



# IBM Data Science

Proyecto Final

Profesor: Jorge Kamlofsky

Tutor: Fernando Culell

# QUIÉNES SOMOS?

---

## Integrantes



Celeste Bertolez



Belén Torres



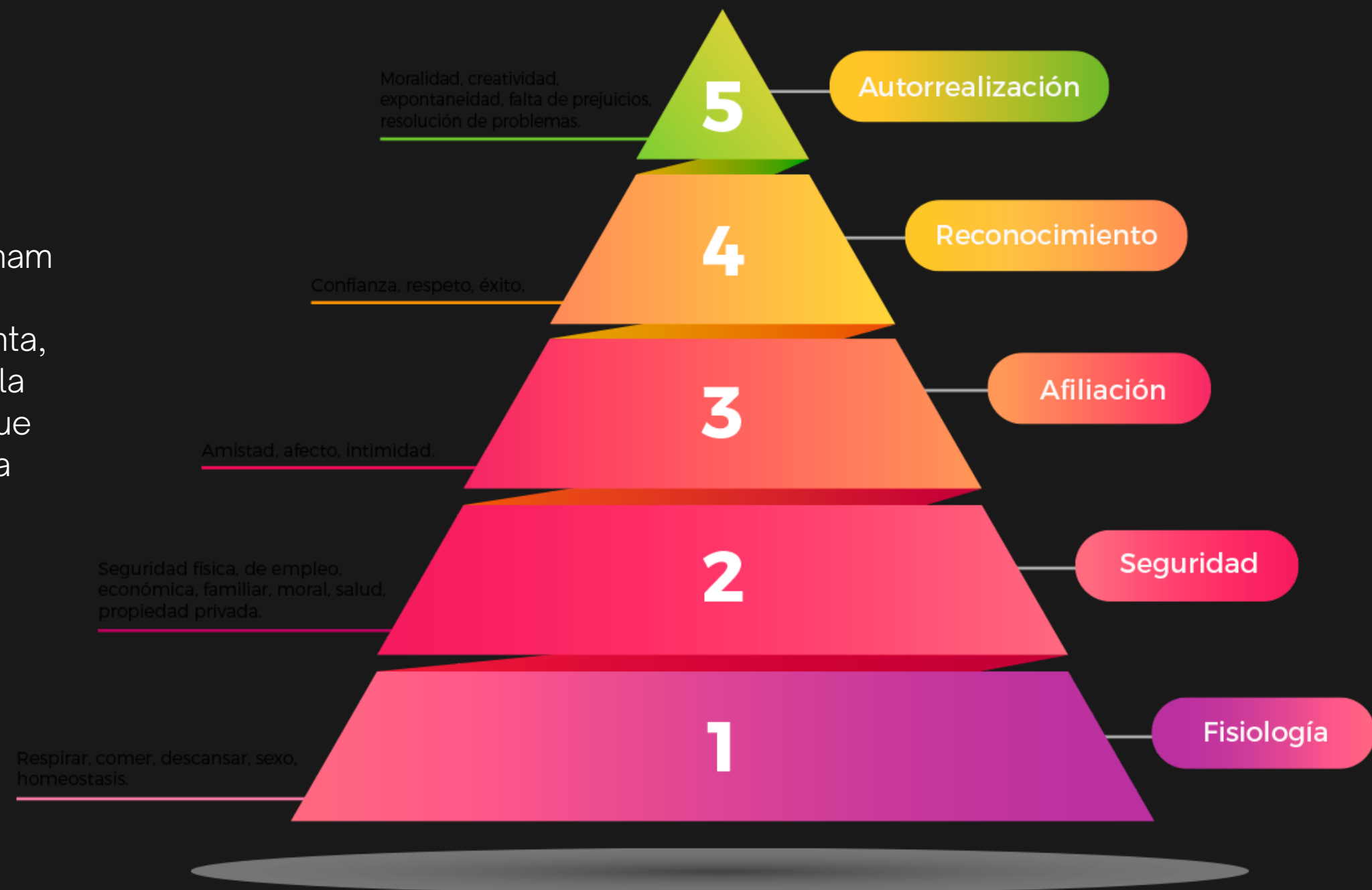
# Tabla de contenidos

- 1 El problema
- 2 Las preguntas de investigación
- 3 Data Acquisition / Data Wrangling
- 4 Análisis exploratorios EDA
- 5 Modelos candidatos - Modelo elegido
- 6 Conclusiones
- 7 Lineas futuras
- 8 Bibliografía

# Satisfacción personal

## La Pirámide de Maslow

La jerarquía de necesidades fue planteada por Abraham Maslow en su libro *Motivation and Personality* (1954, *Motivación y Personalidad*), dicha jerarquía fundamenta, en mucho, el desarrollo de la escuela humanista de la administración y permite adentrarse en las causas que mueven a las personas a trabajar en una empresa y a aportar parte de su vida a ella.





# El problema

El área de RRHH de una empresa tiene entre sus objetivos principales **detectar las oportunidades de mejora** para ajustar el gap entre las competencias de las personal y los requerimientos de los puestos en particular. Para esto se realizan evaluaciones de desempeño periódicas a las personas y a los equipos de trabajo en cuanto a su rendimiento real en comparación a lo planificado según la planificación de RRHH.





# El problema

Una función clave del área, es asegurar de que se viva en un **clima laboral favorable y cómodo** para todos los colaboradores. Entre otros actores a tener en cuenta, esto ayuda a asegurar altos índices de productividad en los procesos de producción/servucción, al mismo tiempo que ayuda a mantener la motivación y la fidelidad de los integrantes de la organización en todos sus niveles.

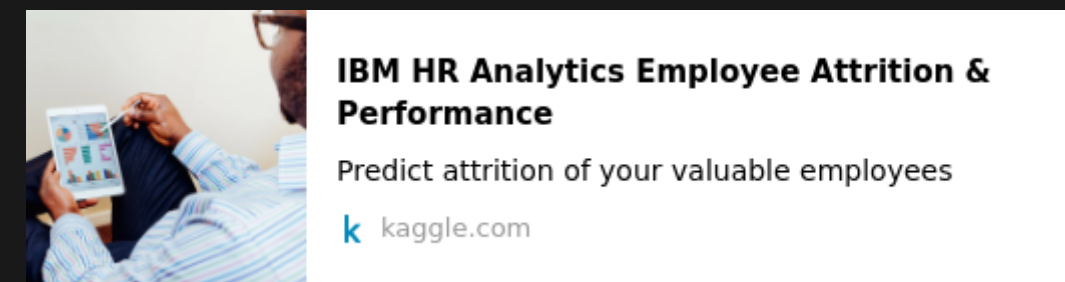
La satisfacción laboral es un constructo interdisciplinar cuya conceptualización ha sido abordado por la psicología, la sociología, **la Administración.**

# El dataset

---

## IBM HR Analytics Employee Attrition & Performance

IBM ha recopilado información sobre la satisfacción de los empleados, los ingresos, la antigüedad y algunos datos demográficos. Incluye los datos de 1470 empleados.





