ITBA
Instituto Tecnológicc
de Buenos Aires

CANDELA PALOMEQUE

ANÁLISIS PREDICTIVO

AGENDA

1. Introducción
2. EDA
3. Modelo Predictivo
4. Conclusiones e Insights

Introducción

Attrition

¿Por qué es un problema?

El *attrition* refiere al retiro o renuncia voluntaria de empleados de una organización, que puede explicarse por múltiples razones.

Esto implica muchos desafíos para las organizaciones:

- Se llevan conocimientos.
- Costos de liquidación.
- Costos de nueva capacitación.
- Procesos de contratación.

Hipótesis

- Existe una tendencia entre los empleados más jóvenes y aquellos que no tienen un plan de carrera establecido a abandonar la organización con mayor frecuencia. Los empleados más jóvenes, en general, pueden ser más propensos a buscar nuevas oportunidades y explorar diferentes opciones profesionales, lo que puede influir en su decisión de dejar la empresa.
- Aquellos empleados cuyos salarios mensuales se sitúan por debajo de los 4 mil dólares, tienen una mayor probabilidad de abandonar la organización. Los empleados que perciben salarios más bajos pueden sentirse insatisfechos con su nivel de remuneración, lo que puede afectar su motivación y compromiso con la empresa.

Problema

El attrition en recursos humanos genera:

- ➔ Altos costos económicos (perdida en reclutamiento, selección y capacitación)
- ➔ Afecta la productividad, estabilidad y clima laboral de la empresa al requerir tiempo y esfuerzo adicional para capacitar a nuevos empleados, generando incertidumbre y desmotivación entre el personal restante
- ➔ La falta de estabilidad dificulta la formación de relaciones sólidas y de confianza, esenciales para el trabajo en equipo y la colaboración efectiva.

Objetivo

Construir un modelo predictivo para que las empresas evalúen el riesgo de *attrition* de sus empleados. Esto se realizará aplicando técnicas supervisadas de machine learning para construir un modelo robusto y confiable.

¿Para que?

1. Evaluar el impacto de incentivos (económicos o de promoción) en la satisfacción de la fuerza laboral.
2. Evaluar el impacto de incentivos en el attrition.
3. Caracterizar grupos de empleados con alto o bajo riesgo de attrition.
4. Evaluar el riesgo de attrition para diferentes grupos de empleados.

Dataset

- ➔ Obtenido de Kaggle.
- ➔ 35 variables y 1.470 observaciones.
- ➔ Tiene variables que brindan información personal de los empleados y variables en relación a la organización.
- ➔ Se seleccionaron finalmente 19 variables

