¿PORQUE SE VAN LOS EMPLEADOS?

PREPROCESAMIENTO (INTRODUCCIÓN A LOS DATOS)

Columnas borradas -->

NúmeroDeEmpleados--> Solo hay valores 1, un solo empleado.

Over18--> Solo tenemos valor YES.

EmployeeID--> Es un valor único para cada empleado.

HorasEstándar --> Solo hay valores 1, 8 horas.

Nulos -->

Categóricas --> Creación de nueva categoría.

Numérica--> imputación por knn_imputer.

Gestión Outlaiers -->

Observamos que los outlaiers son importantes para predecir por lo que hemos decidido quedarnos con ellos.

Escalado-->

Estándar escaler.

Variable Respuesta-->

Yes = 1

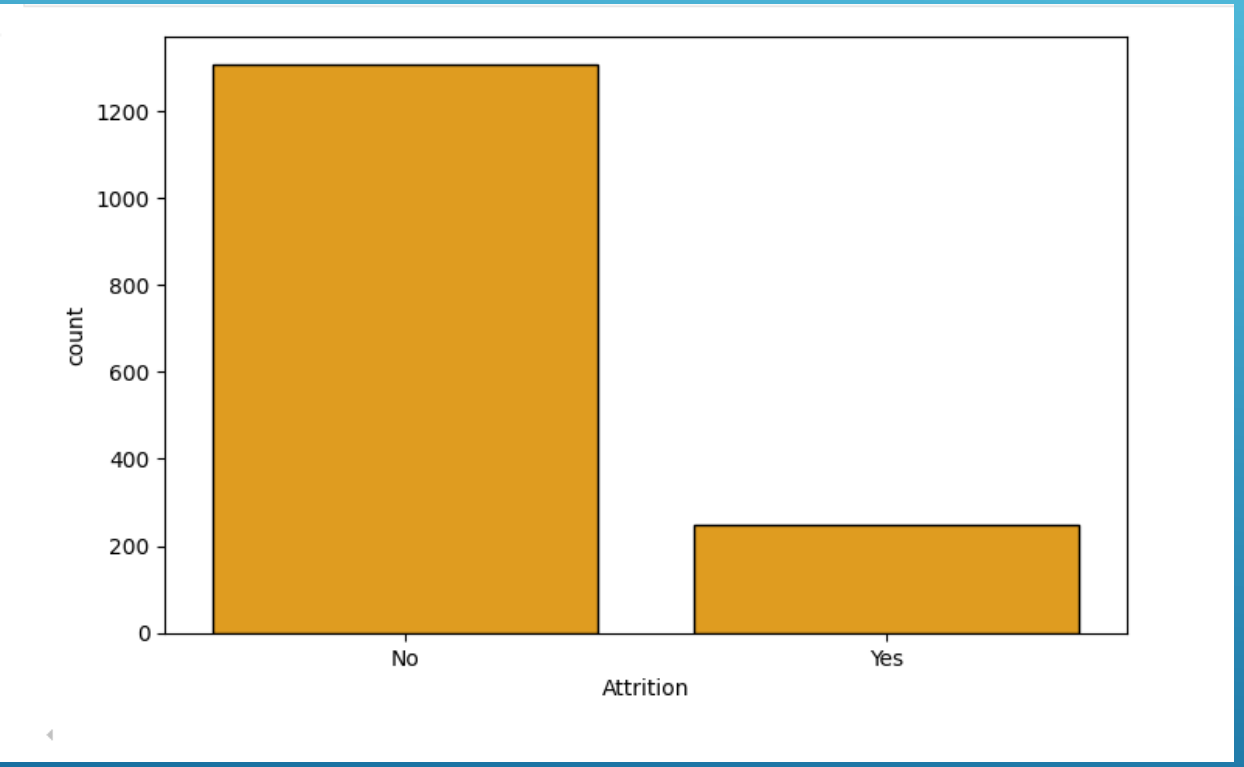
No = 0

Encoding --> Mediante una prueba de chi cuadrado.

