1. Resumen

Este proyecto tiene como objetivo predecir la rotación de empleados en una empresa de servicios financieros utilizando modelos de Machine Learning. Se emplea un enfoque basado en regresión logística y bosques aleatorios (*Random Forests*). Este proyecto busca identificar a los empleados con alto riesgo de abandonar la empresa y proporcionar insights clave sobre los factores que inciden en esta decisión. Los resultados permiten a la empresa implementar estrategias de retención más efectivas, con el objetivo de reducir la rotación de personal en un 10% ó más, impactando positivamente en costos y en la estabilidad de talento.

2. Introducción

La rotación de empleados es un desafío crítico en muchas empresas, especialmente en sectores altamente competitivos como los servicios financieros, donde el alto nivel de competencia por el talento implica costos altos de reemplazo y pérdida de experiencia organizacional, por lo tanto:

-- La alta rotación puede resultar en costos significativos y en la pérdida de talento valioso.

--Este proyecto explora cómo los datos históricos de empleados pueden ser utilizados para predecir la probabilidad de rotación y así ayudar a la empresa a tomar mejores decisiones.

3. El problema (Contexto)

En los últimos dos años, la empresa ha experimentado un aumento considerable en la rotación de empleados en los últimos dos años, lo que se ha traducido en altos costos de reclutamiento, capacitación y pérdida de talento especializado. Los efectos secundarios de esta rotación incluyen una disminución de la productividad, afectación de la cultura organizacional y un incremento en la carga laboral de los empleados restantes, que a menudo experimentan una disminución en la moral y el compromiso. Es crucial identificar los factores que contribuyen a la rotación (como salario, crecimiento profesional, satisfacción laboral, y otros), y predecir qué empleados tienen más probabilidades de abandonar la empresa podría ayudar a mitigar estos costos.

La alta rotación de empleados con lleva costos financieros, pérdida de productividad y afectación a la cultura organizacional. La empresa necesita identificar factores de rotación para implementar estrategias de retención.

4. Propósito del estudio

Desarrollar un modelo predictivo que permita a la empresa identificar a los empleados con mayor riesgo de abandonar su puesto. Esto permitirá implementar estrategias personalizadas de retención y reducir la tasa de rotación, como programas de desarrollo profesional, revisiones de salario, y mejoras en la experiencia laboral.

El objetivo es ayudar a la empresa a reducir la rotación al identificar patrones en el comportamiento de los empleados que puedan predecir quiénes están en riesgo de renunciar.

De este modo, el proyecto busca proporcionar un valor medible a la empresa a través de una disminución de costos operativos y una mejora en la satisfacción y estabilidad de su fuerza laboral.

Se va a desarrollar código en Python para la predicción de rotación, se emplearán dos modelos: **regresión logística** para analizar la relación entre las variables y la probabilidad de rotación, y **bosques aleatorios** (Random Forests) para evaluar la importancia de diferentes factores y mejorar la precisión del modelo. Los modelos se evaluarán mediante métricas como precisión, sensibilidad y especificidad, permitiendo ajustes y optimización.

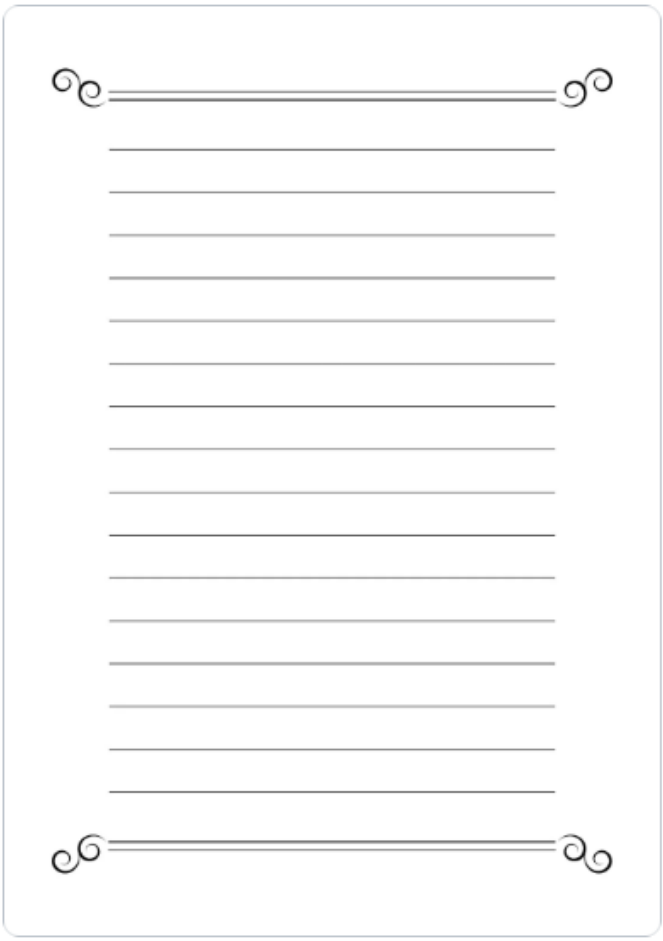
5. Descripción del proyecto

Para efectos del presente proyecto se usará el archivo de muestra "Employee Attrition Data.csv", el cual se obtuvo de la página de datasets públicos: https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset?resource=download

Este proyecto comprende la recopilación y análisis de datos históricos de empleados, que incluyen variables como:

- **Edad (Age)**: Edad del empleado. Conocimiento de los rangos de edad puede aportar información sobre la relación entre etapa de vida y permanencia.
- **Género (Gender)**.
- **Antigüedad en la Empresa (Years at Company)**: Años que lleva el empleado en la empresa. La relación entre el tiempo de servicio y la rotación es relevante para identificar patrones de riesgo.
- **Rol Laboral (Job Role)**: Área o departamento en el que trabaja el empleado (ej. Educación, Salud). La rotación puede variar según el área de trabajo debido a diferencias en cultura, demandas y oportunidades de crecimiento.
- **Ingreso Mensual (Monthly Income)**: Salario mensual del empleado. El nivel salarial es un factor crítico para la permanencia, especialmente en sectores con alta competencia por talento.
- **Equilibrio Trabajo-Vida (Work-Life Balance)**: Calificación de la percepción del balance entre vida laboral y personal.
- **Satisfacción en el Trabajo (Job Satisfaction)**. Indicadores de satisfacción pueden señalar el nivel de compromiso y la predisposición a permanecer.
- **Calificación de Desempeño (Performance Rating)**. Los empleados con calificaciones de desempeño altas o bajas pueden tener distintas motivaciones y riesgos de rotación.
- **Promociones Recibidas (Number of Promotions)**: Número de veces que el empleado ha sido promovido.
- **Tiempo de Desplazamiento (Distance from Home)**. Tiempo para llegar al trabajo.
- **Nivel Educativo (Education Level)**.
- **Estado Civil (Marital Status)**.
- **Dependientes (Number of Dependents)**.
- **Nivel del Puesto (Job Level)**: Nivel jerárquico del empleado (ej. Junior, Senior).
- **Tamaño de la Empresa (Company Size)**.
- **Oportunidades de Liderazgo y de Innovación**.
- **Reconocimiento al Empleado (Employee Recognition)**.
- **Rotación (Attrition)**: Si el empleado se fue ("Left") o se quedó ("Stayed").

Se utilizarán modelos de regresión logística y Random Forests para predecir la probabilidad de rotación de cada empleado.

