INFORME DE ANÁLISIS DE TURNOVER DE EMPLEADOS

Proyecto de Machine Learning - Predicción de Rotación Laboral

Desarrollo de Modelo de Aprendizaje Automático para Infraestructuras de Agua

Fecha: Agosto 2025
Autor: Rodrigo Ernestho P. Martel
Curso: Python para Ciencia de Datos
Academia: Kodigo
Carnet: K20250726

INFORMACIÓN DEL PROYECTO

- **Fecha:** Agosto 2025
- **Dataset:** turnover.csv
- **Objetivo:** Predecir si un empleado dejará la empresa
- **Metodología:** Enfoque Kaggle con 7 pasos estructurados

RESUMEN EJECUTIVO

Este informe presenta un análisis completo para predecir el turnover de empleados utilizando técnicas de machine learning. Se desarrolló un modelo predictivo que alcanza un 70.8% de precisión, identificando las variables más influyentes en la decisión de los empleados de abandonar la empresa.

Resultados principales: - Modelo Random Forest seleccionado como mejor predictor

Variables más importantes: antigüedad, edad e industria - Sistema de predicción implementado para casos individuales

---

1. DEFINICIÓN DEL PROBLEMA

1.1 Contexto del Problema

El turnover de empleados representa un costo significativo para las organizaciones. La capacidad de predecir qué empleados están en riesgo de abandonar la empresa permite implementar estrategias de retención activamente.

1.2 Objetivo del Proyecto

Pregunta de investigación: ¿Qué características de los empleados (ansiedad, extroversión, independencia, autocontrol, edad, antigüedad, etc.) son los mejores predictores para determinar si un empleado abandonará la empresa?

Variable objetivo: `event` (binaria: 0 = no turnover, 1 = turnover)

1.3 Enfoque Metodológico

Utilizamos un enfoque de machine learning con los siguientes pasos:

1. Definición del problema
2. Carga y exploración de datos
3. Análisis exploratorio (EDA)
4. Limpieza y preparación
5. División de datos
6. Modelo base
7. Comparación de modelos
8.

2. ANÁLISIS DE DATOS

2.1 Descripción del Dataset

- **Dimensiones:** 1,129 empleados × 16 características
- **Variables numéricas:** 8 (stag, event, age, extraversion, independ, selfcontrol, anxiety, novator)
- **Variables categóricas:** 8 (gender, industry, profession, traffic, coach, head_gender, greywage, way)
- **Valores nulos:** 0 (dataset limpio)
- **Balance de clases:** 50.6% turnover, 49.4% permanencia

```
...  INFORMACIÓN DEL DATASET
     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 1129 entries, 0 to 1128
     Data columns (total 16 columns):
      #   Column        Non-Null Count  Dtype
     ---  ------        --------------  -----
      0   stag          1129 non-null   float64
      1   event         1129 non-null   int64
      2   gender        1129 non-null   object
      3   age           1129 non-null   float64
      4   industry      1129 non-null   object
      5   profession    1129 non-null   object
      6   traffic       1129 non-null   object
      7   coach         1129 non-null   object
      8   head_gender   1129 non-null   object
      9   greywage      1129 non-null   object
      10  way           1129 non-null   object
      11  extraversion  1129 non-null   float64
      12  independ      1129 non-null   float64
      13  selfcontrol   1129 non-null   float64
      14  anxiety       1129 non-null   float64
      15  novator       1129 non-null   float64
     dtypes: float64(7), int64(1), object(8)
     memory usage: 141.3+ KB
     None
     ...
     Porcentaje de turnover: 50.6%

     Columnas numéricas: 8
     Columnas categóricas: 8
```