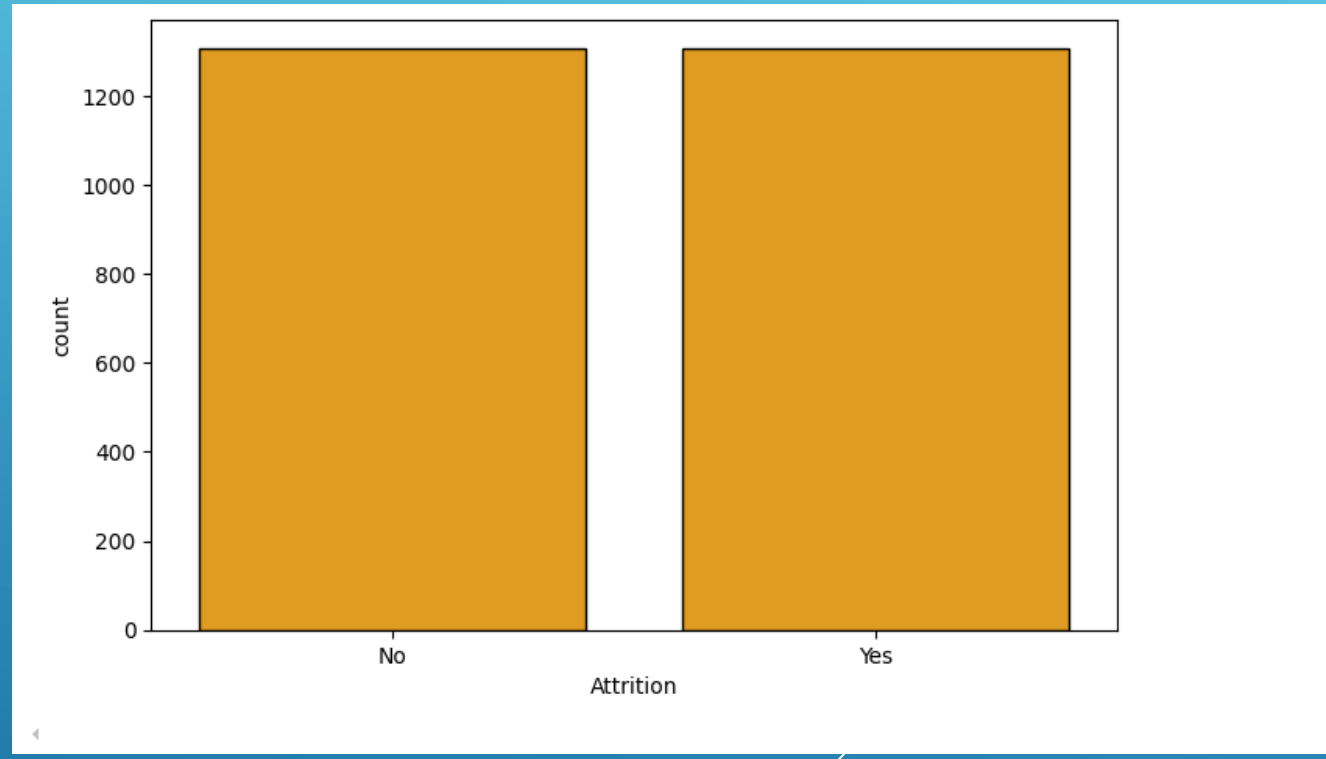
Para las que hay diferencias encodeamos con one hot encoding : MaritalStatus,JobRole,BusinessTravel, Department,EducationField.

Para las que no utilizamos target: Gender,JobLevel.

DESBALANCEO



BALANCEO MEDIANTE SMOTENC



TRAIN

accuracy	precision	recall	f1	kappa	auc	time_seconds	n_jobs	tiempo_computacion(segundos)	nucleos_usados	modelo
0.711345	0.711635	0.711345	0.711242	0.422680	0.773115	0.004988	8.000000	nan	nan	decision tree
0.728489	0.728551	0.728489	0.728478	0.456993	0.776630	nan	nan	0.004988	8.000000	decision tree
0.880326	0.881342	0.880326	0.880244	0.760645	0.943809	0.007980	8.000000	nan	nan	regresion logistica
0.860421	0.861967	0.860421	0.860282	0.720875	0.930011	nan	nan	0.007980	8.000000	regresion logistica
0.802776	0.805457	0.802776	0.802337	0.605535	0.892289	0.034905	nan	nan	nan	random_forest
0.785851	0.786950	0.785851	0.785663	0.571750	0.885467	nan	nan	0.034905	nan	random_forest
0.919100	0.920100	0.919100	0.919051	0.838196	0.969600	0.034875	8.000000	nan	nan	gradient_boosting
0.894837	0.896152	0.894837	0.894757	0.789697	0.954710	nan	nan	0.034875	8.000000	gradient_boosting
0.916707	0.918693	0.916707	0.916606	0.833408	0.966989	0.093748	nan	nan	nan	xgboost
0.898662	0.900791	0.898662	0.898535	0.797351	0.953628	nan	nan	0.093748	nan	xgboost

XGBOOST

Interpretación de las métricas:

Entrenamiento (Train):

- **Accuracy, Precision, Recall, F1 (~0.92):**

El modelo muestra un desempeño excelente al clasificar correctamente la mayoría de los datos de entrenamiento, con un equilibrio adecuado entre precisión y recall. Esto significa que identifica bien tanto los positivos como los negativos.

