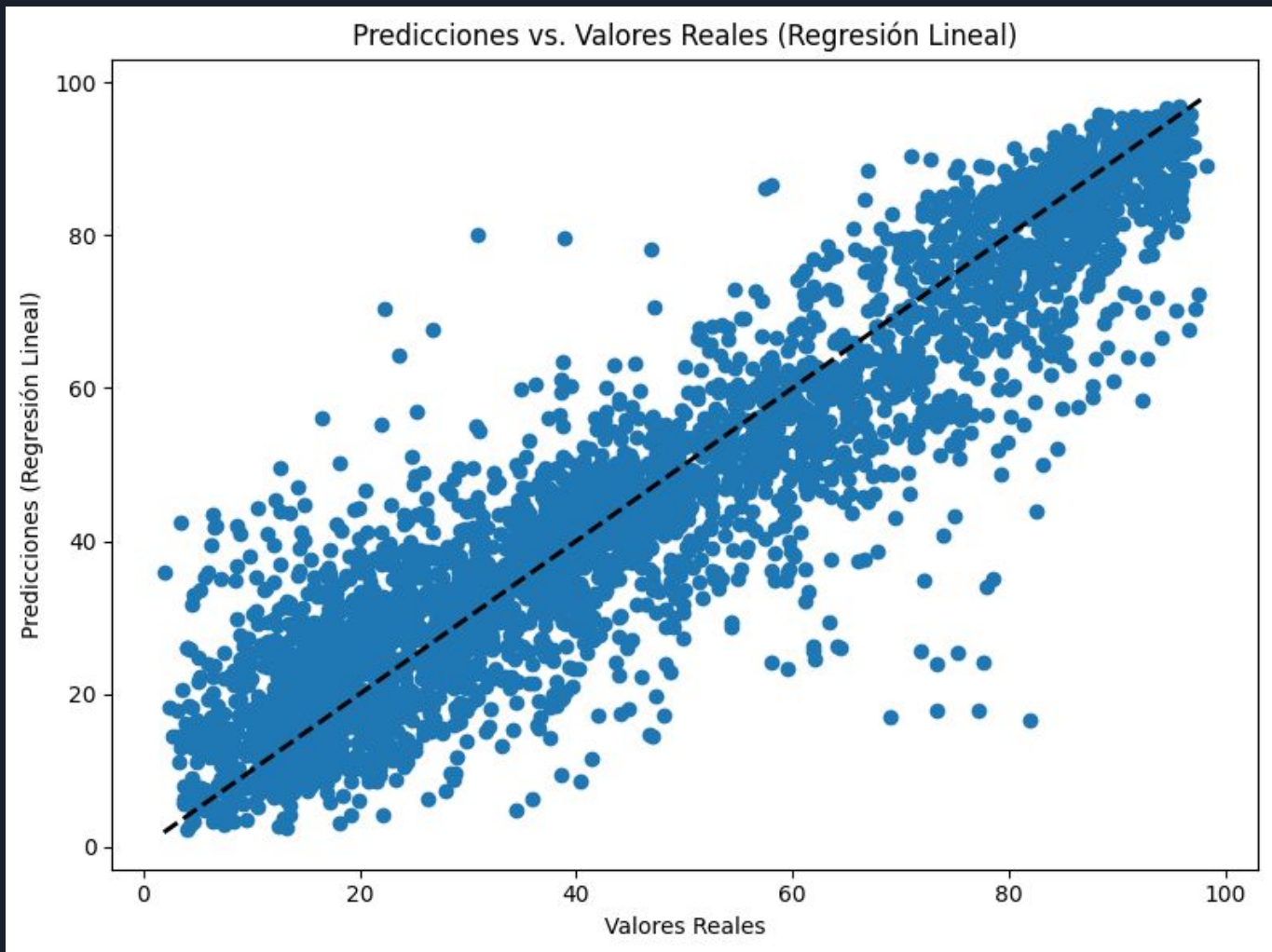
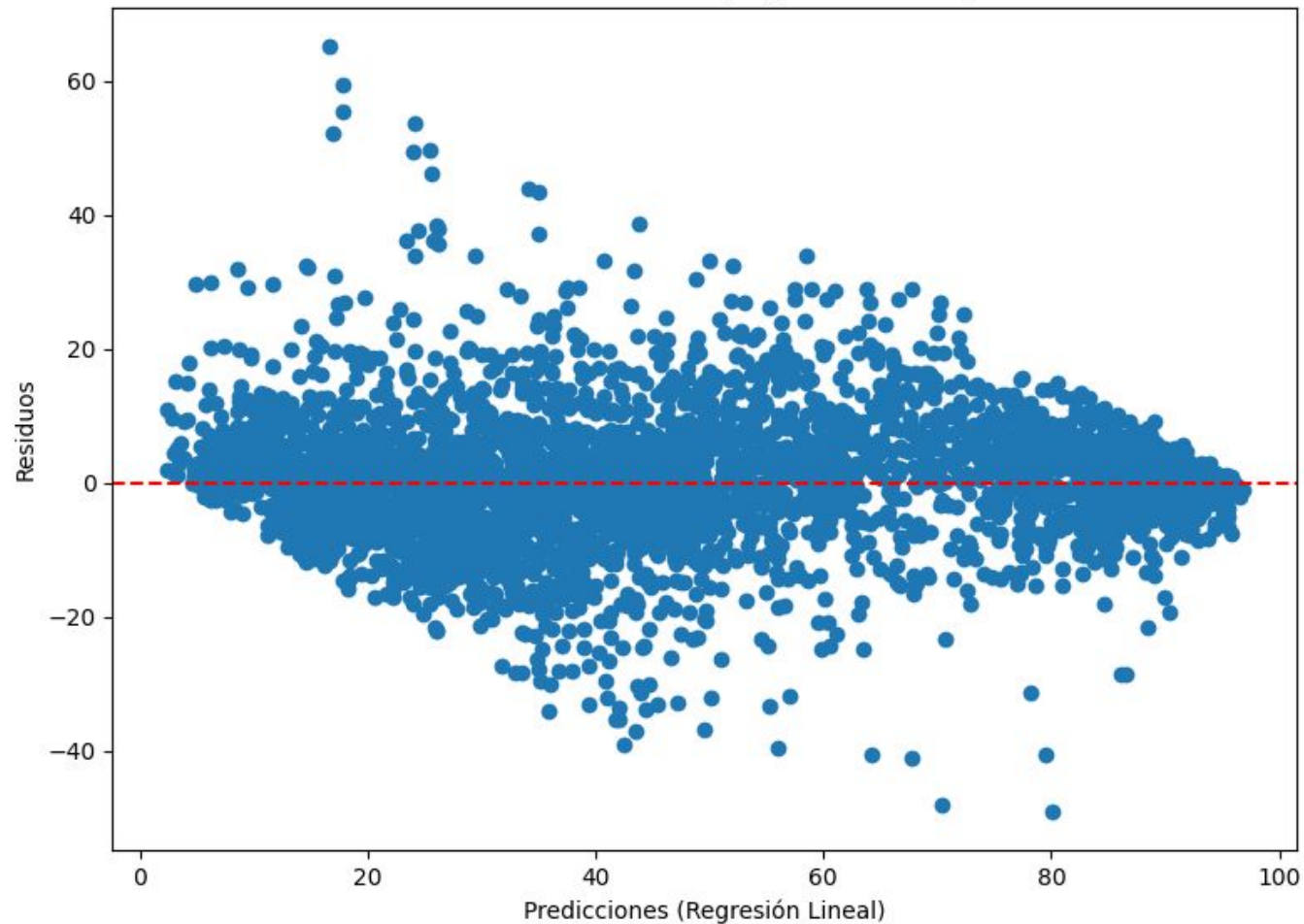
