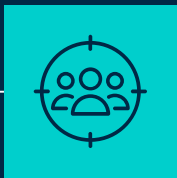


# TRABAJO PRÁCTICO DATA SCIENCE

CoderHouse

# INDICE



01

Conformación del  
equipo de trabajo.



02

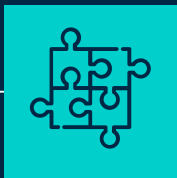
Presentación del  
caso/problema  
específico.



03

Preguntas y  
objetivos de la  
investigación.

# INDICE



04

EDA — Exploratory  
Data Analysis



05

Aplicación de  
algoritmos de ML



06

Conclusiones

# CONFORMACIÓN DEL EQUIPO DE TRABAJO

01

# EQUIPO DE TRABAJO



**TAMARA FAYA**

- Lic. Cs Físicas
- Analista Ssr de datos en Santander

**GUADALUPE  
A. LAMAS**

- Lic. en Administración y Sistemas
- Analista de datos en Santander



**VITTORIA DEL  
SIGNORE**

- Ing. Química
- Analista Data Management en BNP Paribas Cardif

# PRESENTACIÓN DEL CASO

02

# CASO: Dataset de empleados de IBM

- Corresponde a **información de empleados** de la empresa IBM.
- Incluye diferentes variables tanto **personales** como **laborales** de cada uno de los empleados.
- **1470 registros**

Column	Type	Non-Null	Nulls	Unique	Example
Age	int64	1470	0	43	41
Attrition	object	1470	0	2	Yes
BusinessTravel	object	1470	0	3	Travel_Rarely
DailyRate	int64	1470	0	886	1102
Department	object	1470	0	3	Sales
DistanceFromHome	int64	1470	0	29	1
Education	int64	1470	0	5	2
EducationField	object	1470	0	6	Life Sciences
EmployeeCount	int64	1470	0	1	1
EmployeeNumber	int64	1470	0	1470	1
EnvironmentSatisfaction	int64	1470	0	4	2
Gender	object	1470	0	2	Female
HourlyRate	int64	1470	0	71	94
JobInvolvement	int64	1470	0	4	3
JobLevel	int64	1470	0	5	2
JobRole	object	1470	0	9	Sales_Executive
JobSatisfaction	int64	1470	0	4	4
MaritalStatus	object	1470	0	3	Single
MonthlyIncome	int64	1470	0	1349	5993
MonthlyRate	int64	1470	0	1427	19479
NumCompaniesWorked	int64	1470	0	9	8
Over18	object	1470	0	1	Y
Overtime	object	1470	0	2	Yes
PercentSalaryHike	int64	1470	0	15	11
PerformanceRating	int64	1470	0	2	3
RelationshipSatisfaction	int64	1470	0	4	1
StandardHours	int64	1470	0	1	80
StockOptionLevel	int64	1470	0	3	0
TotalWorkingYears	int64	1470	0	39	8
TrainingTimesLastYear	int64	1470	0	6	0
WorkLifeBalance	int64	1470	0	4	1
YearsAtCompany	int64	1470	0	36	6
YearsInCurrentRole	int64	1470	0	18	4
YearsSinceLastPromotion	int64	1470	0	15	0
YearsWithCurrManager	int64	1470	0	17	5

# VARIABLE ATTRITION

## SI

Los empleados deciden renunciar y tenemos que descubrir qué es aquello que los llevó a tomar esa decisión.

## NO

Los empleados aún continúan en la empresa. Tenemos que entender que variables son las que los motivan a quedarse o descubrir quienes podrían ser los siguientes a renunciar para prevenirlo.