- **Kappa (0.83):**

Este valor indica un alto nivel de acuerdo entre las predicciones del modelo y los valores reales, más allá del azar. Es consistente con las demás métricas altas.

- **AUC (0.97):**

Este valor muestra que el modelo discrimina casi perfectamente entre las clases positivas y negativas en el conjunto de entrenamiento.

Prueba (Test):

- **Accuracy (~0.89):**

El modelo mantiene un alto rendimiento en los datos de prueba, lo que indica que generaliza bien y no está sobreajustado.

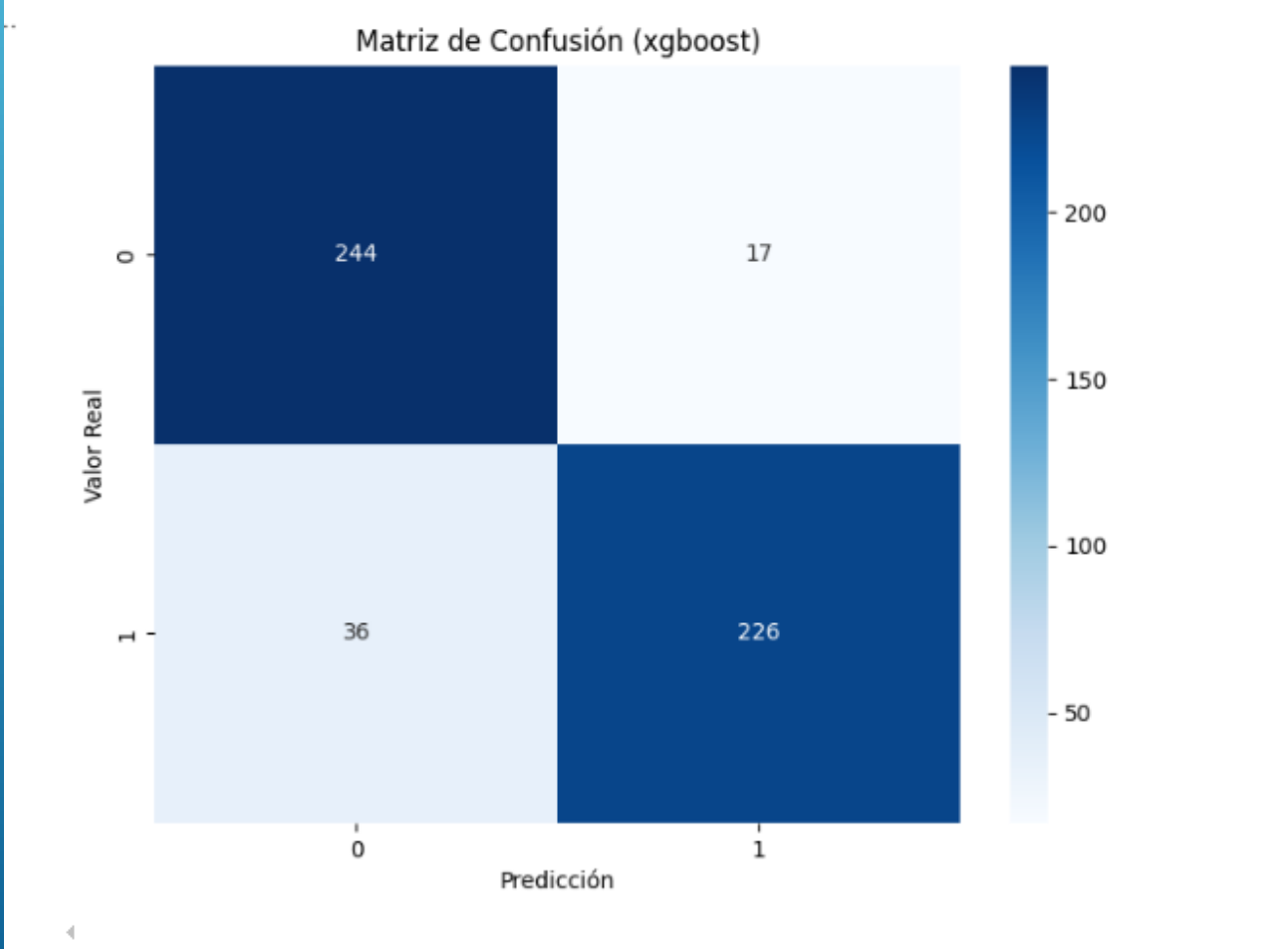
- **AUC (0.95):**

Una AUC muy alta en los datos de prueba sugiere que el modelo sigue siendo excelente para distinguir entre las clases.