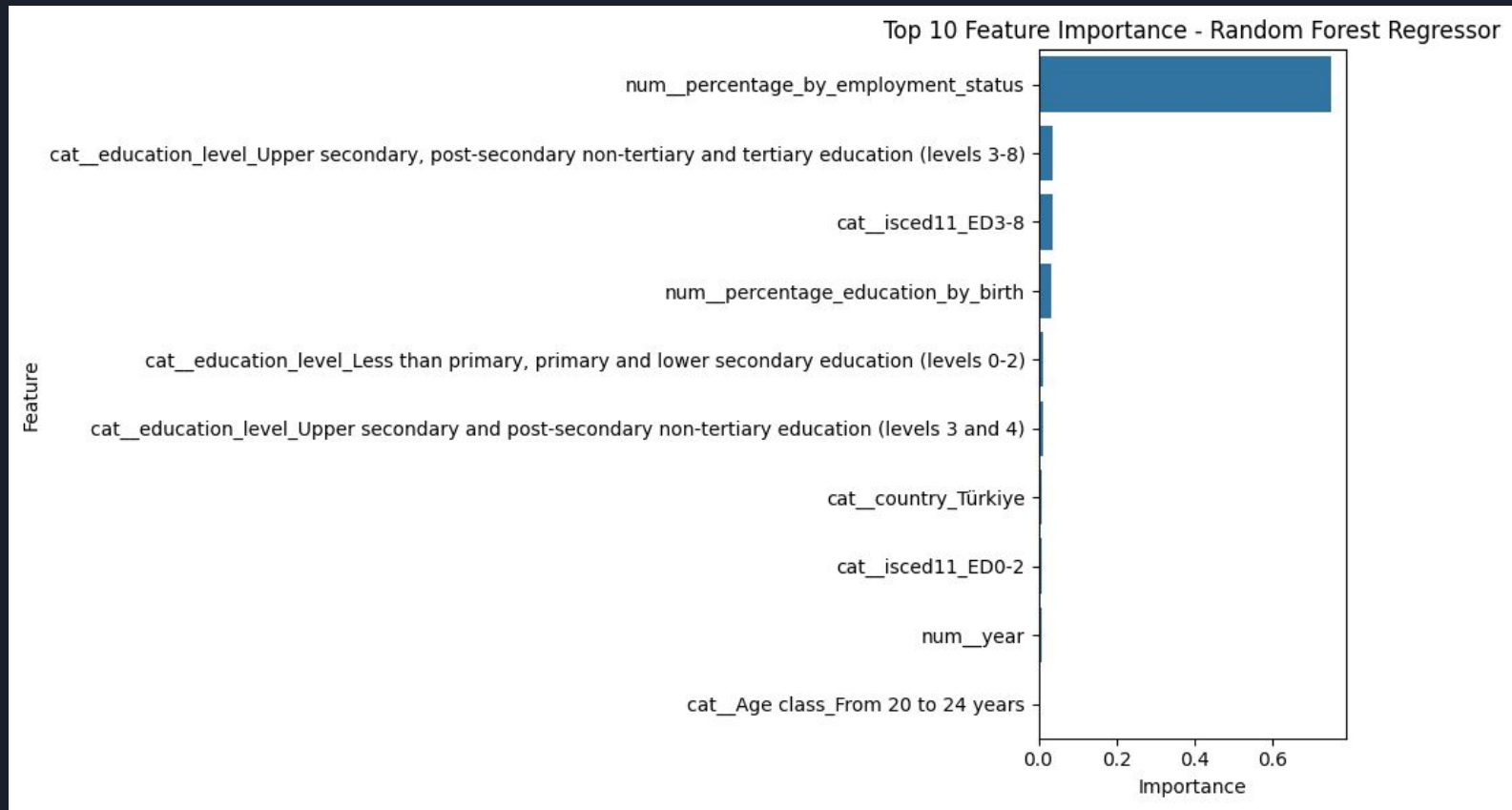