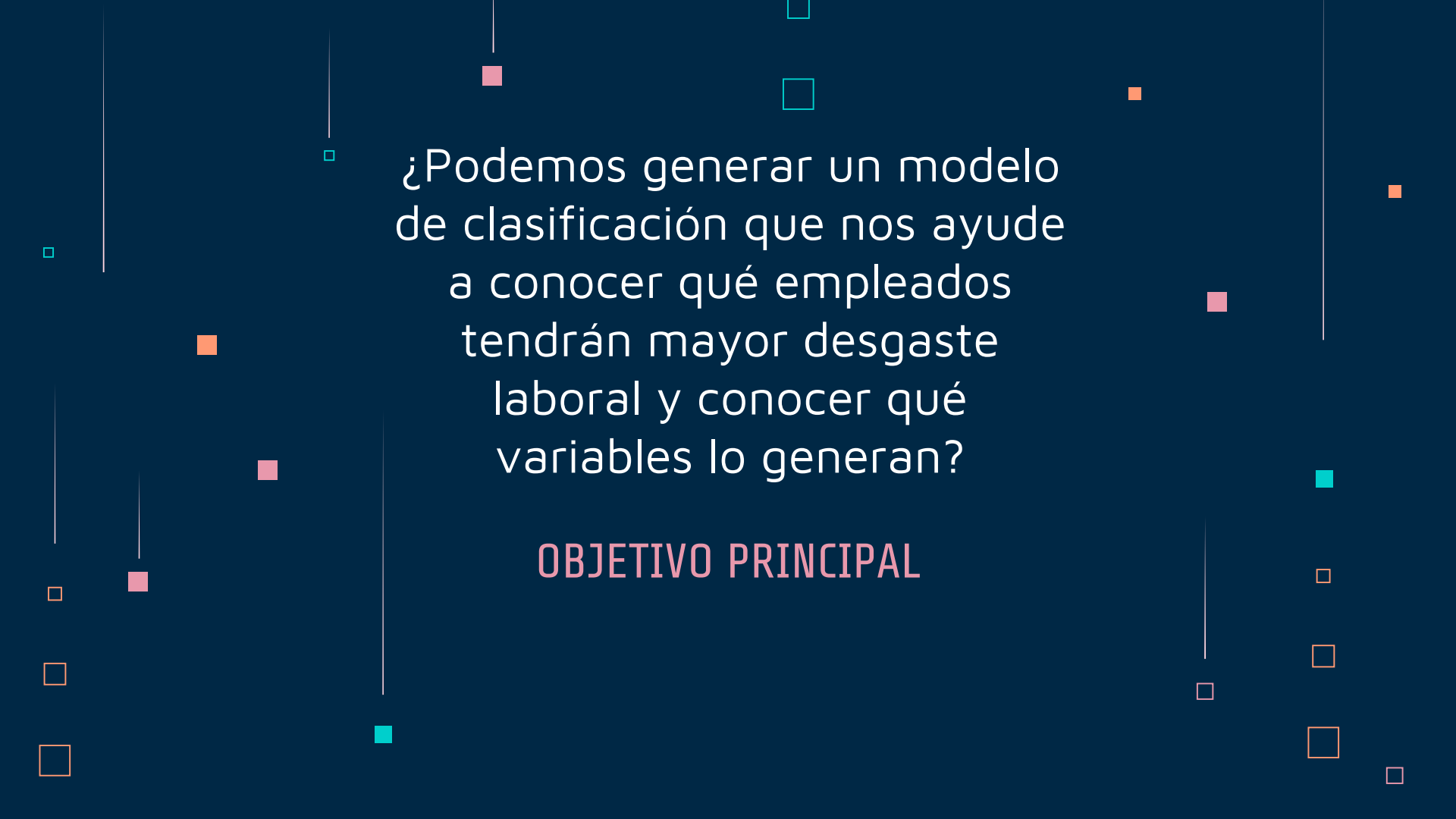


# PREGUNTAS Y OBJETIVOS DE LA INVESTIGACIÓN

03

# ALGUNAS PREGUNTAS SOBRE EL DATASET

1. ¿Qué porcentaje de empleados del estudio afirma tener **desgaste en su trabajo**?
2. ¿Es la **distancia** desde la casa al trabajo un factor determinante para que una persona quiera cambiar de trabajo?
3. ¿Cuántas personas hay de acuerdo a los **niveles de trabajo**?
4. ¿Qué **ramo de estudio** tienen las personas que trabajan en IBM?
5. ¿Los empleados están **satisfechos** con la empresa?
6. ¿Cuál es el **rango de salario** en la empresa?
7. ¿A mayor cantidad de años en la empresa mayor es el **salario**? ¿A mayor edad se incrementa el salario? o ¿El salario aumenta de acuerdo al nivel que tiene cada cargo?
8. ¿Cómo es en general el nivel de **balance entre el trabajo y la vida** de los empleados?
9. ¿Cuánto tiempo tienen los empleados con un **mismo jefe**?