EDA

EDA

Análisis Exploratorio de Datos realizado sobre la base de datos de Attrition

1. Limpieza de datos
2. Análisis de variables

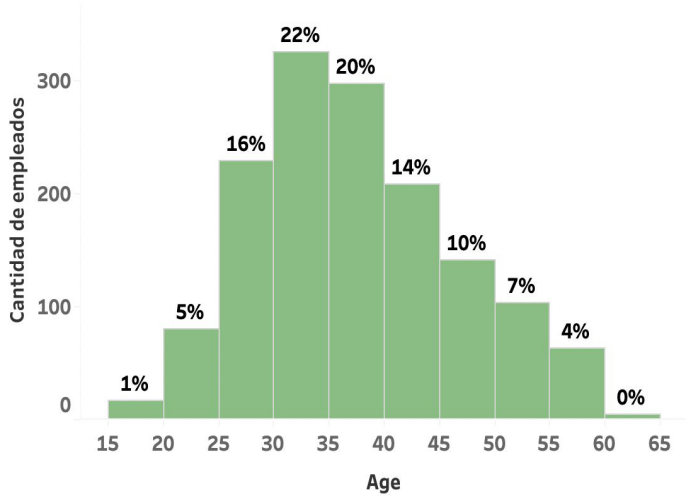
Limpieza de Datos

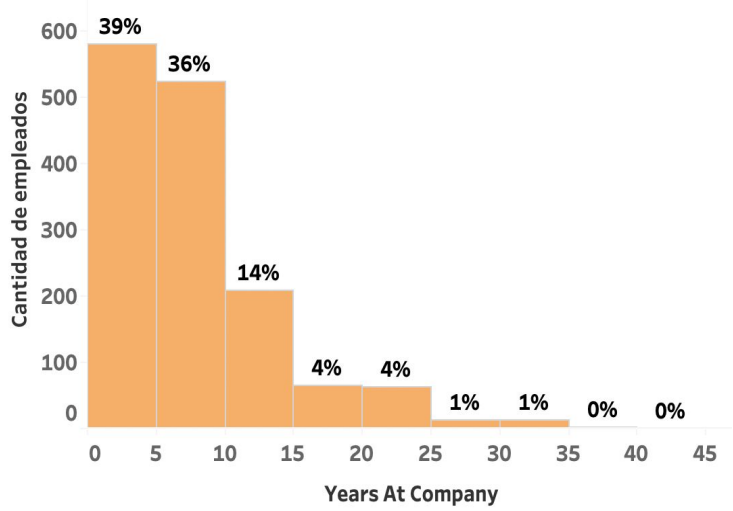
- No hay valores duplicados ni errores en la calidad de datos.
- No se encontraron missings ni outliers.
- No fue necesario realizar una limpieza de la base de datos.
- Posible tratamiento para missings: reemplazarlos por la moda o media, o eliminar dichos registros dependiendo el caso.
- Posible tratamiento para outliers: se pueden eliminar los registros si son erróneos o aplicar algún tipo de transformación a los mismos.

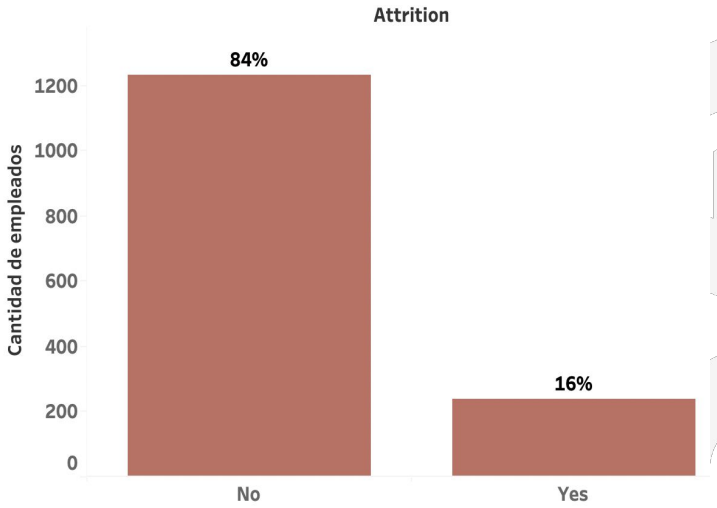
Análisis de Variables

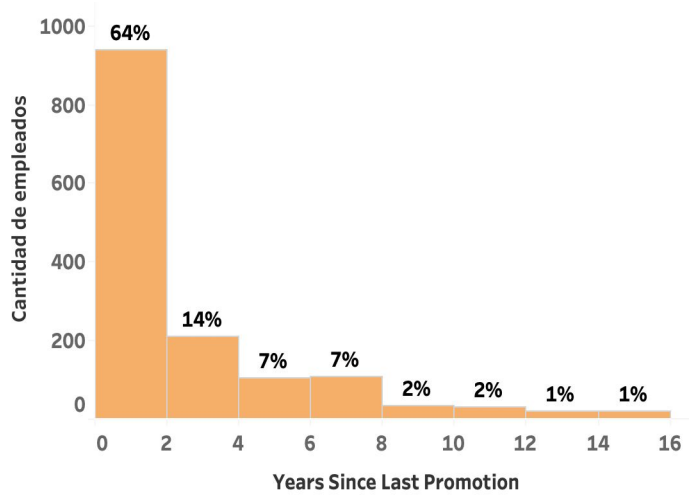
Se estudian las relaciones entre pares de variables. En este caso es importante considerar el impacto de las variables sobre el target de tipo categórica: *attrition*.