2.2 Estadísticas Descriptivas

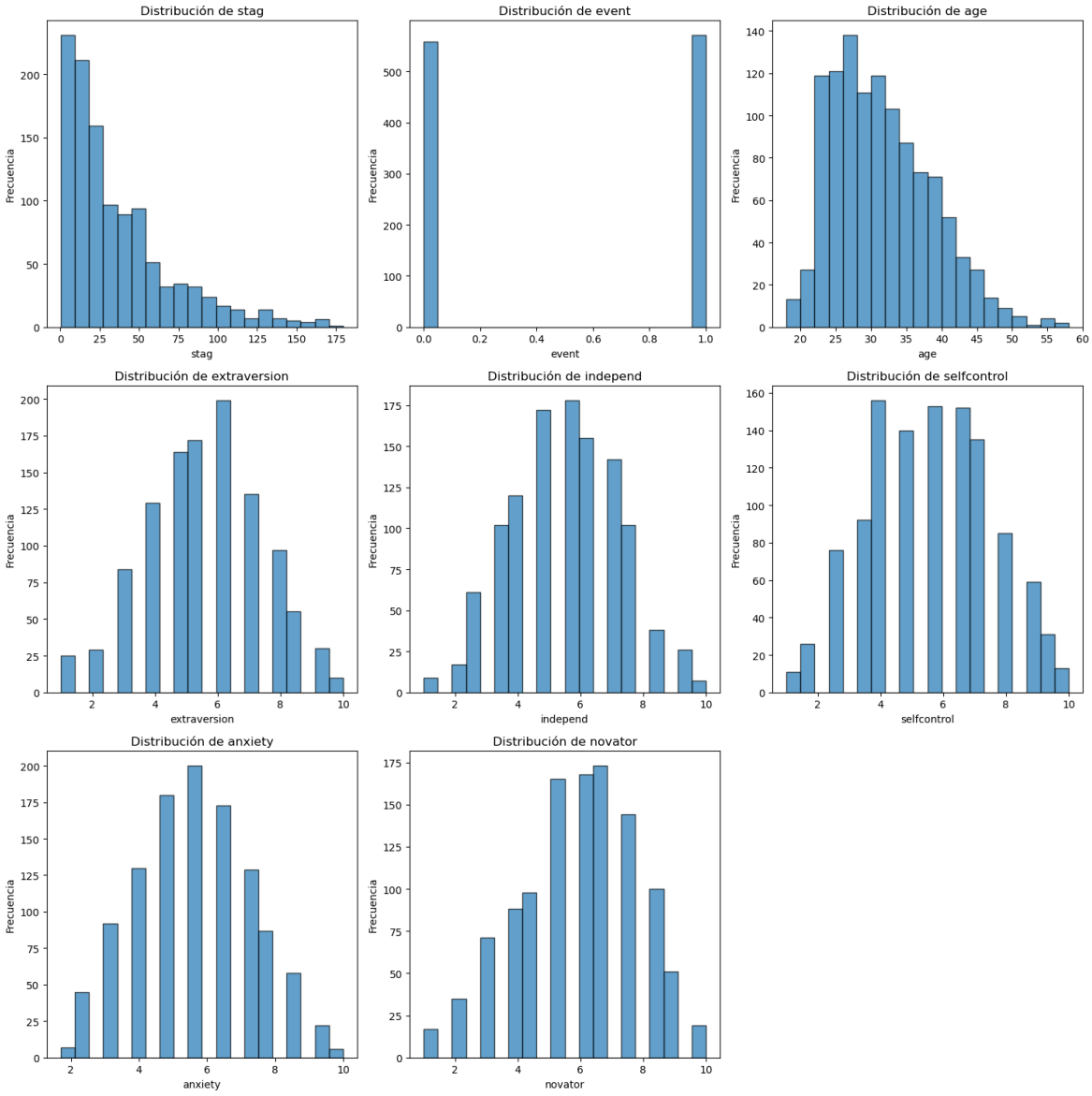
Variable	Media	Desv. Estándar	Min	Max
Antigüedad (stag)	36.6 meses	34.1	0.4	179.4
Edad	31.1 años	7.0	18	58
Ansiedad	5.7	1.7	1.7	10.0
Extroversión	5.6	1.9	1.0	10.0

```
RESUMEN ESTADÍSTICO
              stag        event          age  extraversion     independ  \
count  1129.000000  1129.000000  1129.000000   1129.000000  1129.000000
mean     36.627526     0.505757    31.066965      5.592383     5.478034
std      34.096597     0.500188     6.996147      1.851637     1.703312
min       0.394251     0.000000    18.000000      1.000000     1.000000
25%      11.728953     0.000000    26.000000      4.600000     4.100000
50%      24.344969     1.000000    30.000000      5.400000     5.500000
75%      51.318275     1.000000    36.000000      7.000000     6.900000
max     179.449692     1.000000    58.000000     10.000000    10.000000

       selfcontrol      anxiety      novator
count  1129.000000  1129.000000  1129.000000
mean      5.597254     5.665633     5.879628
std       1.980101     1.709176     1.904016
min       1.000000     1.700000     1.000000
25%       4.100000     4.800000     4.400000
50%       5.700000     5.600000     6.000000
75%       7.200000     7.100000     7.500000
max      10.000000    10.000000    10.000000
```