¿Podemos generar un modelo  
de clasificación que nos ayude  
a conocer qué empleados  
tendrán mayor desgaste  
laboral y conocer qué  
variables lo generan?

**OBJETIVO PRINCIPAL**

# EDA

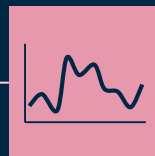
## Exploratory Data Analysis

04

# EDA – Exploratory Data Analysis

## UNIVARIADO

Análisis individual de las variables



## BIVARIADO

Búsqueda de correlaciones

## MULTIVARIADO

Análisis de relación y reducción de dimensionalidad.

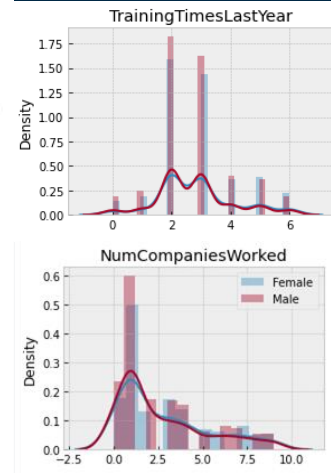
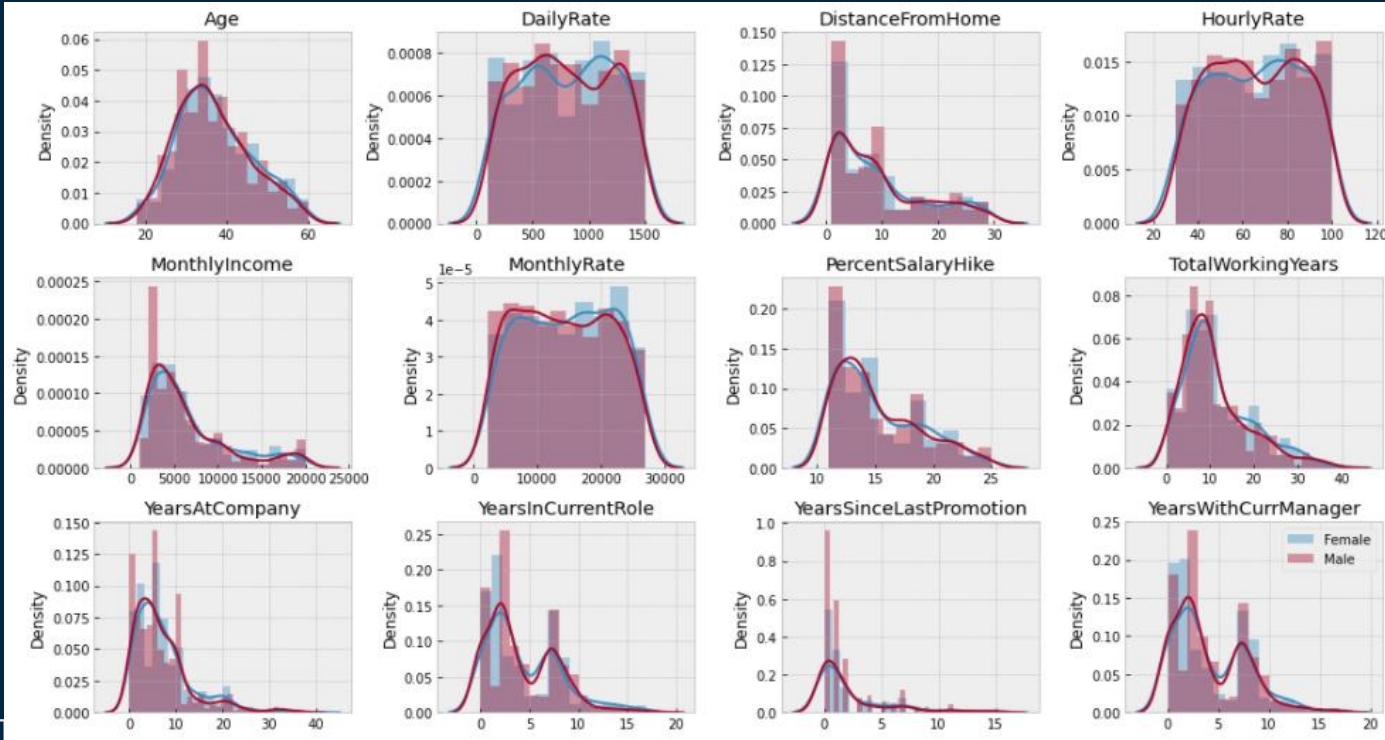