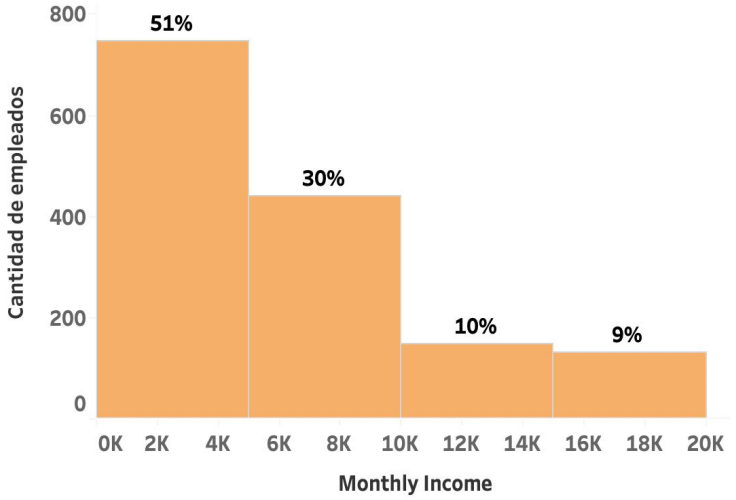
Distribución de las variables

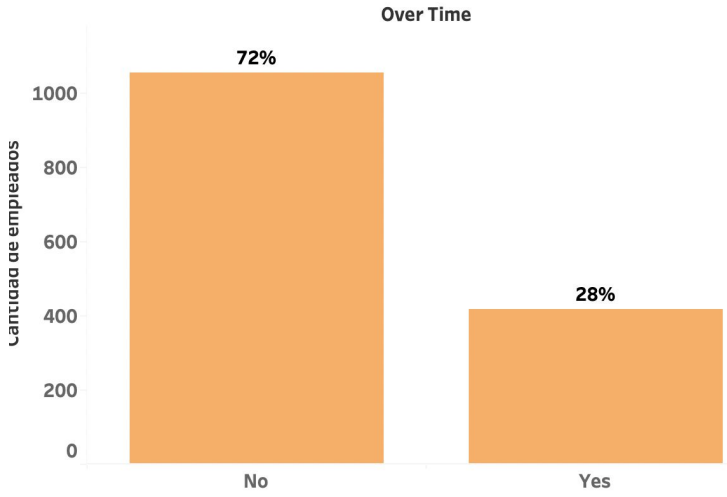




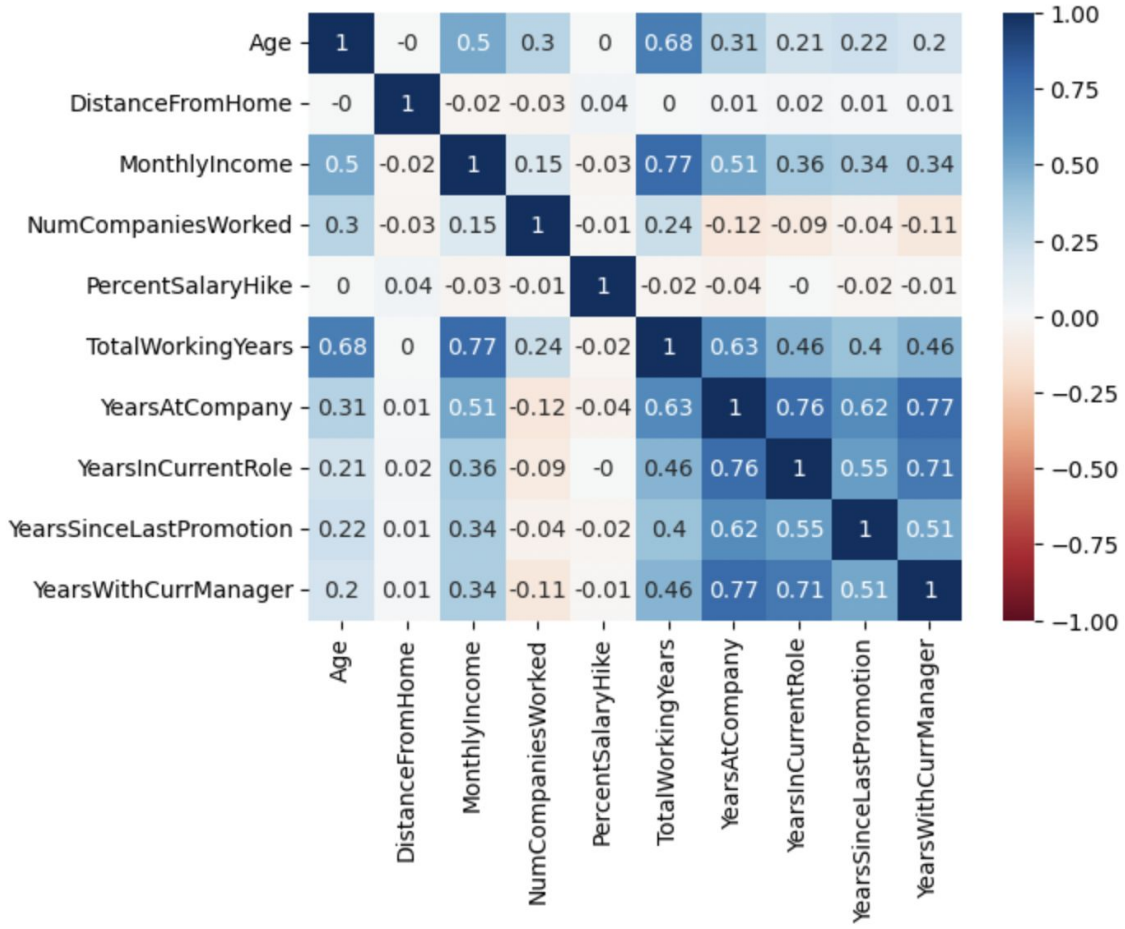








Análisis Bivariado



- ➔ *Age - TotalWorkingYears:* 0.68 de correlación.
- ➔ *MonthlyIncome - TotalWorkingYears:* 0.77 de correlación.
- ➔ *TotalWorkingYears - YearsAtCompany:* 0.68 de correlación.
- ➔ *YearsInCurrentRole - YearsSinceLastPromotion:* 0.55 de correlación.
- ➔ *YearsWithCurrManager - YearsInCurrentRole:* 0.71 de correlación.
- ➔ *YearsAtCompany - YearsSinceLastPromotion:* 0.62 de correlación.
- ➔ *YearsWithCurrManager - YearsAtCompany:* 0.77 de correlación.

Caracterización de los empleados que se van

Monthly Income Grupos

Attrition	0K - 2K	2K - 4K	4K - 6K	>=6K	Grand Total
No	1%	32%	27%	40%	100%
Yes	8%	50%	18%	24%	100%
Grand Total	2%	35%	25%	38%	100%

Age Grupos

Attrition	<=20	21-30	31-40	>40	Grand To..
No	1%	22%	43%	33%	100%
Yes	7%	35%	36%	22%	100%
Grand Total	2%	24%	42%	32%	100%

Years At Company Grupos

Attrition	0-2	3-5	6-10	>10	Grand Total
No	19%	30%	32%	18%	100%
Yes	43%	25%	23%	8%	100%
Grand Total	23%	30%	30%	17%	100%

Over Time

Attrition	No	Yes	Grand Total
No	77%	23%	100%
Yes	46%	54%	100%
Grand Total	72%	28%	100%

Years In Current Role Grupos

Attrition	0-2	3-5	6-10	>10	Grand Total
No	42%	20%	32%	6%	100%
Yes	64%	14%	20%	2%	100%
Grand Total	46%	19%	30%	5%	100%

Grupos con más riesgo de irse

Age Grupos

Attrition	<=20	21-30	31-40	>40	Grand Total
No	43%	77%	86%	89%	84%