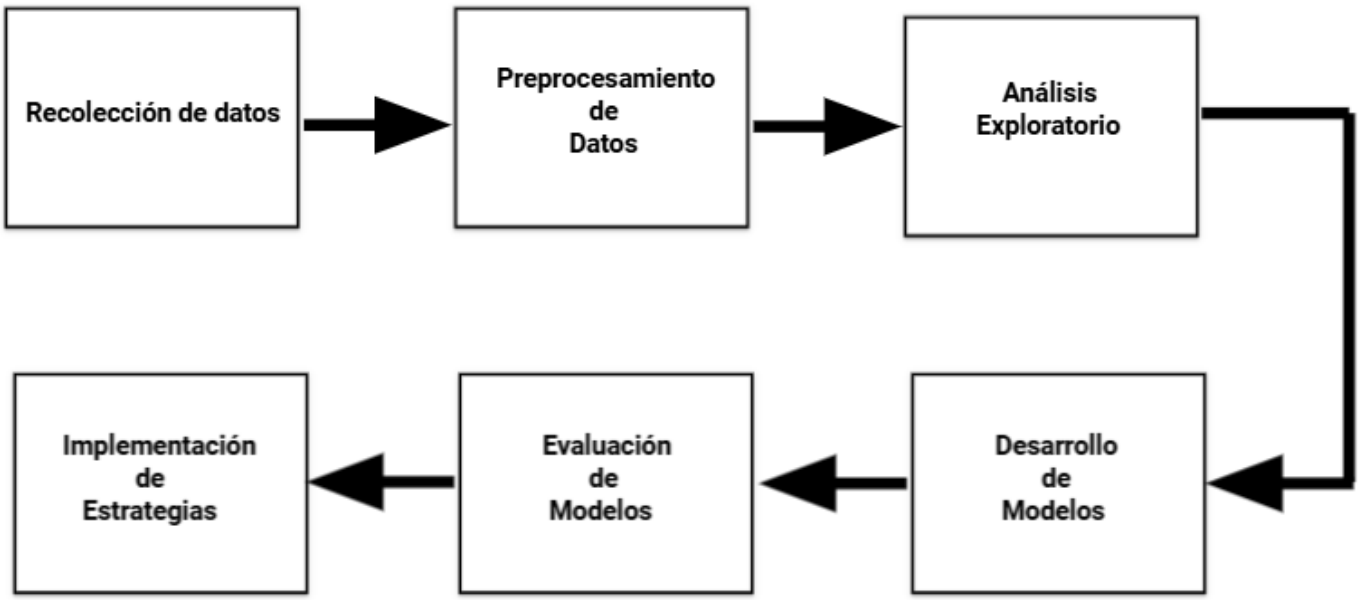


6. Hipótesis

Factores como el salario mensual, la antigüedad en la empresa, el equilibrio trabajo-vida y la satisfacción laboral son determinantes clave para predecir la rotación de empleados.

7. Flujo de trabajo.

7. Flujo de trabajo.



1.-Recolección de datos: Obtención de datos históricos de empleados.

2.- Preprocesamiento de datos: Limpieza y codificación de variables categóricas.

3.-Análisis Exploratorio: Identificación de patrones y relaciones en los datos.

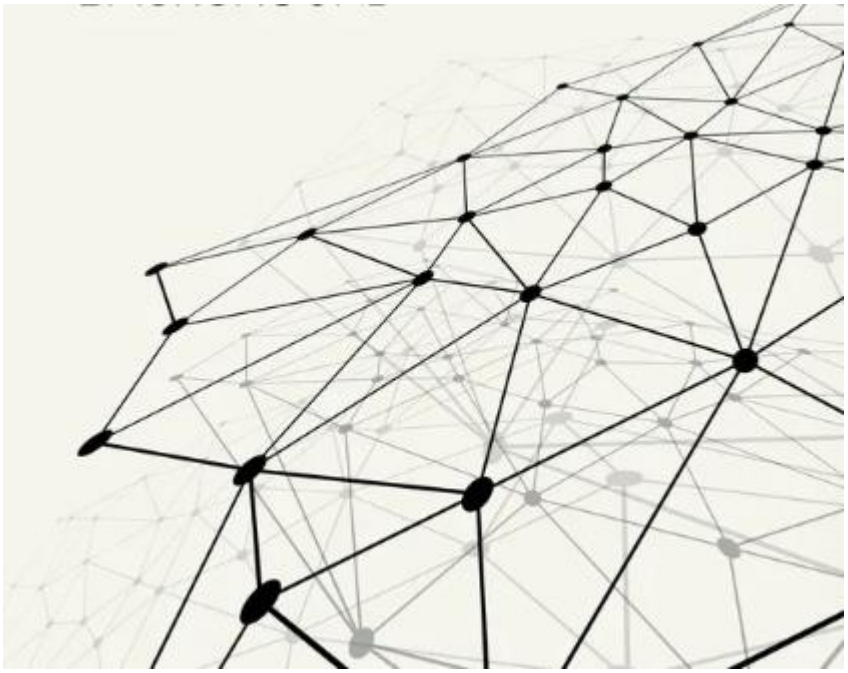
4.-Desarroll de modelos: Entrenamiento de modelos de regresión logística y Random Forest.

5.-Evaluación de Modelos: Validación y comparación de modelos utilizando métricas como la precisión y la curva ROC-AUC.

6.-Implementación de Estrategias: Identificación de empleados en riesgo y desarrollo de estrategias de retención.

8. Mapeo del Sistema

El sistema se estructura en un módulo de predicción, que recibe los datos de los empleados y predice la probabilidad de rotación, y un módulo de visualización, que presenta los resultados de manera comprensible para los gerentes de recursos humanos.



1. **Recolección de datos**:
    - Cargar el archivo "Employee Attrition Data.csv".
    - Limpieza de datos: detección de valores nulos.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.impute import SimpleImputer
import os
os.chdir("/Users/gilgu/OneDrive/Escritorio")


# Cargar datos
data = pd.read_csv('Employee Attrition Data.csv')
```

```
data
```

	Employee ID	Age	Gender	Years at Company	Job Role	Monthly Income	Work-Life Balance	Job Satisfaction	Performance Rating	Number of Promotions	...	Number of Dependents	Job Level	Company Size	Compa Ten
0	52685	36	Male	13	Healthcare	8029	Excellent	High	Average	1	...	1	Mid	Large	
1	30585	35	Male	7	Education	4563	Good	High	Average	1	...	4	Entry	Medium	
2	54656	50	Male	7	Education	5583	Fair	High	Average	3	...	2	Senior	Medium	
3	33442	58	Male	44	Media	5525	Fair	Very High	High	0	...	4	Entry	Medium	
4	15667	39	Male	24	Education	4604	Good	High	Average	0	...	6	Mid	Large	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
14895	16243	56	Female	42	Healthcare	7830	Poor	Medium	Average	0	...	0	Senior	Medium	
14896	47175	30	Female	15	Education	3856	Good	Medium	Average	2	...	0	Entry	Medium	
14897	12409	52	Male	5	Education	5654	Good	Very High	Below Average	0	...	4	Mid	Small	
14898	9554	18	Male	4	Education	5276	Fair	High	Average	0	...	3	Mid	Large	
14899	73042	59	Female	48	Education	3774	Good	High	Below Average	1	...	4	Mid	Large	

14900 rows × 24 columns

```python
# Visualizar las primeras filas
print(data.head())
```

```
   Employee ID  Age Gender  Years at Company    Job Role  Monthly Income  \
0        52685   36   Male                13  Healthcare            8029
1        30585   35   Male                 7   Education            4563
2        54656   50   Male                 7   Education            5583
3        33442   58   Male                44       Media            5525
4        15667   39   Male                24   Education            4604

  Work-Life Balance Job Satisfaction Performance Rating  Number of Promotions  \
0         Excellent             High            Average                     1
1              Good             High            Average                     1
2              Fair             High            Average                     3
3              Fair        Very High               High                     0
4              Good             High            Average                     0

   ... Number of Dependents  Job Level Company Size Company Tenure  \
0  ...                    1        Mid        Large             22
1  ...                    4      Entry       Medium             27
2  ...                    2     Senior       Medium             76
3  ...                    4      Entry       Medium             96
4  ...                    6        Mid        Large             45



   Remote Work Leadership Opportunities Innovation Opportunities  \
0           No                       No                      No
1           No                       No                      No
2           No                       No                     Yes
3           No                       No                      No
4          Yes                       No                      No

   Company Reputation Employee Recognition Attrition
0                Poor                Medium    Stayed
1                Good                  High      Left
2                Good                   Low    Stayed
3                Poor                   Low      Left
4                Good                  High    Stayed

[5 rows x 24 columns]
```

```python
# Identificar y tratar valores nulos
print(data.isnull().sum())
data = data.dropna()  # Si hay pocos nulos, eliminarlos; si no, considerar imputarlos
```

```
Employee ID                0
Age                        0
Gender                     0
Years at Company           0
Job Role                   0
Monthly Income             0
Work-Life Balance          0
Job Satisfaction           0
Performance Rating         0
Number of Promotions       0
Overtime                   0
Distance from Home         0
Education Level            0
Marital Status             0
Number of Dependents       0
Job Level                  0
Company Size               0
Company Tenure             0
Remote Work                0
Leadership Opportunities   0
Innovation Opportunities   0
Company Reputation         0
Employee Recognition       0
Attrition                  0
dtype: int64
```

2.- Preprocesamiento de datos:

- Transformación de datos categóricos en variables dummy para que puedan ser interpretadas por los modelos.