## RESUMEN

Descripción de las variables disponibles y selección de variables irreducibles.

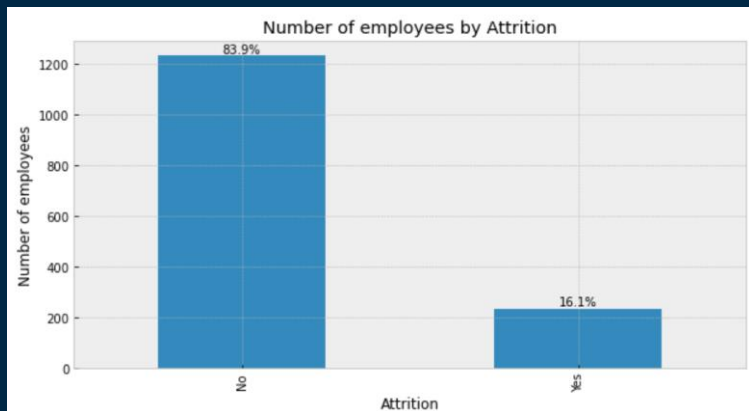


# UNIVARIADO



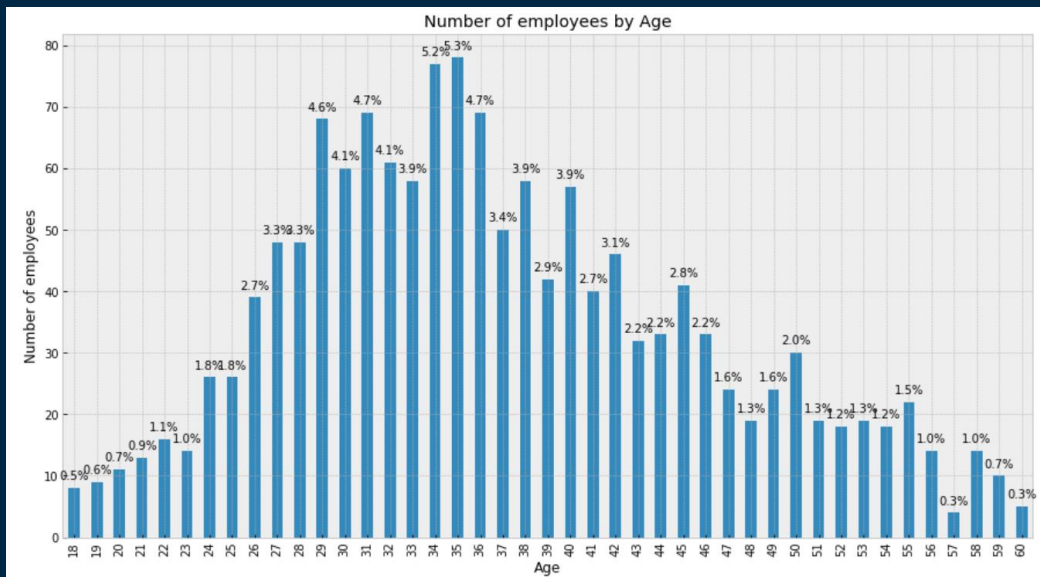


# UNIVARIADO



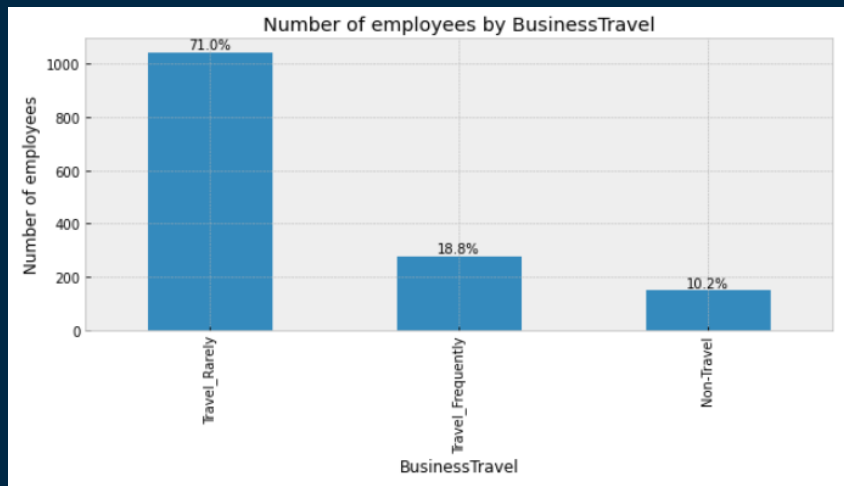
El 68% de los participantes se encuentra en el rango de edades de 28 a 46

Casi el 84 % de los empleados en este dataset dijo no tener desgaste laboral (attrition) mientras que el 16% si tuvo.