Yes	57%	23%	14%	11%	16%
Grand Total	100%	100%	100%	100%	100%

Years At Company Grupos

Attrition	0-2	3-5	6-10	>10	Grand Total
No	70%	86%	88%	92%	84%
Yes	30%	14%	12%	8%	16%
Grand Total	100%	100%	100%	100%	100%

Business Travel

Attrition	Non-Travel	Travel_Rarely	Travel_Freque..	Grand Total
No	92%	85%	75%	84%
Yes	8%	15%	25%	16%
Grand Total	100%	100%	100%	100%

Total Working Years (bin)

Attrition	0	5	10	15	20	25	30	35	40	Grand Total
No	67%	83%	88%	89%	92%	95%	92%	100%		84%
Yes	33%	17%	12%	11%	8%	5%	8%		100%	16%
Grand Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

MODELO PREDICTIVO

Construcción del modelo predictivo

Random Forest y Regresión Logística

1. Selección de variables
2. Split en Train y Test
3. Construcción y evaluación

Preparación del dataset

1. Selección de Variables

```
#   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
0   Age                               1470 non-null   int64
1   MonthlyIncome                     1470 non-null   int64
2   PercentSalaryHike                 1470 non-null   int64
3   TotalWorkingYears                 1470 non-null   int64
4   YearsAtCompany                    1470 non-null   int64
5   YearsInCurrentRole                1470 non-null   int64
6   YearsWithCurrManager              1470 non-null   int64
7   BusinessTravel_Travel_Frequently  1470 non-null   uint8
8   OverTime_Yes                      1470 non-null   uint8
9   Attrition_Yes                     1470 non-null   uint8
```