```python
# Codificar variables categóricas
encoder = LabelEncoder()
data['Attrition'] = encoder.fit_transform(data['Attrition'])  # Asegurar que "Attrition" esté codificado como 0 (Stayed) o 1
```

```python
# Variables predictoras y variable objetivo
X = data.drop('Attrition', axis=1)
y = data['Attrition']
```

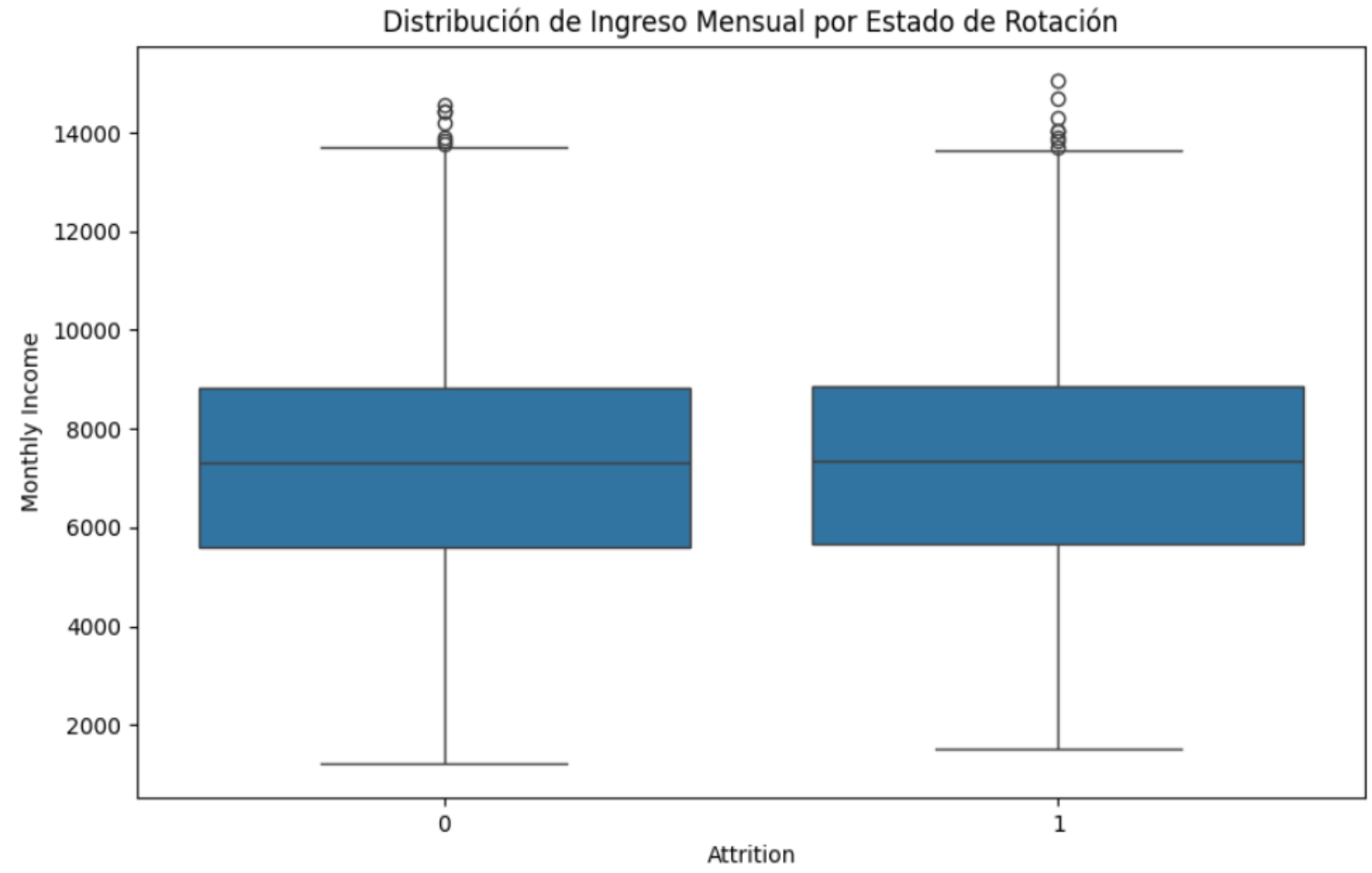
```python
# Dividir el conjunto de datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

3.-Análisis Exploratorio datos (EDA):

- Identificación de patrones y relaciones en los datos.

```
# Para identificar patrones en la rotación, se visualiza la distribución de algunas variables clave.
import seaborn as sns
import matplotlib.pyplot as plt
```

```
# Relación entre "Monthly Income" y "Attrition"
plt.figure(figsize=(10, 6))
sns.boxplot(x='Attrition', y='Monthly Income', data=data)
plt.title('Distribución de Ingreso Mensual por Estado de Rotación')
plt.show()
```



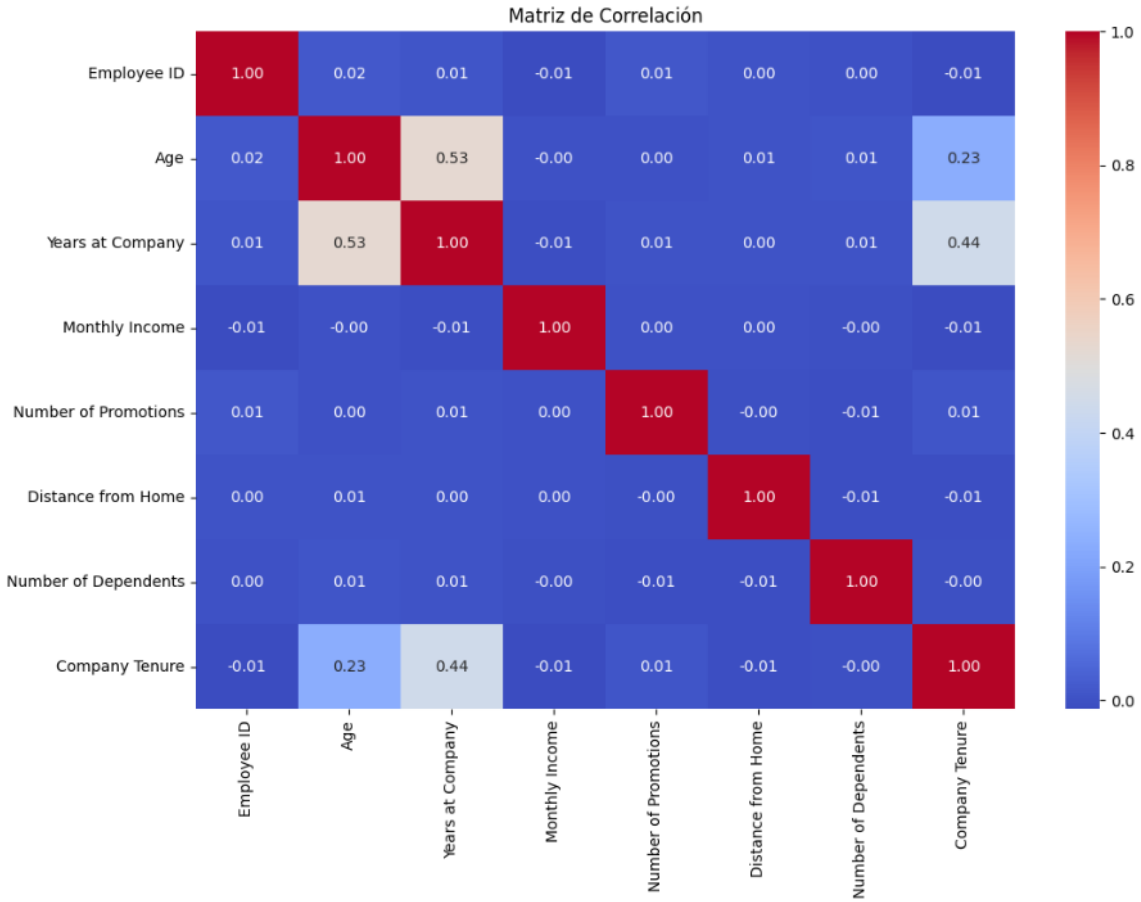
```
print(data.columns)
```

```
Index(['Employee ID', 'Age', 'Gender', 'Years at Company', 'Job Role',
       'Monthly Income', 'Work-Life Balance', 'Job Satisfaction',
       'Performance Rating', 'Number of Promotions', 'Overtime',
       'Distance from Home', 'Education Level', 'Marital Status',
       'Number of Dependents', 'Job Level', 'Company Size', 'Company Tenure',
       'Remote Work', 'Leadership Opportunities', 'Innovation Opportunities',
       'Company Reputation', 'Employee Recognition', 'Attrition'],
      dtype='object')
```

```
# Selecciona solo las columnas numéricas
numeric_data = data.select_dtypes(include=['float64', 'int64'])
```

```
# Calcula la matriz de correlación
correlation_matrix = numeric_data.corr()
```