# Preguntas y objetivos de la investigación

Trabajaremos con un dataset con datos ficticios para determinar qué factores **influyen en el desgaste y deserción** laboral para que el área de recursos humanos pueda tomar acciones a tiempo que permitan evitarlo.

- ¿Cuál es la tasa de rotación por estrato de antigüedad del personal?
- ¿Cuáles son las causas determinantes que hacen que el empleado decida renunciar o disminuya su rendimiento por insatisfacción laboral?

- ¿Cuánto es el tiempo de permanencia en una empresa antes de aceptar una oferta del mercado mejor paga?

# Storytelling

---

01.

**1927**

El profesor de Harvard Elton Mayo se propuso estudiar, junto a un grupo de colaboradores, la relación existente entre las características ambientales y la productividad de los trabajadores

---

02.

**Hawthorne en el trabajo**

Los resultados del experimento de Elton Mayo sirvieron para establecer las bases de lo que hoy se conoce como la Escuela de las Relaciones Humanas



# Storytelling


---



03.

## Relación con el grupo

Los trabajadores, cuando se sienten valorados, motivados y satisfechos con su trabajo y tienen relaciones informales positivas con otros compañeros, son más productivos



04.

## Disminución del abandono

El reconocimiento, la seguridad y el sentido de pertenencia son claves en la determinación de un trabajador, y eso disminuye el abandono de empleos.

Data **Acquisition** / Data **Wrangling**



# Implementación

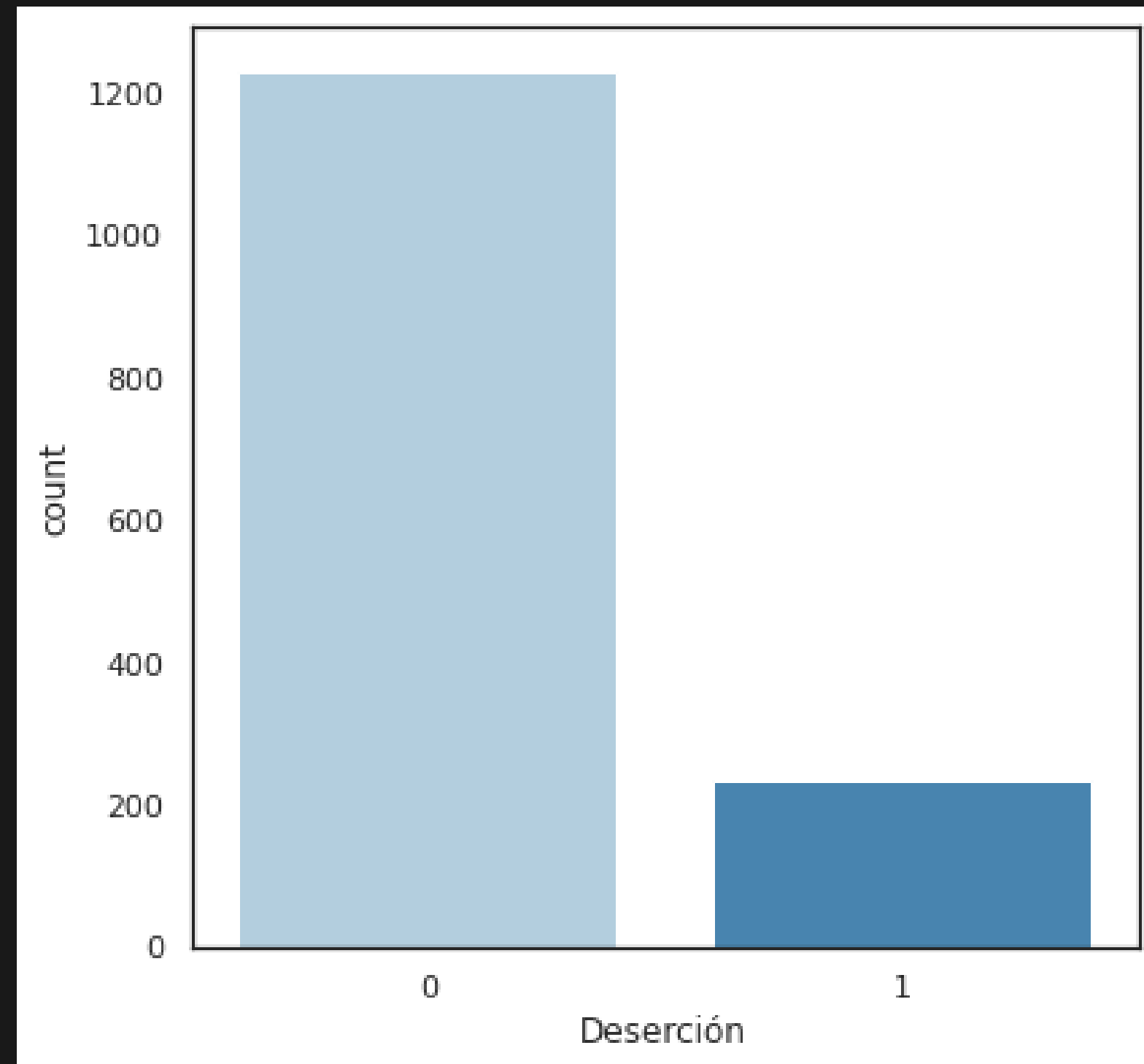
Leímos el dataframe y realizamos Data Wrangling y Limpieza para identificar los datos relevantes, también se borraron algunas columnas por ser irrelevantes para el caso de estudio. El conjunto de datos contiene 1470 registros con 35 features cada uno.

Edad	1470.0	36.92	9.14	18.0	30.0
Tarifa Diaria	1470.0	802.49	403.51	102.0	465.0
Distancia Desde Casa	1470.0	9.19	8.11	1.0	2.0
Educación	1470.0	2.91	1.02	1.0	2.0
Satisfacción Ambiental	1470.0	2.72	1.09	1.0	2.0
Tarifa por hora	1470.0	65.89	20.33	30.0	48.0
Participación en el trabajo	1470.0	2.73	0.71	1.0	2.0
Nivel de trabajo	1470.0	2.06	1.11	1.0	1.0
Satisfacción laboral	1470.0	2.73	1.10	1.0	2.0
Ingreso mensual	1470.0	6502.93	4707.96	1009.0	2911.0
Tarifa mensual	1470.0	14313.10	7117.79	2094.0	8047.0
Número de empresas trabajadas	1470.0	2.69	2.50	0.0	1.0
Porcentaje de aumento de salario	1470.0	15.21	3.66	11.0	12.0
Clasificación de Rendimiento	1470.0	3.15	0.36	3.0	3.0

# Análisis exploratorios EDA

Estamos frente a un problema de Clase desbalanceada.