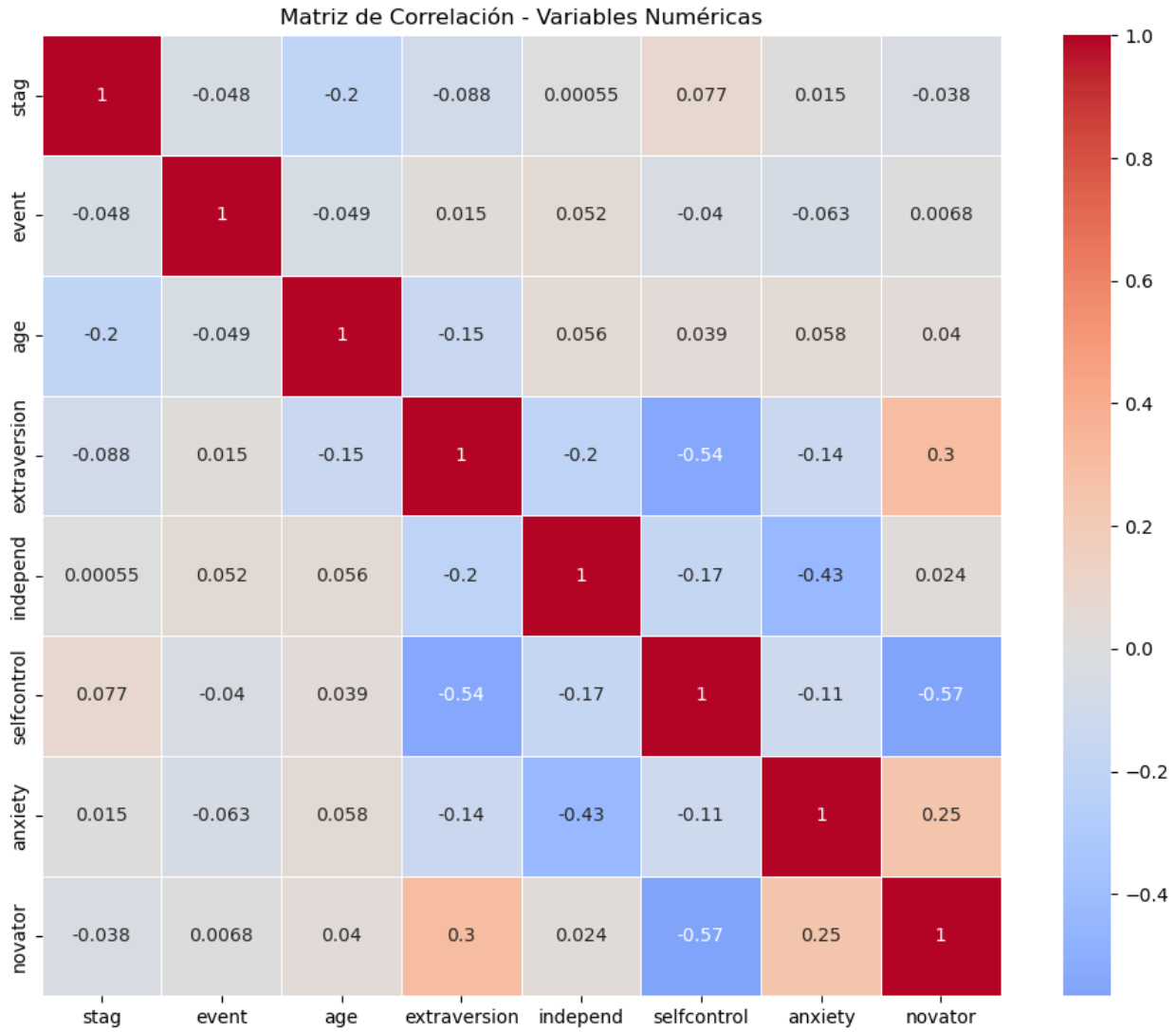
3. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