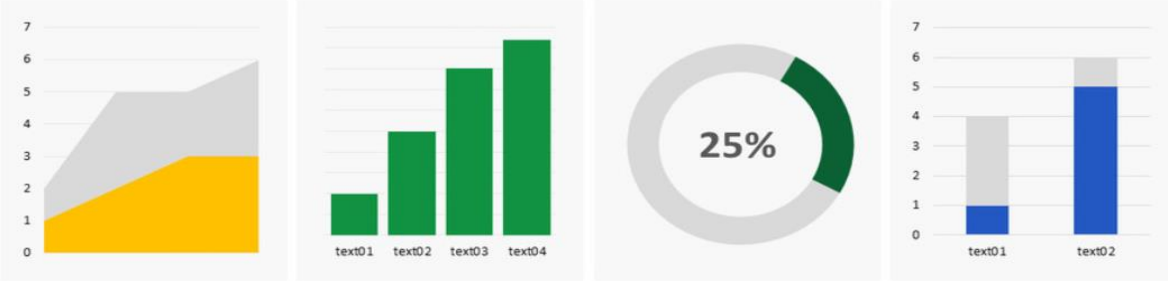
```
# Genera el heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, fmt=".2f", cmap='coolwarm')
plt.title('Matriz de Correlación')
plt.show()
```



9. Definición de Métricas Adecuadas

Para medir la eficacia del modelo de rotación, utilizaremos las siguientes métricas:

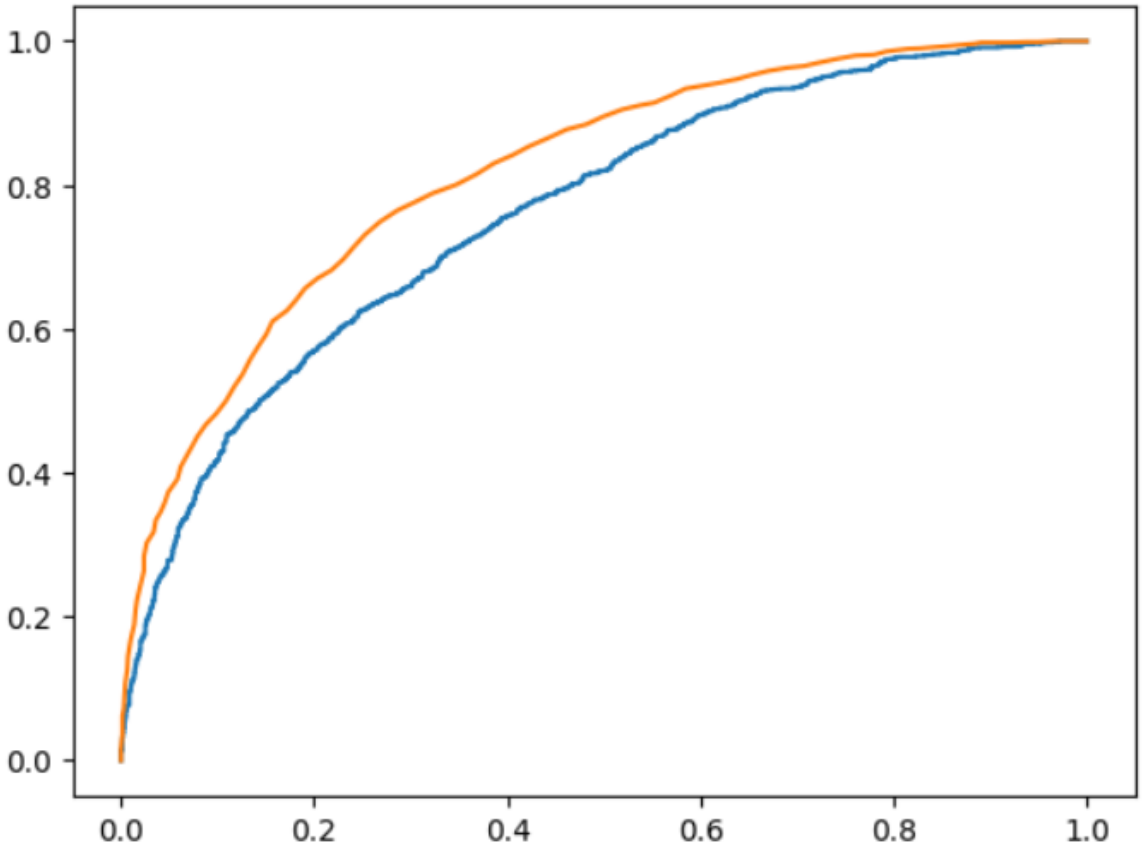
1. **Precisión**:
   - La precisión mide el porcentaje de predicciones correctas realizadas por el modelo, es decir, la proporción de empleados correctamente clasificados como que **abandonan** o **no abandonan** la empresa.
   - En este contexto, la precisión es útil para entender qué tan bien el modelo realiza predicciones generales, pero no diferencia entre las clases (abandono vs. no abandono). Por lo tanto, no proporciona información sobre posibles desbalances entre las clases.
2. **Curva ROC y AUC (Área Bajo la Curva)**:
   - La curva ROC (Receiver Operating Characteristic) evalúa el desempeño del modelo en términos de su capacidad para discriminar entre las dos clases (abandono vs. no abandono). Se genera trazando la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) para diferentes umbrales de decisión.
   - El **AUC (Área Bajo la Curva)** mide la capacidad general del modelo para distinguir entre las clases. Un valor cercano a 1 indica excelente capacidad de discriminación, mientras que un valor cercano a 0.5 indica un modelo sin capacidad predictiva.




```
from sklearn.metrics import accuracy_score, roc_auc_score, roc_curve
import numpy as np
```

```
# Predicciones y métricas
models = {'Logistic Regression': log_reg, 'Random Forest': rf_model}
for name, model in models.items():
    y_pred = model.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    roc_auc = roc_auc_score(y_test, model.predict_proba(X_test)[:, 1])
    print(f"{name} - Precisión: {accuracy:.2f}, ROC-AUC: {roc_auc:.2f}")

    # Curva ROC
    fpr, tpr, _ = roc_curve(y_test, model.predict_proba(X_test)[:, 1])
    plt.plot(fpr, tpr, label=f"{name} (AUC = {roc_auc:.2f})")
```

```
Logistic Regression - Precisión: 0.68, ROC-AUC: 0.76
Random Forest - Precisión: 0.74, ROC-AUC: 0.82
```



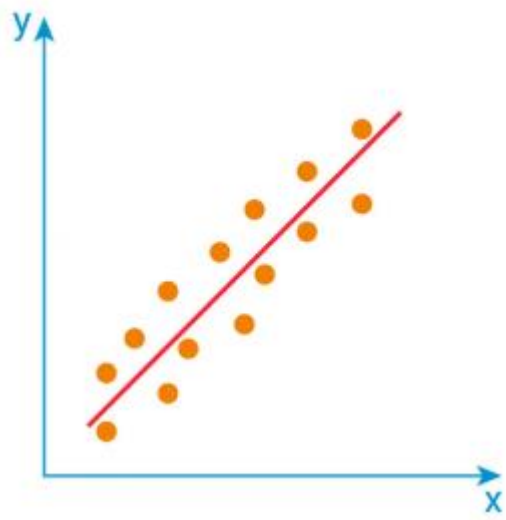
- Precisión: 0.68
- ROC-**AUC**: 0.76

La precisión indica que el modelo clasifica correctamente el 68% de los empleados. El AUC de 0.76 sugiere que el modelo tiene una capacidad moderada para distinguir entre empleados que abandonan y los que no.