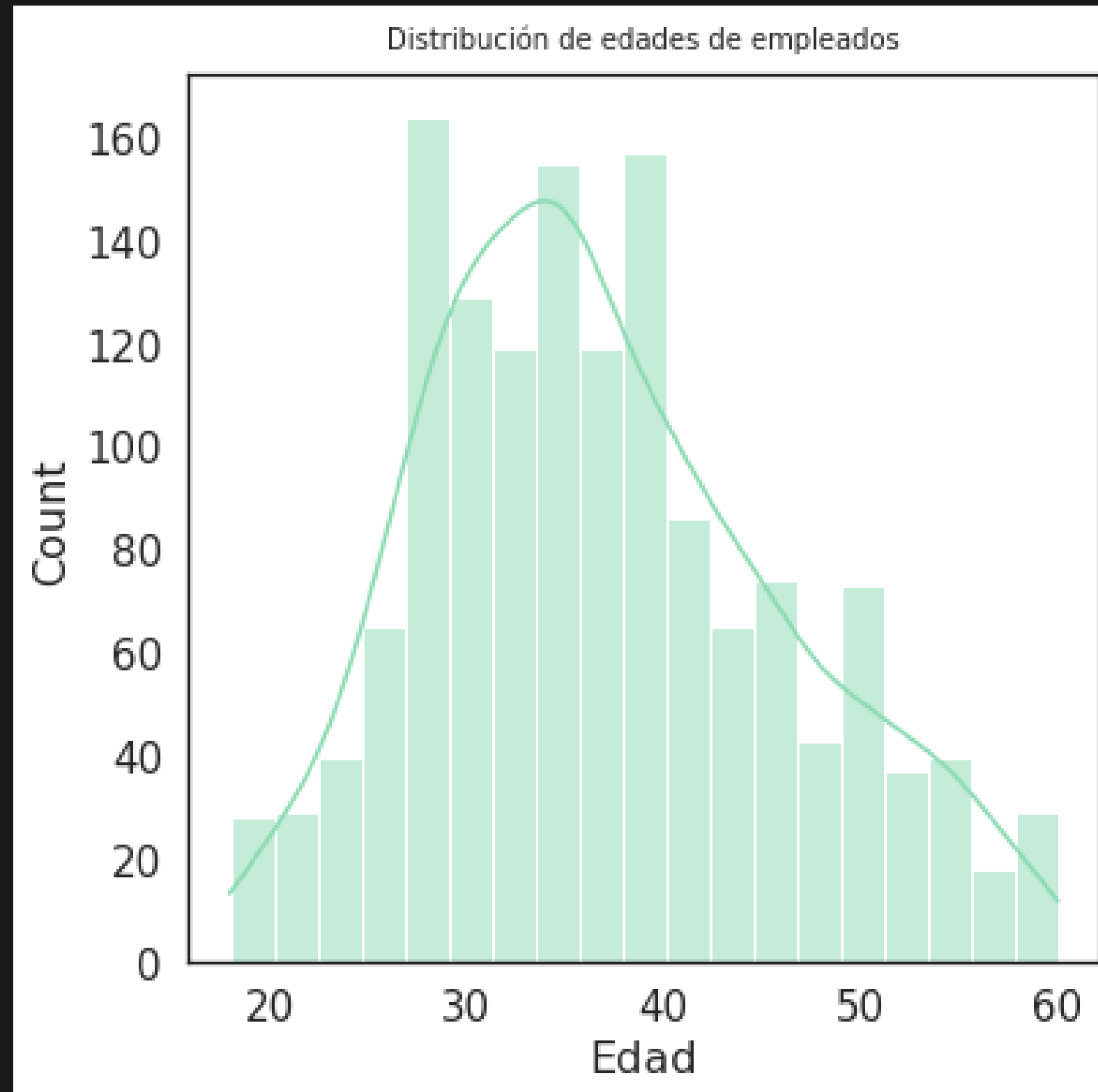
## Análisis univariado



# Análisis univariado

De los 1470, 237 renunciaron (deserción o desgaste), esto puede ser muy bueno o no, dependiendo la industria. Lo que nosotros trataremos de predecir es si una persona renunciará o no de acuerdo a sus características. Nuestra variable target toma como valor NO para el 84% de nuestros empleados y SI (o yes ó 1) para el 16%.

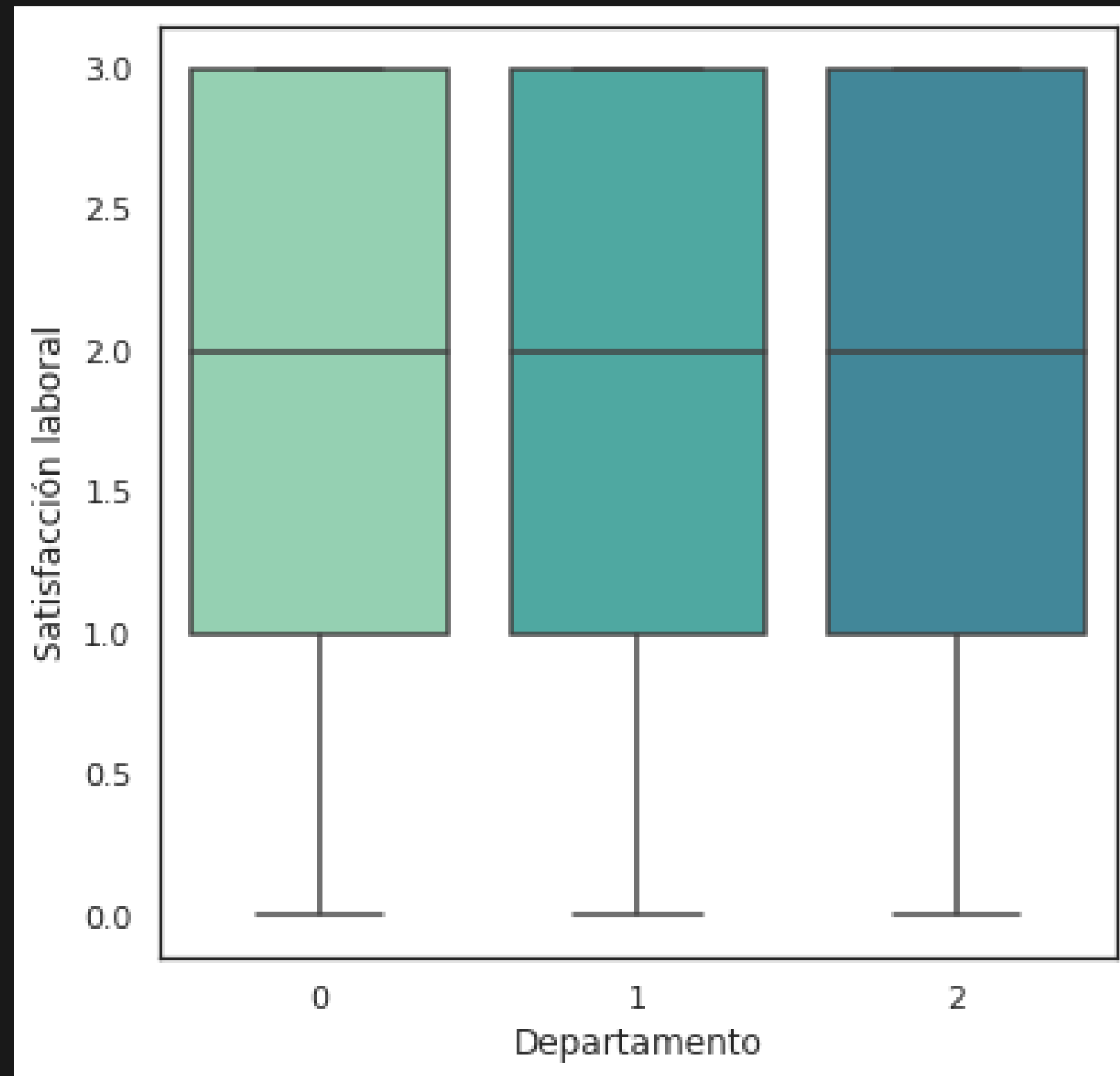
Por otra parte, esta empresa consta de un 40 % de mano de obra femenina y la distribución de las edades están alrededor de la media 36.92 y es similar a la normal