Gráfico de Residuos (Regresión Lineal)





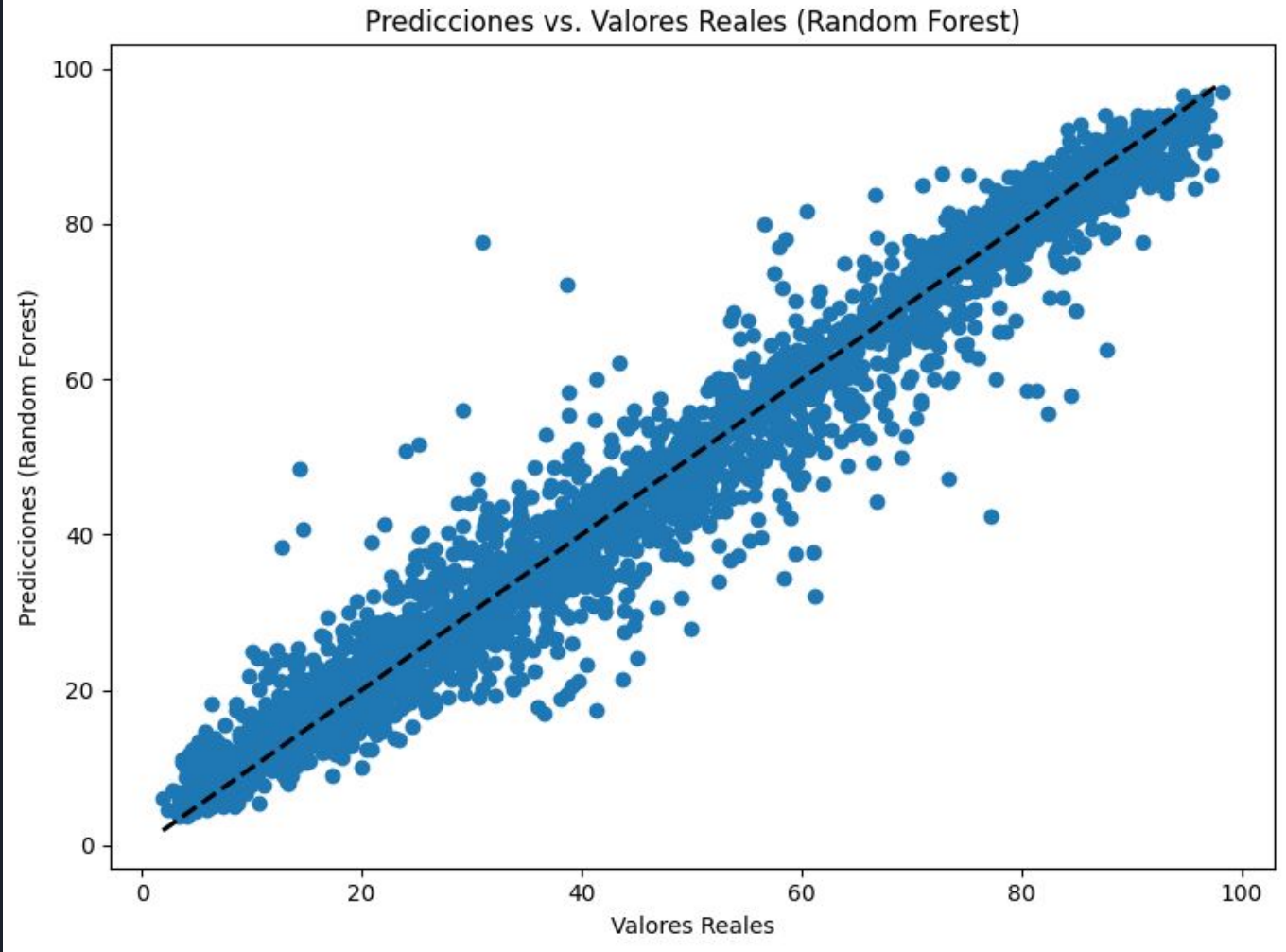
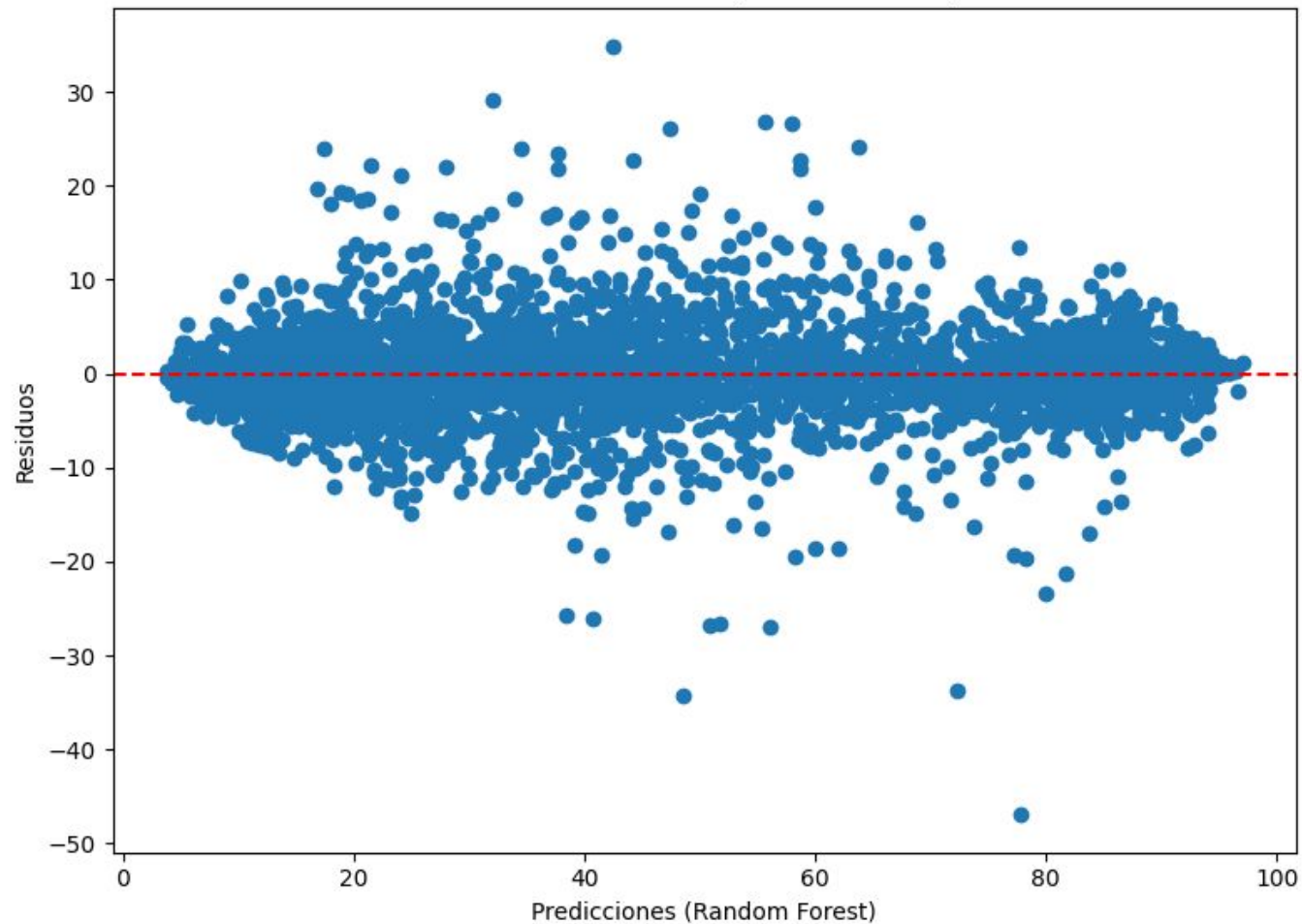