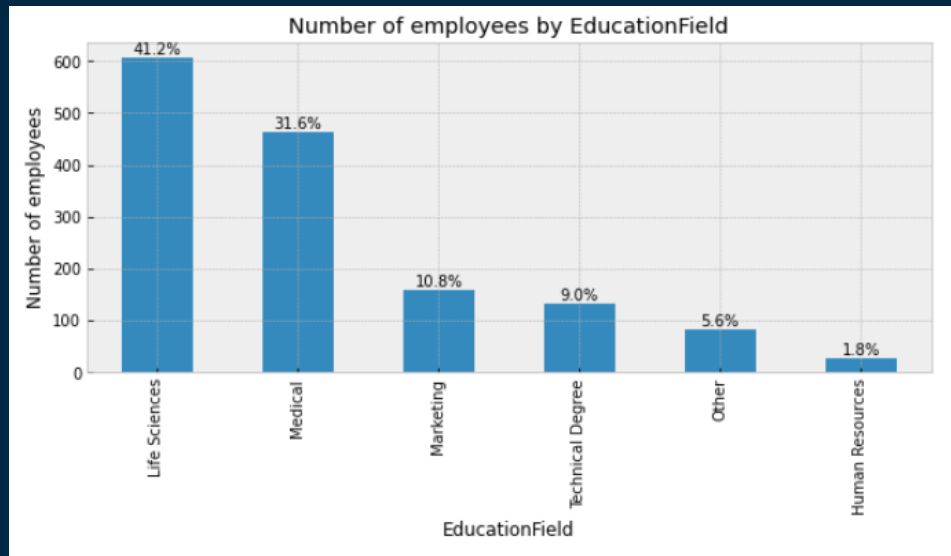


# UNIVARIADO



Tenemos 6 rubros educacionales. Más del 70% tienen formación en Life Sciences o Medical. El 30% restante se dividen en Marketing Technical Degree Human Resources y otros

El 81% de los empleados raramente viaja o no lo hace directamente. Solo el 18.8% viaja frecuentemente éste es un porcentaje parecido al indicado en attrition puede haber alguna relación?