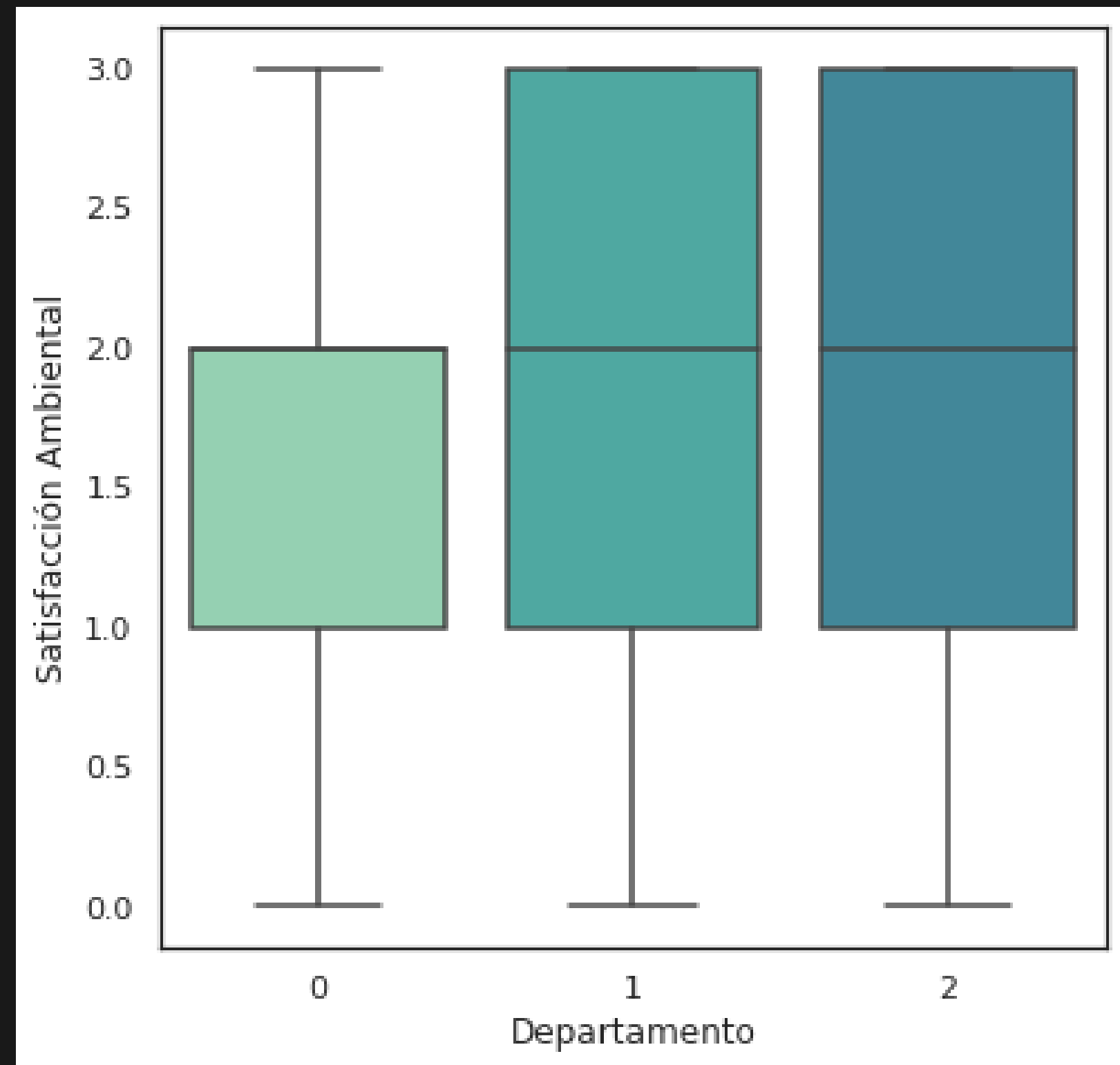
# Análisis bivariado

Departamento -Satisfacción laboral



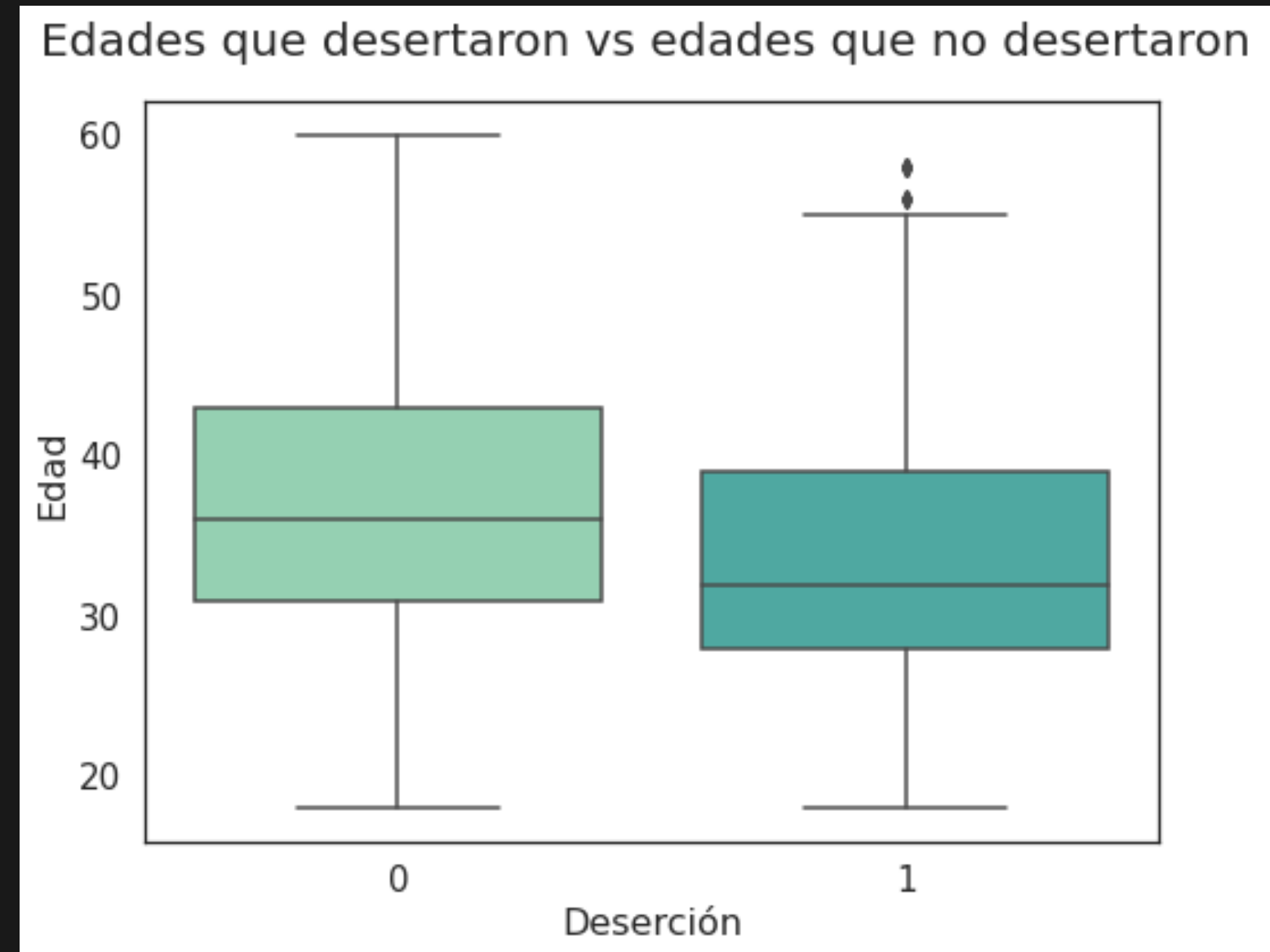
# Análisis bivariado

Departamento -Satisfacción Ambiental



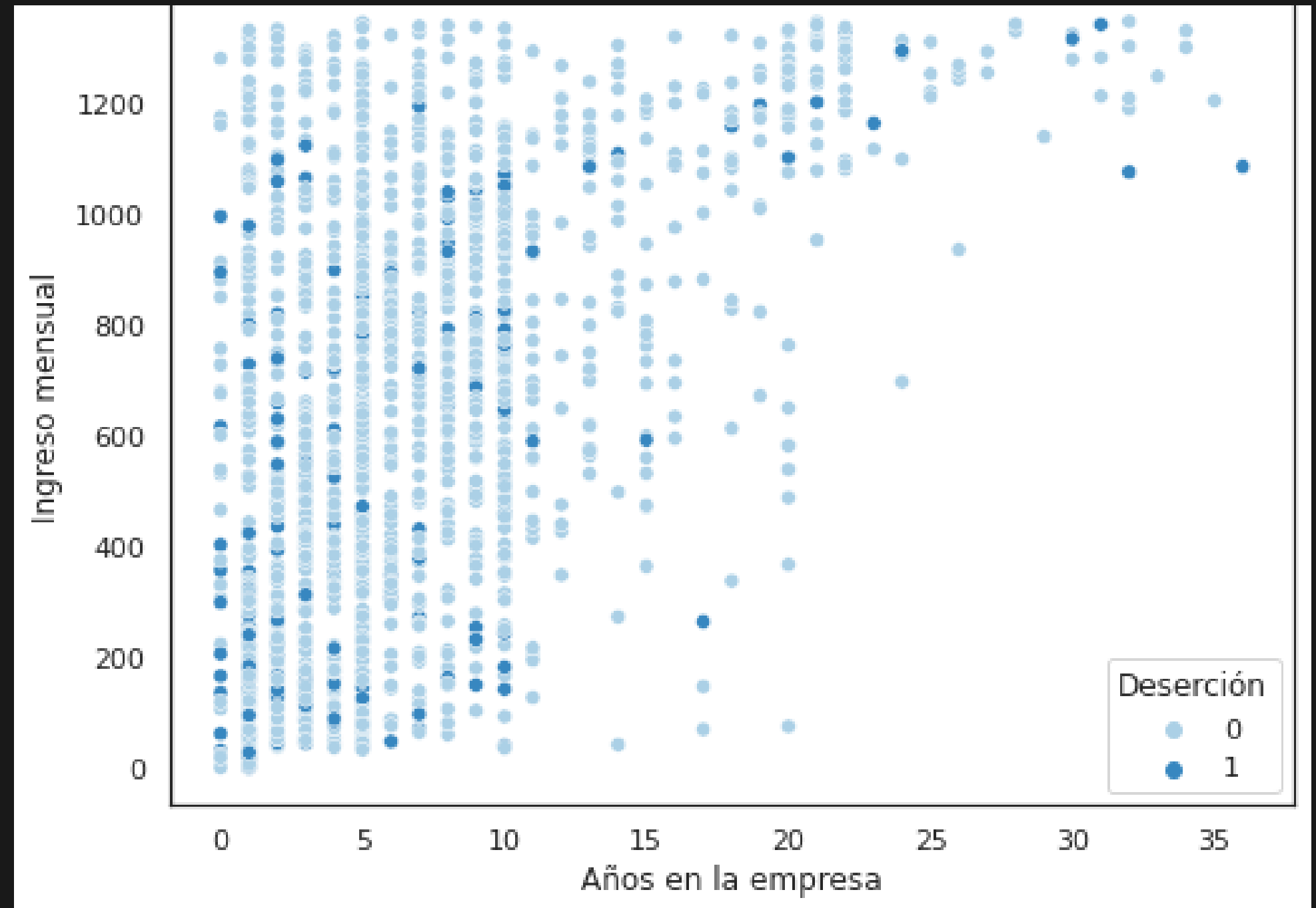
# Análisis bivariado

No hay diferencias significativas entre la satisfacción de los distintos departamentos, por lo cual no valdria la pena hacer una subdivisión de ellos para un analisis particular. Puede observarse que la distribución de edades entre quienes desertan y quienes no es bastante similar. Se nota más concentración de edad alrededor de la media entre quienes no desertan. No se observa que la edad sea un factor determinante en la decisión de desertar.