Conclusión general:

El modelo **XGBoost** tiene un desempeño sobresaliente, muy similar al de **Gradient Boosting**, pero con un ligero margen superior en cuanto a capacidad de generalización. Los valores de las métricas de prueba (test) y entrenamiento (train) están cercanos, lo que indica que no hay señales significativas de overfiting.

MATRIZ DE CONFUSION



Matriz de Confusión:

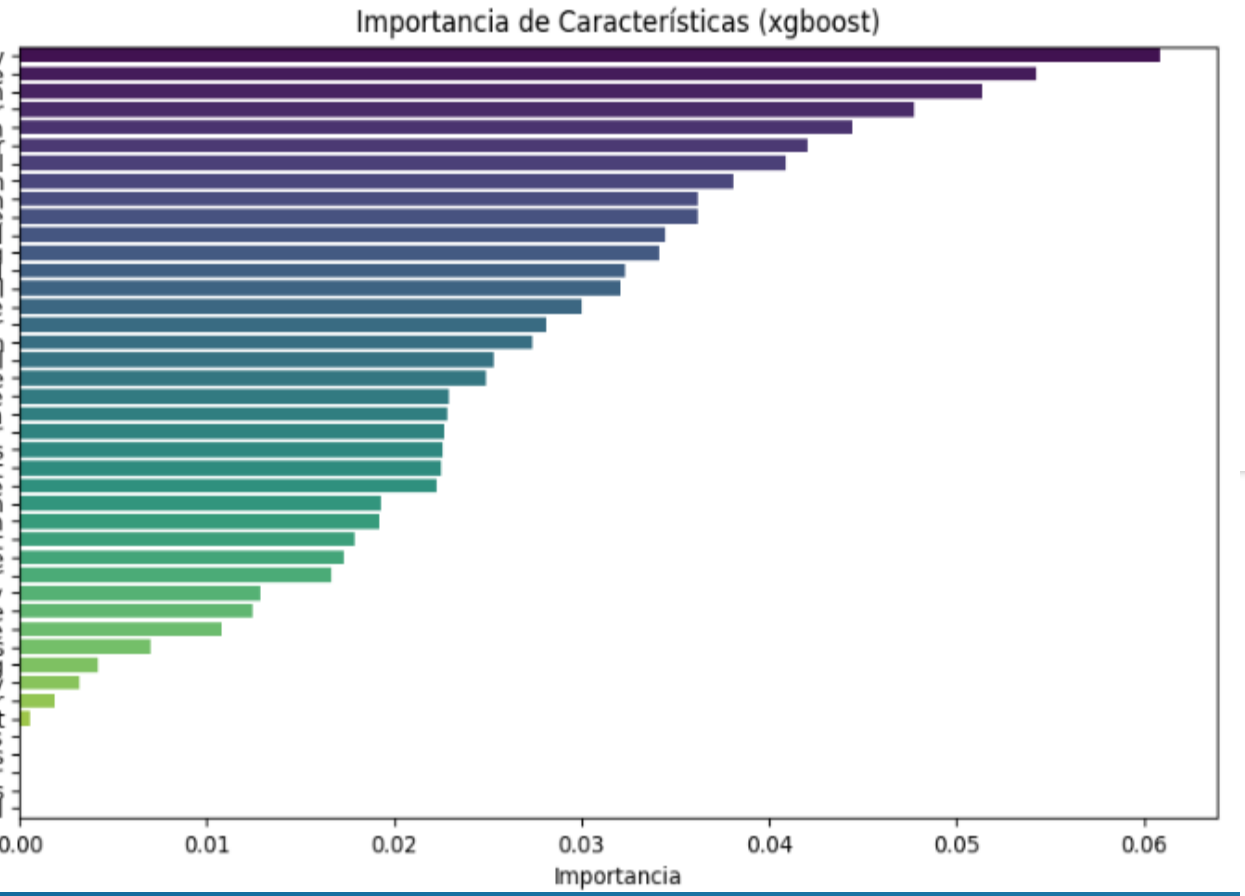
Verdaderos Negativos (TN): 244 --> El modelo predijo "No" (0) y el valor real también era "No" (0).