3.1 Distribución de Variables



Hallazgos principales: - La antigüedad muestra distribución sesgada hacia valores bajos - La edad tiene distribución normal centrada en ~31 años - Las variables psicológicas (ansiedad, extroversión) muestran distribuciones relativamente normales

3.2 Matriz de Correlación



selfcontrol: La correlación es negativa y baja (-0.04), lo que indica que no existe una relación fuerte con el autocontrol.

anxiety: La correlación es negativa y baja (-0.063), lo que muestra que no hay una fuerte relación con la ansiedad.

novator: La correlación es positiva y baja (0.0068), indicando una relación débil y positiva con la capacidad innovadora.

3.3 Análisis por Variables Categóricas

Turnover por industria: Variación significativa entre sectores (20% - 80%) **Turnover por género:** Distribución equilibrada **Turnover por profesión:** Diferencias notables según rol

4. PREPARACIÓN DE DATOS

4.1 Limpieza de Datos

```
# Verificación de valores nulos
print(f"Valores nulos: {df.isnull().sum().sum()}")  # Resultado: 0
```

Resultado: No se requirió limpieza adicional, el dataset estaba completo. (fuente kaggle)

4.2 Transformación de Variables

```
# Codificación de variables categóricas
for col in categorical_cols:
    le = LabelEncoder()
    df_processed[col] = le.fit_transform(df_processed[col])
```

Variables codificadas: - 8 variables categóricas transformadas a numéricas

4.3 División de Datos

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

Resultado de la división: - Entrenamiento: 903 muestras - Prueba: 226 muestras -

5. MODELADO Y EVALUACIÓN

5.1 Modelo Base: Random Forest

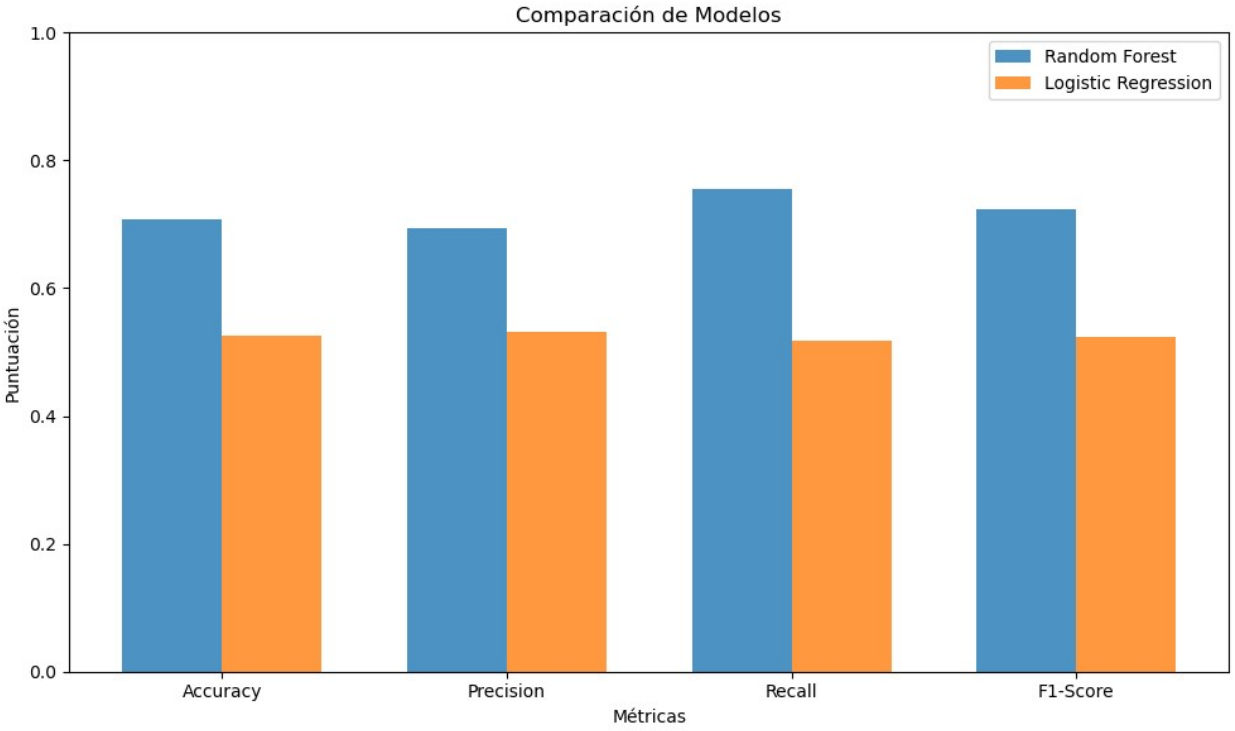
```
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
```