# Análisis multivariado

Se puede observar en el diagrama de dispersión que existe mayor cantidad de deserción por desgaste en los primeros años trabajados relacionados con bajos ingresos. La satisfacción laboral es cambiante porque crecen y decrecen los sentimientos satisfactorios a medida que los motivos de logro se van cubriendo, en caso de que esto no ocurra, las consecuencias suelen ser como se muestran en el diagrama





# Correlación de variables

---

Se realizó un primer análisis para intentar hacer un filtro de características o variables que eran independientes, para esto se eliminó la variable target del dataset, que es la deserción.

La variable target o a predecir es Deserción ,es una variable CATEGÓRICA que tiene dos posibilidades YES o NO. Se normalizaron esos valores de modo que contengan 1 y 0 respectivamente.

Todas las variables son numéricas, ya sea enteras o reales y no tiene valores nulos. También nos dimos cuenta que no había datos ausentes en el dataset ni "sucios".

Se realizó un segundo análisis para intentar hacer un filtro de características o variables que no sean importantes, ahora si se incluyó la variable Deserción.

# Correlación de variables



# Correlación de variables

---

Del mapa de calor de correlación, podemos ver que algunas de nuestras variables parecen estar muy correlacionadas entre sí y otras con una correlación casi nulas.

La mayoría de las features están mal correlacionadas entre si, salvo:

- Años en la compañía
- Años en el rol actual
- Años desde la última promoción
- Años con el actual manager
- Total de años trabajados

Existe muy buena relación entre

- Clasificación de rendimiento y Porcentaje de aumentos de salario : 0.77
- Ingreso mensual y Nivel de trabajo 0.89

Como no tenemos un gran número de variables correlacionadas entre si decidimos no aplicar la técnica PCA y decidimos eliminar variables altamente correlacionadas.

**Modelos candidatos**



# Modelos candidatos

---

## Seleccionando posibles algoritmos candidatos

- Decision Tree
- Random Forest
- Ada Boost
- Xgboost
- Logistic Regresion

# Comparación entre modelos

Modelo	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score	Tiempo de ejecución
DecisionTree	0.874773	0.807065	0.457143	0.235294	0.310680	0.803616
RandomForestOptimizado	0.905626	0.836957	0.900000	0.132353	0.230769	134.650529
AdaBoostClassifier	0.952813	0.820652	0.520000	0.382353	0.440678	8.031217
modeloXgboost	0.881125	0.823370	0.571429	0.176471	0.269663	0.114809
Regresionlogistica	0.768603	0.747283	0.403101	0.764706	0.527919	2.339383

- Accuracy: El modelo ajusta razonablemente bien con más de 75% en accuracy de tanto en train como en test en todos los modelos
- Precisión: El modelo que mejor precision tiene es Random Forest fue optimizado buscando los mejores hiperparámetros con RandomizedSearchCV
- Recall: El modelo que tiene mejor recall o sensibilidad, es decir proporción de casos positivos bien identificados por el algoritmo es el modelo de Regresión Logística.
- f1-Score: De todos los modelos el que mejor se ajusta en f1 es el modelo de AdaBoostClassifier.

# Comparación entre modelos

---

Con la métrica de precisión podemos medir la calidad del modelo de machine learning en tareas de clasificación, en nuestro caso teniendo en cuenta los resultados el modelo más preciso es Random Forest.

Si observamos los resultados respecto a la métrica de sensibilidad (recall), que nos informa sobre la cantidad que el modelo de machine learning es capaz de identificar, podemos concluir que este modelo tiene un valor de 0.11 respecto a Regresión Logística que tiene considerablemente superior con 0.76 como valor.

# Comparación entre modelos

---

Nos interesa analizar más profundamente esta medida, ya que Recall es la proporción de **positivos reales** que se identificó que van a abandonar su posición laboral, aunque ese número también podría incluir falsos positivos, los cuáles serían los que no se quieren ir realmente de su puesto de trabajo o lo que están en duda si en abandonar la posición, para lo cuál deberíamos modificar la precisión para que esto disminuya.

Sin embargo, si quisieramos comparar el rendimiento combinado de la precisión y la exhaustividad del modelo elegido, el valor F1 score sería el utilizado para combinar las medidas de precisión y recall en un sólo valor. En resumen, sabemos que **Recall** es la métrica que usamos para seleccionar nuestro mejor modelo cuando hay un alto costo asociado con el **Falso Negativo**. Debido a que el modelo Regresión logística obtuvo los mejores resultados con respecto a esta métrica, lo usaremos para los próximos pasos.

# ROC

---

**Para interpretar correctamente las predicciones realizadas por modelos de clasificación binarios (dos clases) utilizaremos las curvas ROC y las curvas de precisión-sensibilidad (Precision-Recall)**

ROC es un acrónimo que viene del inglés **Receiver Operating Characteristic** (Característica Operativa del Receptor). Es una gráfica que enfrenta el ratio de falsos positivos (eje x) con el ratio de falsos negativos (eje y). Estos ratios los va obteniendo en función de una serie de umbrales definidos entre **0 y 1**. En palabras comunes y referenciando al ejemplo anterior, enfrenta la «falsa alarma» vs la tasa de éxito.