Gráfico de Residuos (Random Forest)





Entrenamiento y Evaluación de Modelos de Regresión

El **Random Forest Regressor** demostró ser significativamente mejor que la Regresión Lineal para predecir el porcentaje de educación general (R^2 de 0.97 vs. 0.86).

La variable '**percentage_by_employment_status**' fue el predictor más importante para ambos modelos, sugiriendo una fuerte conexión entre el empleo y el nivel educativo de una población.

Los **niveles educativos** también fueron consistentemente identificados como características importantes.

La capacidad del Random Forest para modelar relaciones no lineales probablemente contribuyó a su mejor rendimiento.



Resumen, Conclusión y Trabajo Futuro

Sección: Resumen de los Modelos

- **Clasificación (Predicción del Nivel de Éxito):**
 - El **Random Forest sin PCA** fue el modelo con mejor rendimiento (Accuracy del 89% en el test set).
 - La reducción de dimensionalidad con PCA no mejoró los resultados para los modelos de clasificación.
 - Los niveles educativos y el grupo de edad fueron características importantes para predecir el nivel de éxito.
- **Regresión (Predicción del Porcentaje de Educación General):**
 - El **Random Forest Regressor** superó a la Regresión Lineal (R^2 de 0.97 vs. 0.86 en el test set).
 - El porcentaje de empleo fue el predictor más importante en ambos modelos.
 - Los niveles educativos también fueron predictores relevantes.



Resumen, Conclusión y Trabajo Futuro

Sección: Conclusiones Generales


- Los modelos de **Random Forest** (tanto para clasificación como para regresión) demostraron ser más efectivos para capturar las relaciones en los datos, probablemente debido a su capacidad para modelar no linealidades.
- Las **características socioeconómicas y demográficas**, como el nivel educativo y el empleo, juegan un papel crucial en la predicción tanto del nivel de éxito como del porcentaje de educación general.
- La **reducción de dimensionalidad** (PCA) no siempre conduce a mejoras en el rendimiento y puede resultar en pérdida de información importante en ciertos casos (como en la clasificación).



Resumen, Conclusión y Trabajo Futuro

Sección: Trabajo Futuro

- **Explorar modelos más avanzados:** Investigar otros algoritmos de machine learning que puedan mejorar aún más la precisión de las predicciones (e.g., Gradient Boosting Machines, Redes Neuronales).
- **Ingeniería de características más profunda:** Crear nuevas características a partir de las existentes o explorar interacciones entre ellas para ver si se puede extraer más información predictiva.
- **Recopilación de datos adicionales:** Incorporar otras fuentes de datos relevantes que puedan influir en el nivel de éxito o el porcentaje de educación general (e.g., datos económicos a nivel regional, políticas educativas).
- **Análisis de interpretabilidad:** Profundizar en la interpretabilidad de los modelos más complejos para entender mejor las relaciones subyacentes (e.g., utilizando técnicas de interpretabilidad de modelos de caja negra).
- **Validación robusta:** Realizar una validación cruzada más exhaustiva para asegurar la generalización de los modelos a datos no vistos.
- **Despliegue y monitorización:** Considerar cómo se podrían desplegar estos modelos en un entorno real y cómo se monitorizaría su rendimiento a lo largo del tiempo.



CODEOP - BARCELONA
2025