Falsos Positivos (FP): 17 --> El modelo predijo "Sí" (1) pero el valor real era "No" (0)

Falsos Negativos (FN): 36 --> El modelo predijo "No" (0) pero el valor real era "Sí" (1).

Verdaderos Positivos (TP): 226 --> El modelo predijo "Sí" (1) y el valor real también era "Sí" (1).

IMPORTANCIA



IMPORTANCIA ALTA

YearsAtCompany:

La característica más influyente en el modelo.

Los años en la empresa son cruciales, sugiriendo que la antigüedad afecta significativamente el resultado modelado.

MaritalStatus_Single:

Estar soltero es una característica importante que influye notablemente en las predicciones del modelo.

EnvironmentSatisfaction_knn:

La satisfacción con el entorno laboral también juega un papel clave, indicando que un ambiente de trabajo positivo tiene un impacto significativo en el resultado.

Importancia baja

EducationField_Medical:

Sugiriere que el campo educativo en medicina no es tan relevante para el resultado modelado.

EducationField_Human Resources:

La educación en recursos humanos tiene un impacto menor en el modelo.

JobRole_Research Director:

El rol de director de investigación también tiene una influencia mínima en el modelo.

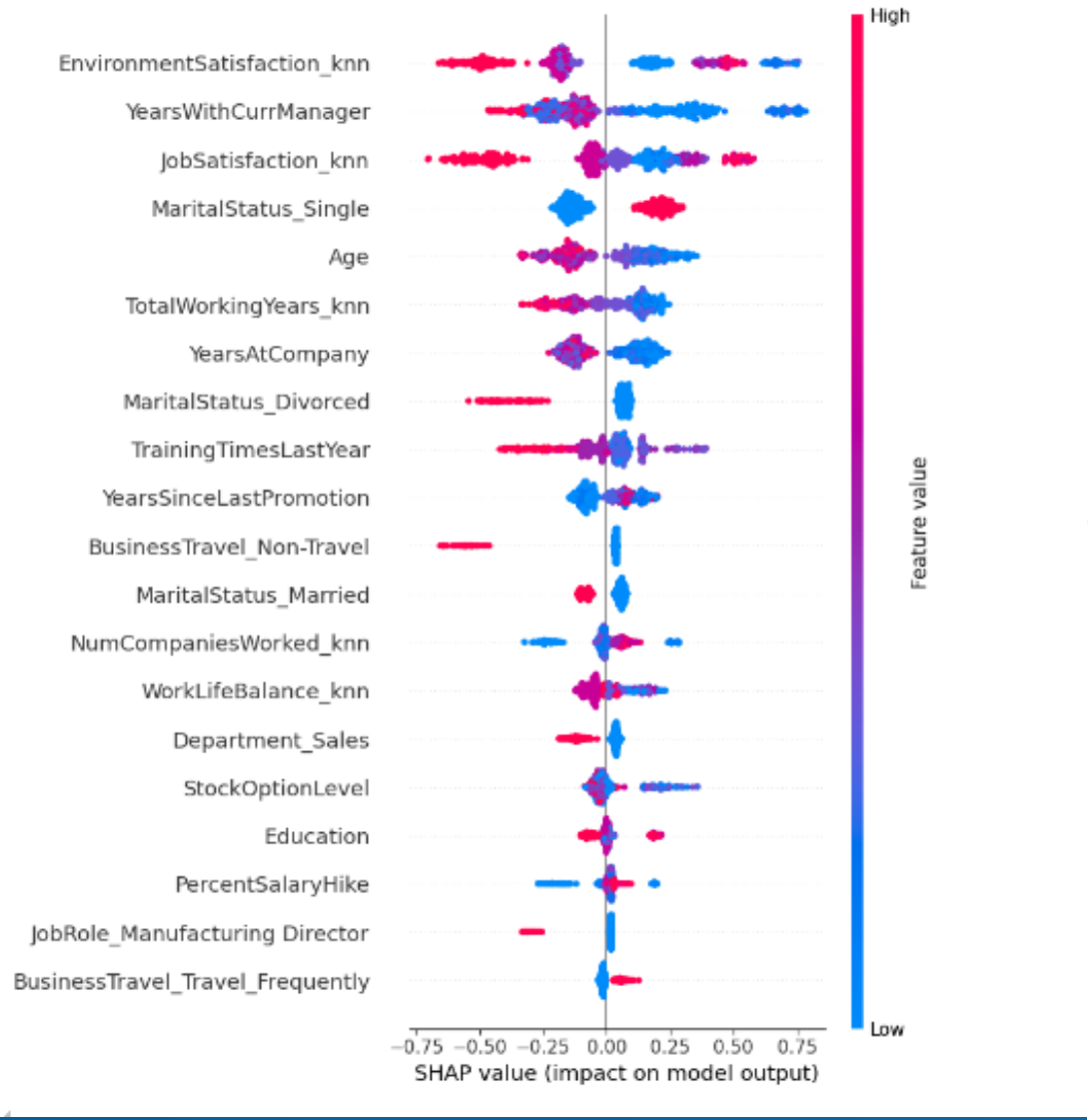
Department_Human Resources:

El departamento de recursos humanos tiene un impacto menor en el modelo.

JobRole_Human Resources:

Similar a las otras características de recursos humanos, el rol específico en este departamento no es muy influyente en el modelo.

SHAP



EnvironmentSatisfaction_knn:

La satisfacción con el entorno laboral tiene un impacto significativo en las predicciones del modelo. Los puntos rojos (altos) aumentan la probabilidad del resultado positivo, indicando que una alta satisfacción con el entorno laboral es beneficiosa.

YearsWithCurrManager:

Los años con el gerente actual también tienen un impacto significativo. Los puntos rojos (más años) tienden a aumentar la probabilidad del resultado positivo, sugiriendo que una relación estable con el gerente es importante.

JobSatisfaction_knn:

La satisfacción laboral influye considerablemente en las predicciones. Los valores altos (rojos) tienden a tener un impacto positivo, indicando que empleados satisfechos con su trabajo son más propensos a resultados positivos.

Age:

La edad de los empleados influye de manera importante en las predicciones. Los valores altos (rojos) muestran un impacto significativo, lo que sugiere que la edad está correlacionada con el resultado modelado.

YearsAtCompany:

Los años en la empresa son relevantes. Valores altos (más años) tienen un impacto positivo en las predicciones del modelo.

TotalWorkingYears_knn:

La experiencia laboral total también es una característica crucial. Los puntos rojos (más años de experiencia) tienden a tener un impacto positivo en las predicciones del modelo.

EJEMPLO DE PREDICCIÒN

```python
data = {
    'Age': 51,
    'Attrition': 'Yes',
    'BusinessTravel': 'Travel_Rarely',
    'Department': 'Sales',
    'DistanceFromHome': 6,
    'Education': 2,
    'EducationField': 'Life Sciences',
    'Gender': 'Female',
    'JobLevel': 1,
    'JobRole': 'Healthcare Representative',
    'MaritalStatus': 'Married',
    'MonthlyIncome': 131160,
    'PercentSalaryHike': 11,
    'StockOptionLevel': 0,
    'TrainingTimesLastYear': 6,
    'YearsAtCompany': 1,
    'YearsSinceLastPromotion': 0,
    'YearsWithCurrManager': 0,
    'JobInvolvement': 3,
    'PerformanceRating': 3,
    'EnvironmentSatisfaction_knn': 3.0,  # Proxy of 'EnvironmentSatisfaction'
    'JobSatisfaction_knn': 4.0,  # Proxy of 'JobSatisfaction'
    'WorkLifeBalance_knn': 4.0,
    'NumCompaniesWorked_knn': 1.0,  # Proxy of 'NumCompaniesWorked'
    'TotalWorkingYears_knn': 1.0  # Proxy of 'TotalWorkingYears'
}
```

```python
# Llamamos la función
prediction, prob = sp.predecir_attrition(data)

# Mostrar los resultados
print(f"Predicción: {prediction} (0: No, 1: Sí)")
print(f"Probabilidades: {prob}")
print(f"La probabilidad de que el empleado no se vaya es:{prob[0] * 100:.2f}%")
print(f"La probabilidad de que el empleado si se  vaya es: {prob[1] * 100:.2f}%")
```

✓ 0.0s

```
Predicción: 0 (0: No, 1: Sí)
Probabilidades: [0.8222095  0.17779054]
La probabilidad de que el empleado no se vaya es:82.22%
La probabilidad de que el empleado si se  vaya es: 17.78%
```