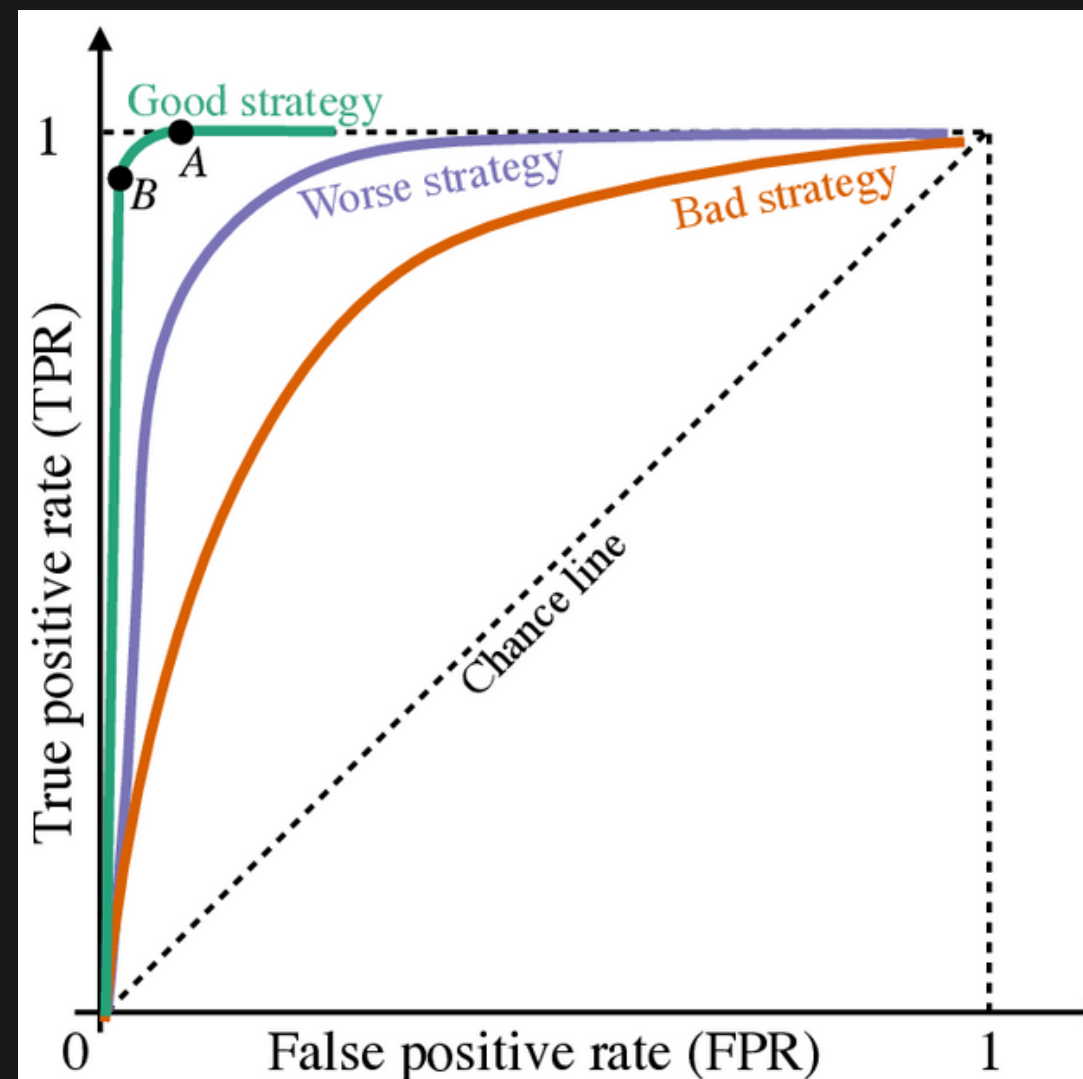
Es útil por dos principales motivos:

- Permite comparar diferentes modelos para identificar cual otorga mejor rendimiento como clasificador.
- El área debajo de la curva (AUC) puede ser utilizado como resumen de la calidad del modelo.

# ROC

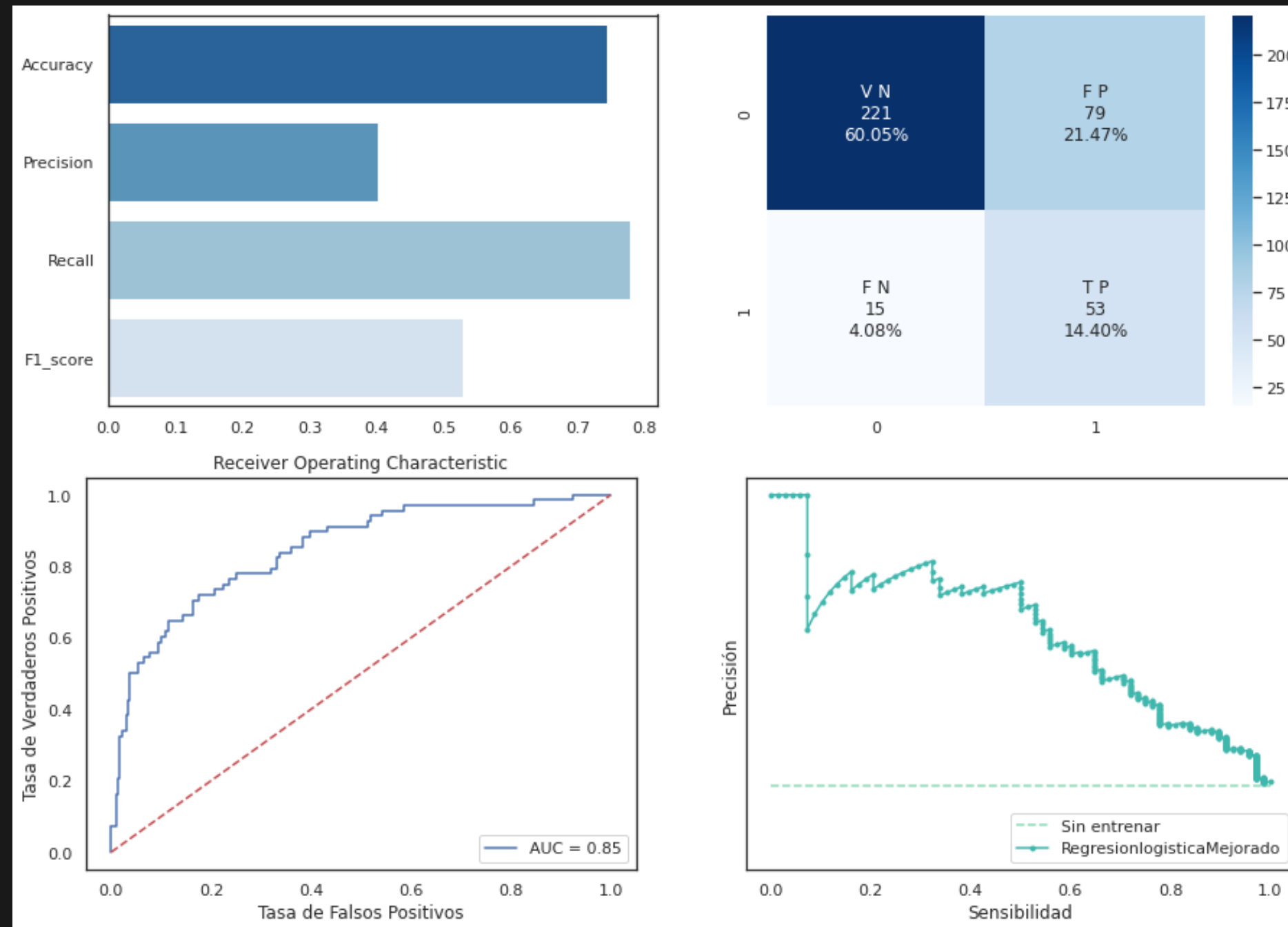
## Criterios:

- Valores pequeños en el eje X indican pocos falsos positivos y muchos verdaderos negativos
- Valores grandes en el eje Y indican elevados verdaderos positivos y pocos falsos negativos