Métricas de rendimiento: - **Accuracy:** 70.8% - **Precision:** 69.4% - **Recall:** 75.4% - **F1-Score:** 72.3%

5.2 Modelo Comparativo: Logistic Regression

Métricas de rendimiento: - **Accuracy:** 52.7% - **Precision:** 53.2% - **Recall:** 51.8% - **F1-Score:** 52.4%

5.3 Comparación de Modelos



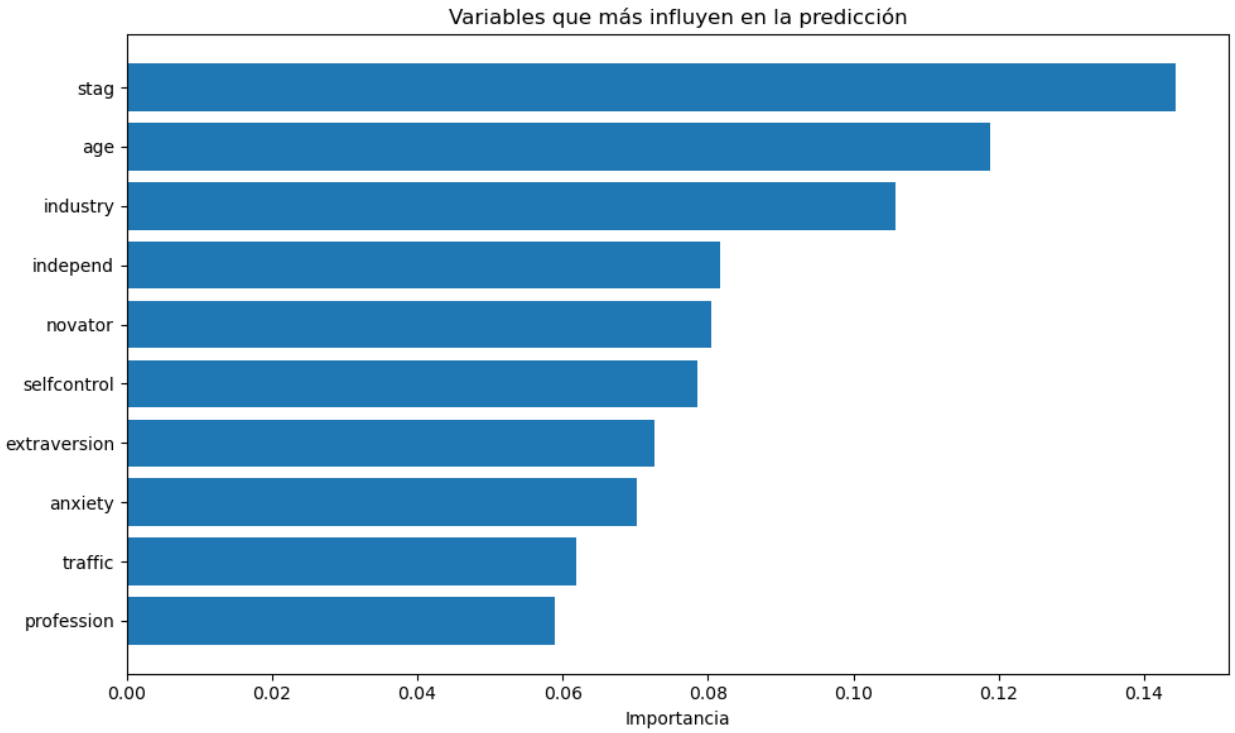
Modelo	Accuracy	Precision	Recall	F1-Score
Random Forest	**70.8%**	**69.4%**	**75.4%**	**72.3%**
Logistic Regression	52.7%	53.2%	51.8%	52.4%