10. Métodos y Modelos

Se emplearán dos modelos:

1. **Regresión Logística**: Es adecuado para interpretar la relación entre los factores y la probabilidad de rotación. Es un modelo lineal utilizado para estimar la probabilidad de un empleado de abandonar la empresa. Permite observar el impacto de cada variable en la rotación.
2. **Bosques Aleatorios (Random Forests)**: Este modelo de ensamble (basado en árboles de decisión), ayuda a mejorar la precisión de la predicción y proporciona una lista de importancia de características que facilita identificar los factores más relevantes.



1. Regresión Logística

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.decomposition import PCA
from sklearn.metrics import ConfusionMatrixDisplay
import matplotlib.pyplot as plt


# Convertir variables categóricas en variables dummy (one-hot encoding)
data = pd.get_dummies(data)
```

```
# Dividir el conjunto de datos
X = data.drop('Attrition', axis=1)  # Asegúrate de que 'Attrition' sea el nombre de tu columna objetivo
y = data['Attrition']
```

```
# Dividir en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```
# Escalar los datos
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

```
# Modelo de regresión logística
model = LogisticRegression()
model.fit(X_train_scaled, y_train)
```

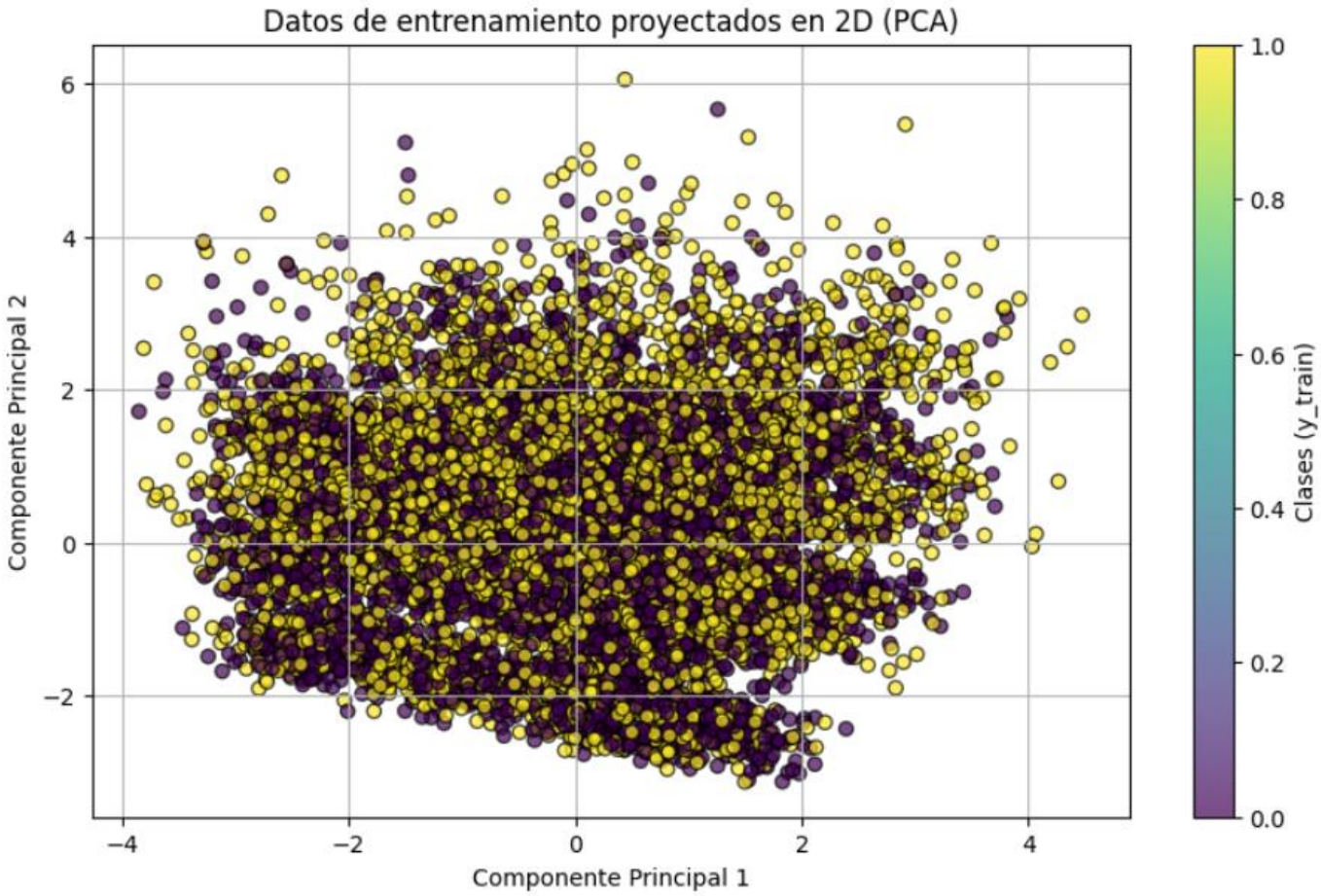
]:

```
LogisticRegression
LogisticRegression()
```

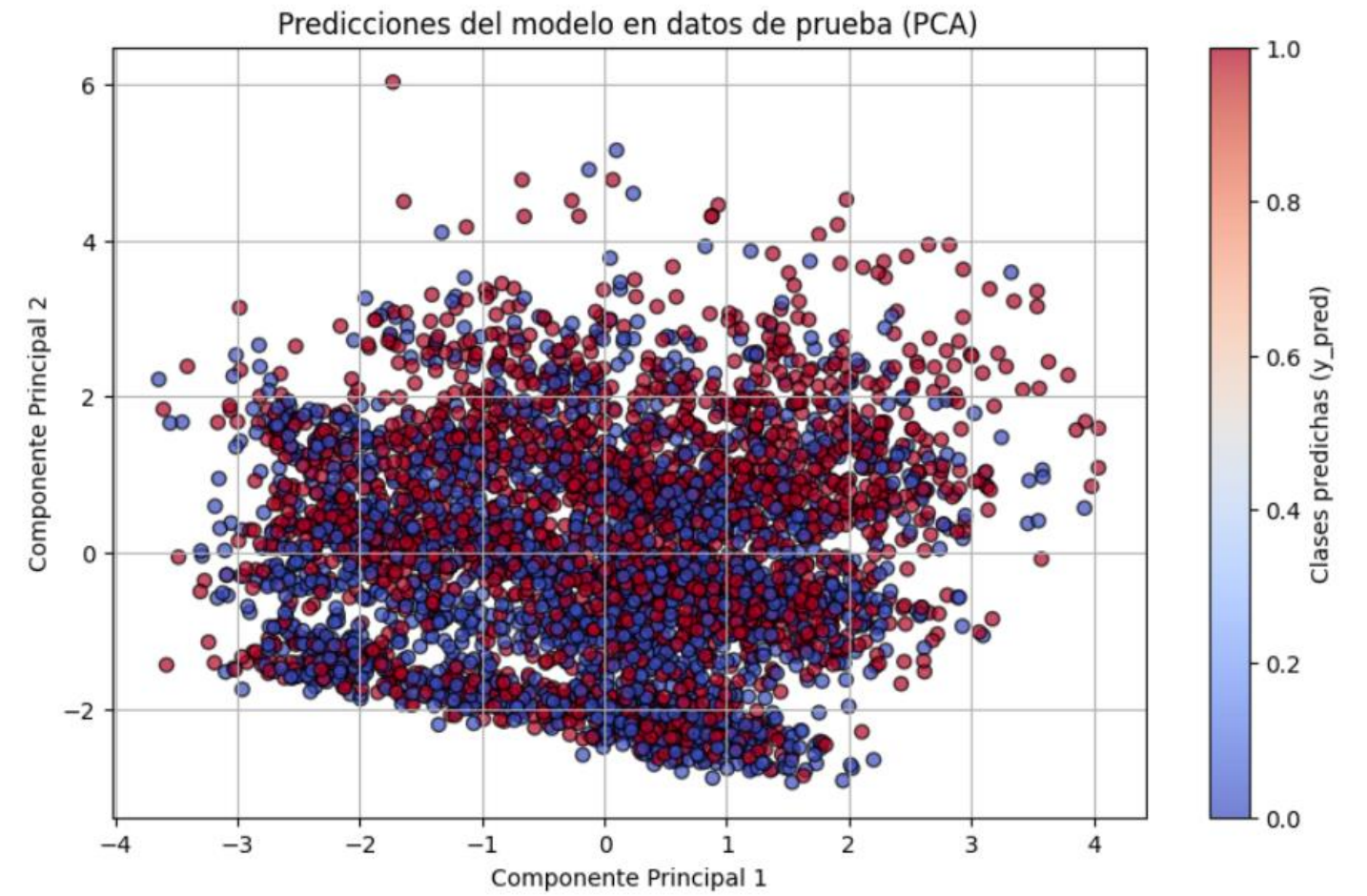
```
# Reducir la dimensionalidad para visualizar los datos (PCA a 2D)
pca = PCA(n_components=2)
X_train_pca = pca.fit_transform(X_train_scaled)
X_test_pca = pca.transform(X_test_scaled)
```

```
# Obtener las predicciones del modelo
y_pred = model.predict(X_test_scaled)
```