→ **Dummies**

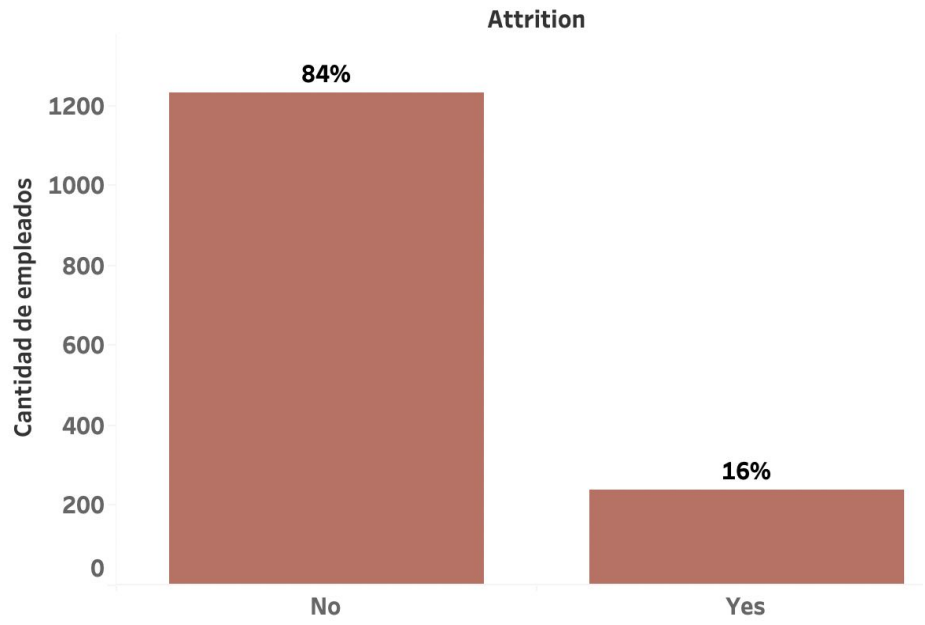
2. Split en Train y Test

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.2,
                                                    random_state=123)
print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

Entrenamiento con el 80% de los datos

Preparación del dataset

PROBLEMA! → dataset desbalanceado



Recordando la distribución de la variable target, observamos que las clases están muy desbalanceadas.

Esto genera problemas a la hora de entrenar modelos, que funcionan mejor cuando las clases tienen un número similar de observaciones.

SOLUCIONES:

- **Oversampling (SMOTE)**
- **Undersampling (riesgo de pérdida de información)**
- **Hiperparámetro de class_weights**

Construcción del modelo predictivo

RANDOM FOREST

Tuneo de hiperparámetros:

- **n_estimators = [100, 200, 300]**
- **max_depth = [None, 5, 10]**
- **min_samples_split = [2, 5, 10]**

class_weights = {0:1, 1:5}

MEJOR MODELO		
ACCURACY	SENSIBILIDAD	ESPECIFICIDAD
83%	50%	88%

REGRESIÓN LOGÍSTICA

Tuneo de hiperparámetros:

- **penalty = ['l1' , 'l2']**

class_weights = {0:1, 1:5}

MEJOR MODELO		
ACCURACY	SENSIBILIDAD	ESPECIFICIDAD
63%	74%	62%

Elección del modelo predictivo

REGRESIÓN LOGÍSTICA

MEJOR MODELO		
ACCURACY	SENSIBILIDAD	ESPECIFICIDAD
63%	74%	62%

Se prioriza mejorar la detección de las observaciones de la clase positiva en lugar de aumentar el porcentaje de aciertos a expensas de disminuir la sensibilidad. La sensibilidad es crucial para evitar falsos negativos, y al optar por el modelo de Regresión Logística se asegura una mayor capacidad para identificar correctamente los casos positivos.

Conclusiones e insights

- El análisis desarrollado permite caracterizar a los empleados que deciden dejar de trabajar en la organización.
- Además, se analizó el riesgo de attrition para dichas características de los empleados.
- Este análisis junto con el modelo de Regresión Logística seleccionado permitirán a la organización identificar los empleados con mayor riesgo de irse y a partir de eso elaborar planes de acción.
- Se recomienda a la empresa llevar adelante planes de acción como: Mejorar el ambiente laboral, implementar programas de retención, establecer programas de integración efectivos para nuevos empleados, entre otros.
- Realizar actualización de la base por lo menos 1 vez al año.

ITBA
Instituto Tecnológicc
de Buenos Aires

¡MUCHAS GRACIAS!

MÁS INFORMACIÓN > **www.itba.edu.ar**