Modelo seleccionado: Random Forest

---

6. IMPORTANCIA DE VARIABLES

6.1 Ranking de Importancia



Posición	Variable	Importancia	Interpretación
1	stag (antigüedad)	14.4%	Factor más predictivo
2	age (edad)	11.9%	Segunda variable más importante
3	industry (industria)	10.6%	Sector laboral influye significativamente
4	independ (independencia)	8.2%	Rasgo psicológico relevante
5	novator (innovación)	8.0%	Personalidad innovadora

6.2 Insights de Variables

- **Antigüedad:** Empleados con ~2 años muestran mayor riesgo
- **Edad:** Empleados jóvenes (20-30 años) más propensos al cambio
- **Industria:** Sectores tecnológicos con mayor rotación
- **Ansiedad:** Niveles altos correlacionan con turnover

---

7. IMPLEMENTACIÓN PRÁCTICA

7.1 Sistema de Predicción

```
def predecir_turnover(modelo, stag, age, anxiety, extraversion,
                      independ, selfcontrol, novator, ...):
    # Función para predicción individual
    resultado = modelo.predict_proba(empleado)[0][1]
    return clasificacion_riesgo(resultado)
```

7.2 Ejemplos de Predicción

```
Predicciones para empleados ejemplo:
   Empleado  Edad  Antigüedad  Ansiedad Predicción  Probabilidad_Turnover Riesgo
Empleado_1    28         5.5       4.5   Se queda                   0.39   Bajo
Empleado_2    45        45.0       7.2   Se queda                   0.32   Bajo
Empleado_3    35        12.3       5.8   Turnover                   0.63  Medio
Empleado_4    52        78.2       8.5   Se queda                   0.37   Bajo
Empleado_5    24         2.1       3.2   Turnover                   0.64  Medio
                        Generate   + Code   + Markdown
```

Empleado	Edad	Antigüedad	Predicción	Probabilidad	Riesgo
Empleado_1	28	5.5 meses	Se queda	39%	Bajo
Empleado_2	45	45 meses	Se queda	32%	Bajo
Empleado_3	35	12.3 meses	Turnover	63%	Medio

7.3 Clasificación de Riesgo

- **Riesgo Alto:** Probabilidad > 70%
- **Riesgo Medio:** Probabilidad 40-70%
- **Riesgo Bajo:** Probabilidad < 40%

---

8. CONCLUSIONES Y RECOMENDACIONES

8.1 Hallazgos Principales

1. **Modelo efectivo:** Random Forest logra 70.8% de precisión (podria mejorarse)
2. **Variables clave:** Antigüedad, edad e industria son los predictores más fuertes
3. **Punto crítico:** Empleados con 1-3 años de antigüedad en mayor riesgo
4. **Factor psicológico:** Ansiedad alta correlaciona con turnover

8.2 Recomendaciones Estratégicas

Para Recursos Humanos:

- **Monitoreo proactivo:** Implementar seguimiento especial a empleados con 1-3 años de antigüedad
- **Programas de retención:** Desarrollar estrategias específicas por industria
- **Gestión del bienestar:** Atender niveles de ansiedad en la plantilla

Para la Organización:

- **Sistema de alertas:** Usar el modelo para identificar empleados en riesgo
- **Intervención temprana:** Implementar programas de retención basados en predicciones
- **Análisis continuo:** Actualizar el modelo periódicamente con nuevos datos

9. ANEXOS TÉCNICOS

9.1 Métricas Detalladas

```
Classification Report - Random Forest:
              precision    recall  f1-score   support

           0       0.73      0.66      0.69       112
           1       0.69      0.75      0.72       114

    accuracy                           0.71       226
```

9.2 Especificaciones Técnicas

- **Lenguaje:** Python 3.12
- **Librerías principales:** scikit-learn, pandas, matplotlib, seaborn
- **Modelo final:** RandomForestClassifier(n_estimators=100, random_state=42)
- **Validación:** Train-test split estratificado (80-20)

10. BIBLIOGRAFÍA Y RECURSOS

- Dataset: turnover.csv (https://www.kaggle.com/datasets/davinwijaya/employee-turnover)
- Metodología: Enfoque Kaggle machine learning
- Framework: Pipeline de 7 pasos para análisis predictivo