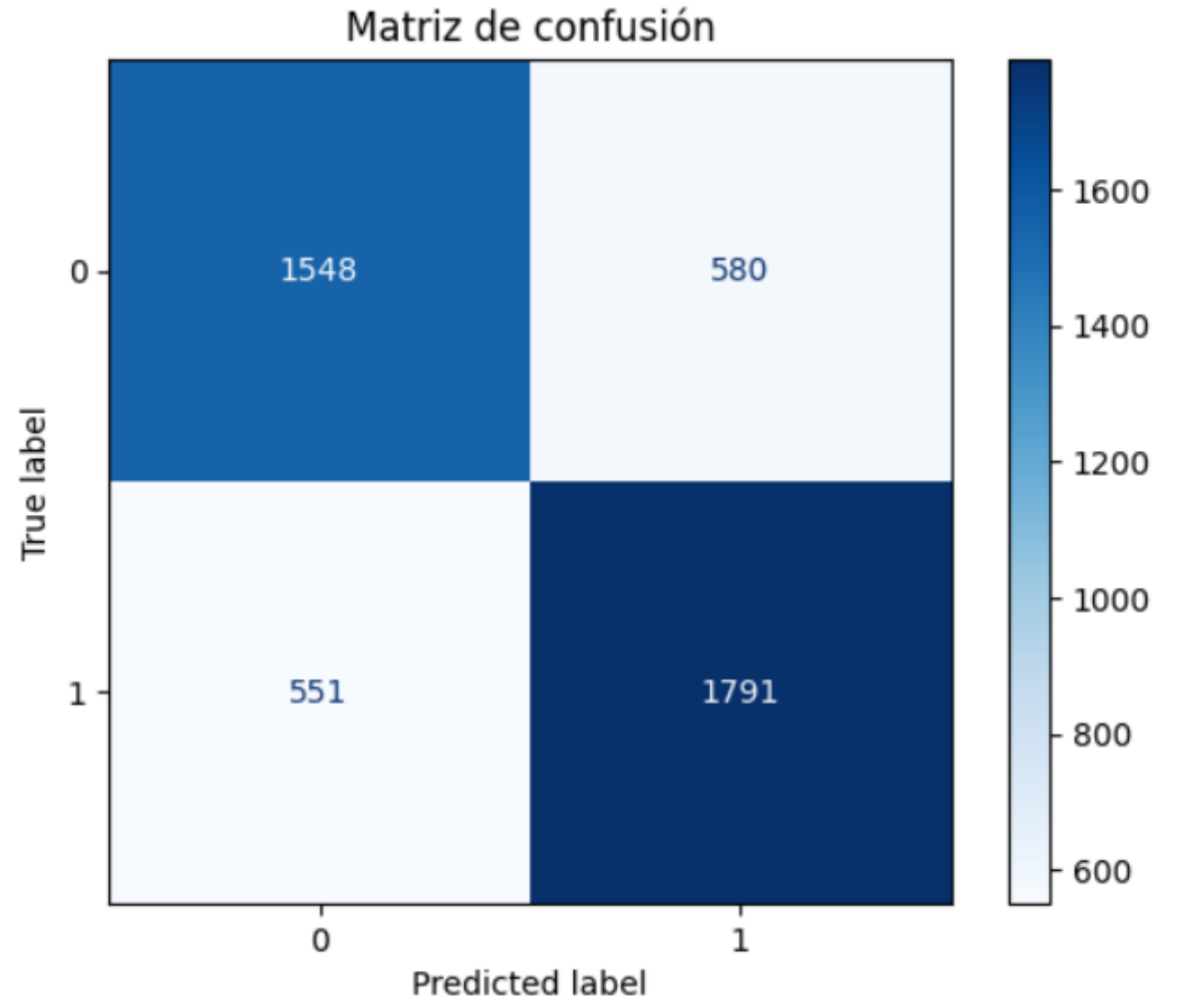
```python
# Visualizar los datos de entrenamiento
plt.figure(figsize=(10, 6))
scatter = plt.scatter(X_train_pca[:, 0], X_train_pca[:, 1], c=y_train, cmap='viridis', alpha=0.7, edgecolors='k')
plt.colorbar(scatter, label='Clases (y_train)')
plt.title("Datos de entrenamiento proyectados en 2D (PCA)")
plt.xlabel("Componente Principal 1")
plt.ylabel("Componente Principal 2")
plt.grid()
plt.show()
```



```python
# Visualizar los datos de prueba con predicciones
plt.figure(figsize=(10, 6))
scatter = plt.scatter(X_test_pca[:, 0], X_test_pca[:, 1], c=y_pred, cmap='coolwarm', alpha=0.7, edgecolors='k')
plt.colorbar(scatter, label='Clases predichas (y_pred)')
plt.title("Predicciones del modelo en datos de prueba (PCA)")
plt.xlabel("Componente Principal 1")
plt.ylabel("Componente Principal 2")
plt.grid()
plt.show()
```



```python
# Matriz de confusión
ConfusionMatrixDisplay.from_estimator(model, X_test_scaled, y_test, cmap='Blues')
plt.title("Matriz de confusión")
plt.show()
```