# Regresión logística mejorado

El modelo elegido es de Regresión logística



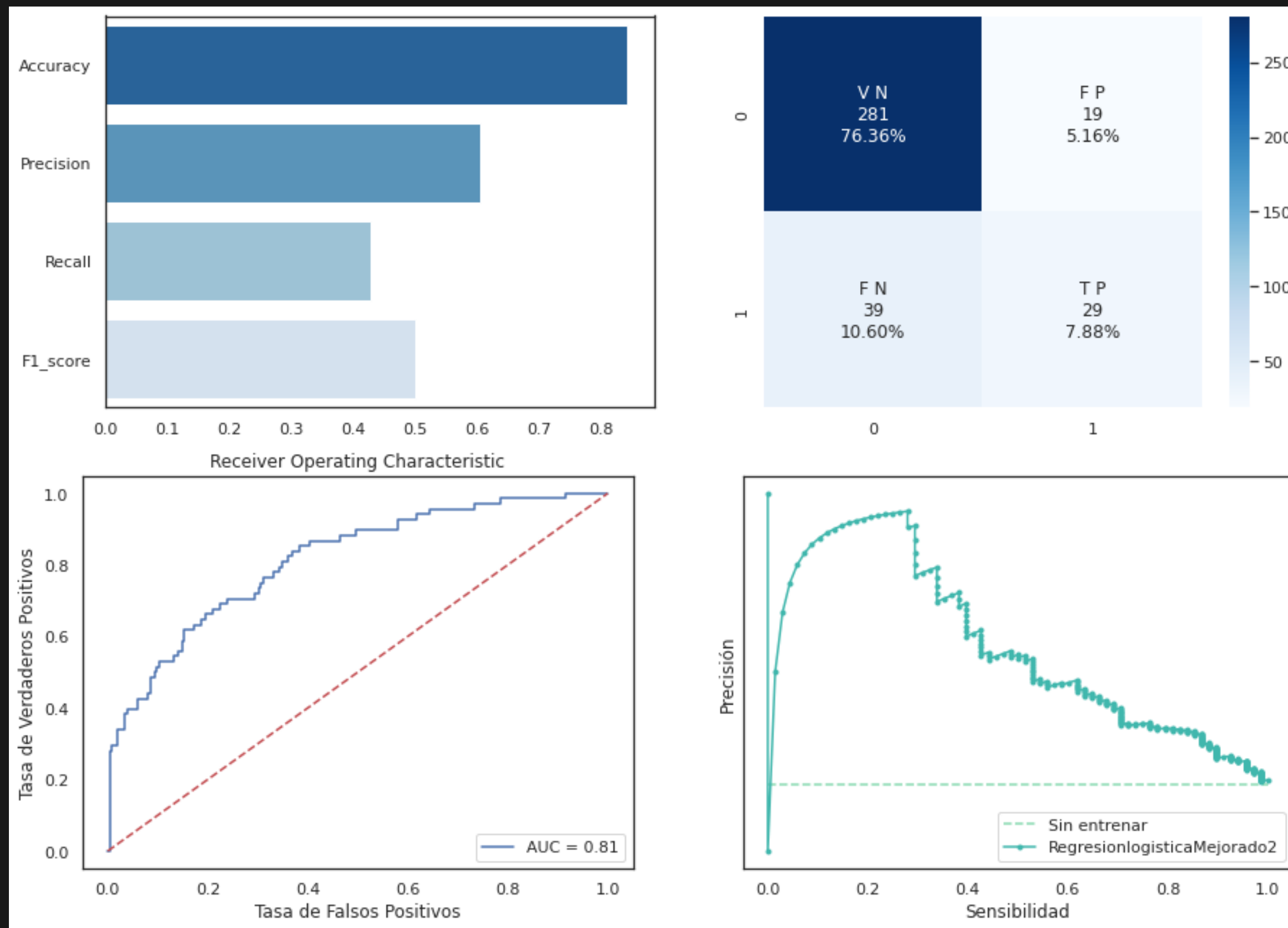
# Regresión logística mejorado

---

Modelo	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score	Tiempo de ejecución
RegresionlogisticaMejorado	0.766788	0.744565	0.401515	0.779412	0.53	0.267132



# Regresión logística mejorado V2



# Regresión logística mejorado V2

---

Modelo	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score	Tiempo de ejecución
RegresionlogisticaMejorado2	0.854809	0.842391	0.604167	0.426471	0.5	0.558594

# Regresión logística mejorado V2

---

Modelo	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score	Tiempo de ejecución
Regresionlogistica	0.768603	0.747283	0.403101	0.764706	0.527919	2.339383
RegresionlogisticaMejorado	0.766788	0.744565	0.401515	0.779412	0.530000	0.219904
RegresionlogisticaMejorado2	0.854809	0.842391	0.604167	0.426471	0.500000	0.394769

# Iteraciones de Optimización

Modelo	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score	Tiempo de ejecución
Regresionlogistica	0.768603	0.747283	0.403101	0.764706	0.527919	2.339383
RegresionlogisticaMejorado	0.766788	0.744565	0.401515	0.779412	0.530000	0.219904
RegresionlogisticaMejorado2	0.854809	0.842391	0.604167	0.426471	0.500000	0.394769

Seguiríamos seleccionando al modelo de Regresión logística con la mejora 1 aplicada, es decir con tuning de hiperparámetros ya que si bien aplicando la técnica **SMOTE** mejora considerablemente la exactitud (casi 10 puntos) perdemos mucho de sensibilidad (**recall**) y nuestro objetivo es obtener buenos resultados respecto a esta medida. Además es mas performante.

# Conclusiones

---

- Si bien el salario no es competitivo, el mal ambiente de trabajo o la mala relación con el jefe pueden ser razones para que un trabajador renuncie, aunque estas no son razones suficientes para que un empleado renuncie.
- La renuncia laboral es provocada por una combinación de múltiples factores que pueden o no ser parte de las características de este conjunto de datos, sin embargo, se debe tener en cuenta que cada empresa presentará diversos factores y formas de calificar al trabajador, por lo que este conjunto de datos debe tomarse como una visión general.
- La regresión logística demostró ser una la mejor herramienta para clasificar y predecir qué empleados no renunciarán, y eso permite a las empresas seguir tomando medidas y beneficios necesarias para expandir la satisfacción laboral y mejorar la calidad de vida del empleado.
- Pudimos responder algunas preguntas del problema inicial planteado pero a lo largo de la investigación y el desarrollo surgieron otras nuevas.

# Futuras líneas

---

- Rasgos de personalidad del empleado
- Antecedentes laborales anteriores que influyan en sus decisiones actuales
- Condiciones externas que influyan en su satisfacción laboral



# Recursos

---

## Algunos sitios web que nos sirven de referencia:

- <https://www.comparably.com/companies/ibm/employee-engagement>
- <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
- <https://www.questionpro.com/blog/es/como-hacer-encuestas-de-satisfaccion-de-empleados/>
- [http://bibliotecadigital.econ.uba.ar/download/tpos/1502-0921\\_BenglerAA.pdf](http://bibliotecadigital.econ.uba.ar/download/tpos/1502-0921_BenglerAA.pdf)