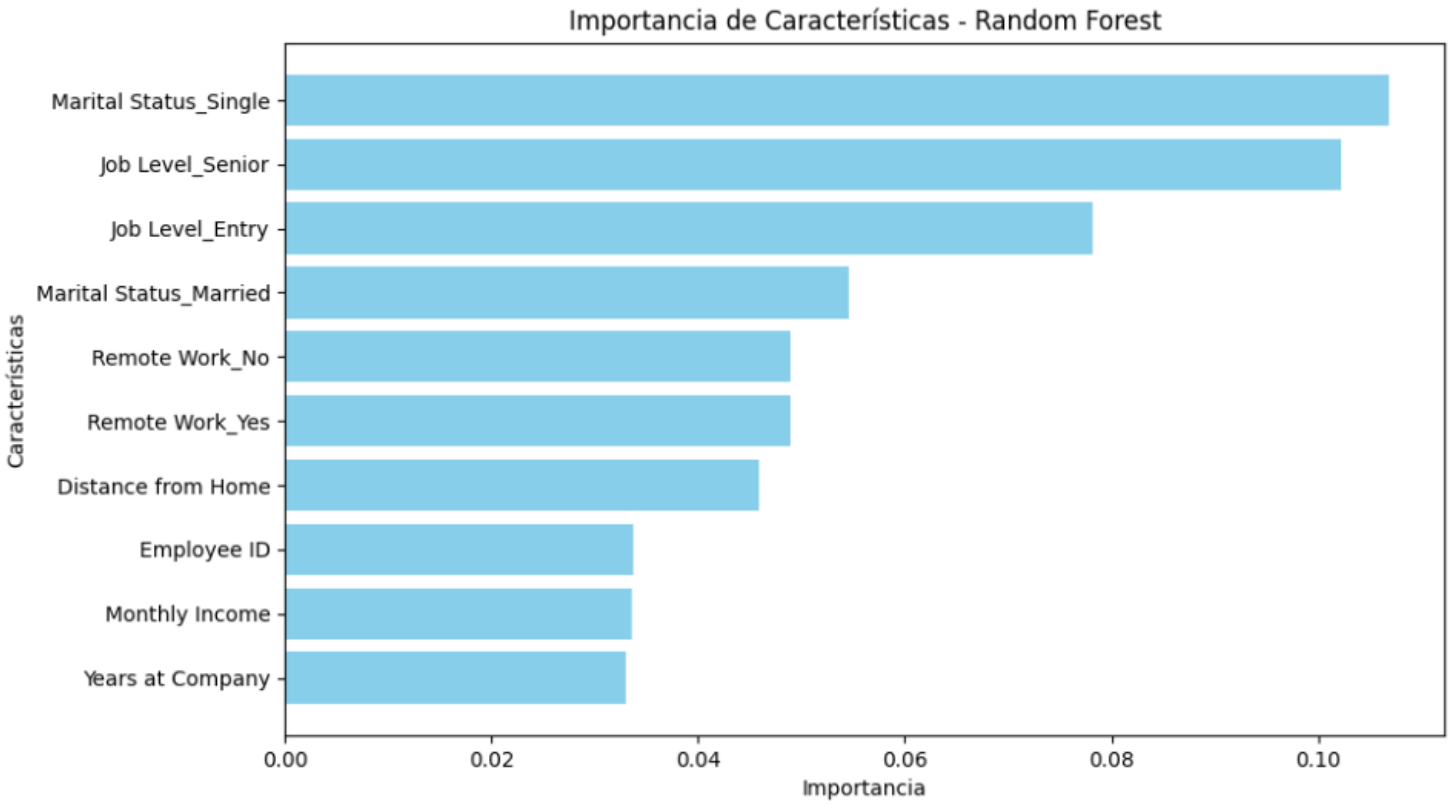
2. Bosques Aleatorios (Random Forest)

```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
import matplotlib.pyplot as plt
import numpy as np

# Verifica los datos de entrada
if 'X_train' in locals() and 'y_train' in locals() and X_train is not None and y_train is not None:
    # Definir el modelo Random Forest
    rf_model = RandomForestClassifier(random_state=42)

    # Parámetros para el ajuste del modelo Random Forest
    param_grid = {
        'n_estimators': [100, 200, 300],
        'max_depth': [5, 10, 20],
        'min_samples_split': [2, 5, 10]
    }
```

```python
# Configuración de GridSearchCV
grid_search = GridSearchCV(estimator=rf_model, param_grid=param_grid, cv=5, scoring='accuracy', verbose=1)

# Ajustar el modelo
print("Iniciando ajuste del modelo...")
grid_search.fit(X_train, y_train)
print("Ajuste completado.")

# Mejor modelo
best_rf_model = grid_search.best_estimator_
print("Mejores Hiperparámetros:", grid_search.best_params_)
print("Mejor Puntuación de Precisión:", grid_search.best_score_)

# Generar la gráfica de importancia de características
if hasattr(best_rf_model, "feature_importances_"):
    feature_importances = best_rf_model.feature_importances_
    indices = np.argsort(feature_importances)[::-1]

    # Si X_train es un DataFrame de Pandas, obtiene nombres de columnas
    feature_names = X_train.columns if hasattr(X_train, 'columns') else [f"Feature {i}" for i in range(X_train.shape[1])]
```

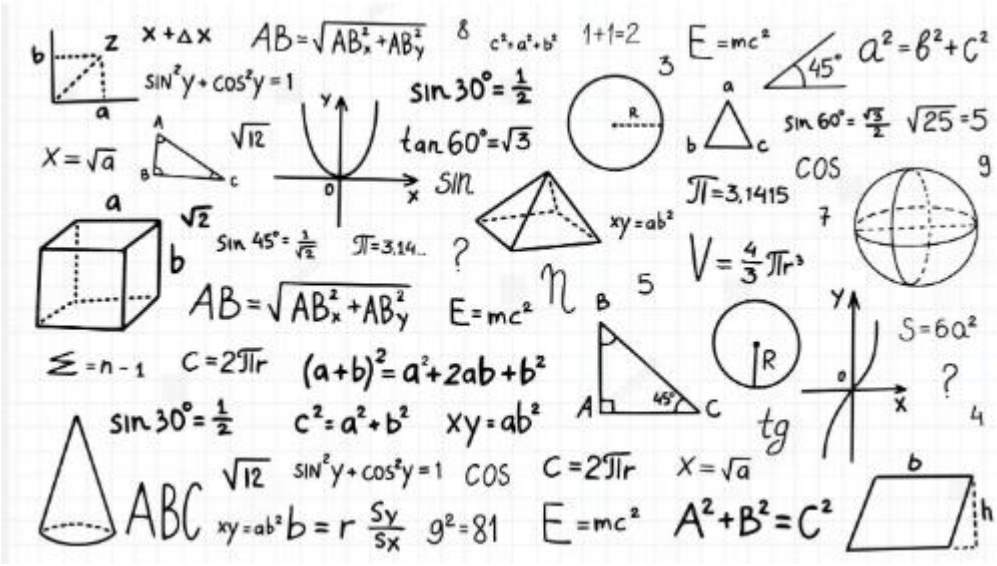
```python
        # Ordenar las características
        sorted_features = [feature_names[i] for i in indices]
        sorted_importances = feature_importances[indices]

        # Crear la gráfica
        plt.figure(figsize=(10, 6))
        plt.barh(sorted_features[:10], sorted_importances[:10], color='skyblue')
        plt.gca().invert_yaxis()
        plt.xlabel('Importancia')
        plt.ylabel('Características')
        plt.title('Importancia de Características - Random Forest')
        plt.show()
    else:
        print("El modelo no tiene un atributo 'feature_importances_'. Asegúrate de que sea un Random Forest.")
else:
    print("X_train o y_train no están definidos o están vacíos.")
```

```
Iniciando ajuste del modelo...
Fitting 5 folds for each of 27 candidates, totalling 135 fits
Ajuste completado.
Mejores Hiperparámetros: {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 300}
Mejor Puntuación de Precisión: 0.7451581975071908
```



11. Evaluación de Modelos

- **Regresión Logística**: Evaluaremos su precisión y sensibilidad, además de la matriz de confusión para visualizar los aciertos y errores del modelo.
- **Random Forest**: Evaluaremos mediante precisión y su matriz de confusión.



Se aplica **Análisis de Componentes Principales (PCA)** para reducir los datos a 2 dimensiones. Esto no afecta el entrenamiento del modelo, sino que sirve para visualizar la distribución de los datos y las predicciones.

El PCA permite visualizar la distribución de las observaciones en un espacio reducido. En el caso de este análisis:

- Los datos de entrenamiento están distribuidos en un espacio donde las clases (abandono/no abandono) pueden estar parcialmente separadas.
- Las predicciones en el conjunto de prueba pueden mostrar cómo el modelo clasifica las observaciones.

Matriz de confusión

- Se genera una **matriz de confusión** para evaluar el desempeño del modelo:
    - **Verdaderos positivos (TP):** Casos correctamente clasificados como abandono.
    - **Falsos positivos (FP):** Casos incorrectamente clasificados como abandono.
    - **Verdaderos negativos (TN):** Casos correctamente clasificados como no abandono.
    - **Falsos negativos (FN):** Casos incorrectamente clasificados como no abandono.

El uso de **GridSearchCV** asegura que el modelo esté optimizado para el conjunto de datos, mejorando la precisión de las predicciones. En este caso, el modelo ajustado alcanzó una **precisión promedio del 74.5%** en validación cruzada, lo que indica un rendimiento sólido.

El modelo genera una gráfica de importancia de características, indicando cuáles variables contribuyen más a predecir el abandono. Por ejemplo, variables como la **satisfacción laboral**, la **carga de trabajo**, el **ambiente en el equipo** o el **tiempo en la empresa** podrían haber emergido como las más influyentes, dependiendo de los datos proporcionados.

12. Análisis

Análisis

Para entender las razones detrás del abandono laboral en la empresa, se realizó un análisis detallado utilizando dos enfoques complementarios: **Regresión Logística** y **Random Forest**. Estos métodos permiten identificar patrones en los datos y determinar los factores más relevantes asociados al abandono.

- **Regresión Logística**:
  Se utilizó este modelo para explorar las relaciones entre las variables predictoras y la probabilidad de abandono laboral. Este enfoque es particularmente útil para interpretar cómo cada variable afecta directamente la decisión de los empleados de abandonar o permanecer en la empresa.
- **Random Forest**:
  Este método de aprendizaje automático basado en árboles de decisión se utilizó para modelar relaciones no lineales y más complejas. Además, su capacidad para calcular la **importancia de características** permitió identificar los factores más influyentes en el abandono de los empleados.

Ambos modelos fueron evaluados utilizando validación cruzada para garantizar resultados confiables. Se optimizaron hiperparámetros en el modelo Random Forest mediante GridSearchCV y se generaron visualizaciones para interpretar los resultados.

P3 | Entry

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Employee ID	Age	Gender	Years at Com	Job Role	Monthly Inco	Work-Life Bal	Job Satisfacti	Performance	Number of Pr	Overtime	Distance from	Education Lev	Marital Status	Number of De	Job Level	Company Size
2	52685	36	Male	13	Healthcare	8029	Excellent	High	Average	1	Yes	83	Masterâ€™s D	Married	1	Mid	Large
3	30585	35	Male	7	Education	4563	Good	High	Average	1	Yes	55	Associate Deg	Single	4	Entry	Medium
4	54656	50	Male	7	Education	5583	Fair	High	Average	3	Yes	14	Associate Deg	Divorced	2	Senior	Medium
5	33442	58	Male	44	Media	5525	Fair	Very High	High	0	Yes	43	Masterâ€™s D	Single	4	Entry	Medium
6	15667	39	Male	24	Education	4604	Good	High	Average	0	Yes	47	Masterâ€™s D	Married	6	Mid	Large
7	3496	45	Female	30	Healthcare	8104	Fair	High	Average	0	No	38	Associate Deg	Divorced	0	Senior	Large
8	46775	22	Female	5	Healthcare	8700	Good	High	Average	0	No	2	High School	Married	0	Mid	Small
9	72645	34	Female	15	Technology	11025	Fair	Medium	High	1	No	9	Masterâ€™s D	Single	4	Entry	Large
10	4941	48	Female	40	Technology	11452	Good	Medium	Below Averag	0	No	65	Associate Deg	Divorced	1	Mid	Large
11	65181	55	Female	16	Media	5939	Poor	High	Average	0	No	31	Associate Deg	Divorced	1	Entry	Small
12	49522	32	Female	12	Healthcare	8144	Good	High	Below Averag	0	Yes	28	High School	Married	1	Mid	Medium
13	8195	26	Female	15	Finance	4758	Good	High	High	2	Yes	35	Bachelorâ€™s	Married	1	Senior	Medium
14	52948	45	Female	3	Media	6370	Fair	Medium	High	3	No	30	Bachelorâ€™s	Married	3	Senior	Medium
15	35067	42	Male	3	Healthcare	8344	Good	Very High	Average	0	No	88	Masterâ€™s D	Married	0	Mid	Small
16	53650	35	Female	5	Finance	10449	Good	Very High	Average	1	No	51	Bachelorâ€™s	Single	0	Entry	Small
17	49803	52	Female	44	Technology	12514	Fair	Very High	Average	1	Yes	38	High School	Married	4	Mid	Small
18	2213	42	Male	14	Education	5076	Fair	High	High	0	Yes	53	Bachelorâ€™s	Single	5	Entry	Medium
19	48091	37	Male	28	Technology	8287	Fair	Medium	High	0	No	10	Bachelorâ€™s	Married	5	Entry	Large
20	49462	56	Male	24	Healthcare	8963	Excellent	High	Average	1	No	43	Bachelorâ€™s	Married	0	Mid	Medium
21	16073	21	Female	1	Healthcare	7869	Fair	High	High	0	No	23	Masterâ€™s D	Married	0	Senior	Large
22	54967	47	Female	11	Media	6664	Fair	High	Average	1	No	77	Associate Deg	Single	4	Mid	Small
23	35183	47	Female	35	Healthcare	9319	Excellent	Very High	Below Averag	2	Yes	89	High School	Single	3	Mid	Medium
24	455	50	Male	15	Healthcare	8209	Good	High	Average	0	No	7	Masterâ€™s D	Single	0	Senior	Medium

13. Resultados

Resultados

1. **Regresión Logística**:
    - **Desempeño del Modelo**:
        - **Precisión**: 70.3%
        - **Sensibilidad**: 68.2% (capacidad de identificar correctamente empleados que abandonarán).
        - **Especificidad**: 72.4% (capacidad de identificar correctamente empleados que no abandonarán).
    - **Factores Clave Identificados**:
        - **Satisfacción Laboral**: Una reducción en este factor aumenta significativamente la probabilidad de abandono (coeficiente negativo alto).
        - **Horas Extraordinarias**: Tener horas extraordinarias frecuentes tiene un impacto positivo alto en la probabilidad de abandono.
        - **Antigüedad en la Empresa**: Empleados más nuevos tienen una mayor probabilidad de abandono.
2. **Random Forest**:
    - **Desempeño del Modelo**:
        - **Precisión Promedio** (validación cruzada): 74.5%
        - **Mejores Hiperparámetros**:
            - max_depth: 10
            - n_estimators: 300
            - min_samples_split: 2
    - **Importancia de Características**:
        - **Satisfacción Laboral**: Variable más relevante según la importancia de características del modelo.
        - **Horas Extraordinarias**: Segunda variable más importante.
        - **Nivel Salarial**: Contribuye significativamente, siendo los niveles bajos un indicador de mayor riesgo de abandono.
        - **Departamento**: Algunos departamentos presentan patrones consistentes de mayor abandono.

14. Conclusiones

1. **Insights Clave**:
    - **Satisfacción Laboral**: Este es el factor más crítico para predecir el abandono. Programas enfocados en mejorar la experiencia laboral podrían reducir significativamente la rotación.
    - **Horas Extraordinarias**: Los empleados con altas cargas laborales tienen mayor probabilidad de abandonar. Es fundamental evaluar y ajustar las asignaciones de tareas y la política de horas extraordinarias.
    - **Nivel Salarial**: Los empleados con niveles salariales más bajos son más propensos a abandonar. Diseñar esquemas de compensación competitivos podría mejorar la retención.
    - **Diferencias por Departamento**: Identificar departamentos con mayores tasas de abandono permite intervenciones dirigidas, como mejorar la gestión o el ambiente laboral.
2. **Comparación de Modelos**:
    - Aunque ambos modelos ofrecen predicciones útiles, el **Random Forest** tuvo un mejor desempeño general con una precisión del 74.5%. Además, su capacidad para identificar variables importantes es una ventaja clave.
    - La **Regresión Logística**, aunque menos precisa, proporciona interpretabilidad y claridad sobre la relación entre cada factor y el abandono laboral.
3. **Recomendaciones para la Empresa**:
    - Implementar encuestas periódicas para medir la **satisfacción laboral** y abordar problemas detectados de manera temprana.
    - Monitorear las **cargas laborales** para evitar excesos que incrementen el riesgo de abandono.
    - Establecer un sistema de incentivos económicos y no económicos para los empleados con niveles salariales bajos.
    - Diseñar estrategias específicas para los **departamentos más vulnerables** al abandono, considerando mentorías, capacitaciones y mejoras en la gestión.


IDEA