# BIVARIADO

AGE (1)

DAILYRATE (2)

DISTANCEFROMHOME (3)

HOURLYRATE (4)

MONTHLYINCOME (5)

MONTHLYRATE (6)

PERCENTSALARYHIKE (7)

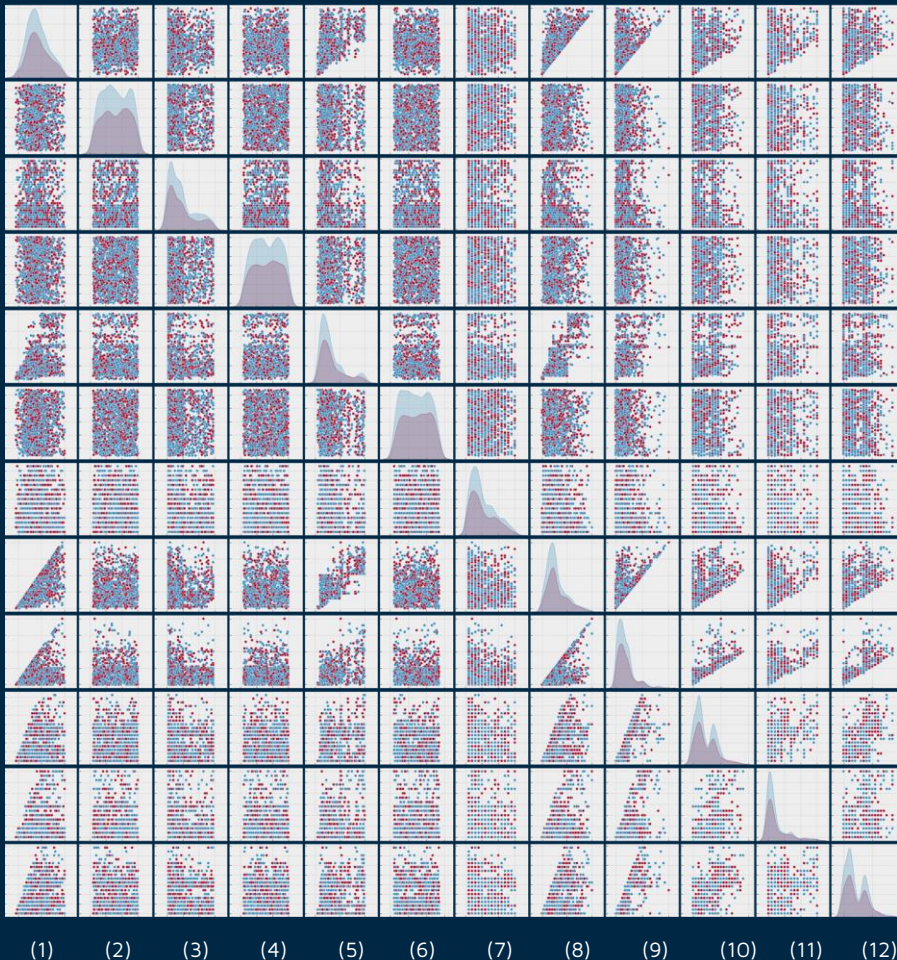
TOTALWORKINGYEARS (8)

YEARSATCOMPANY (9)

YEARSINCURRENTROLE (10)

YEARSSINCELASTPROMOTION (11)

YEARSWITHCURRMANAGER (12)

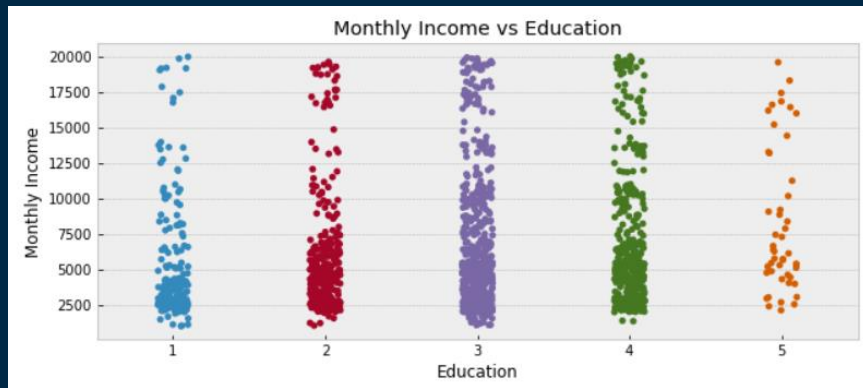


Gender  
● Male  
● Female





# BIVARIADO



La participación laboral es media en general independientemente del nivel de sueldos.

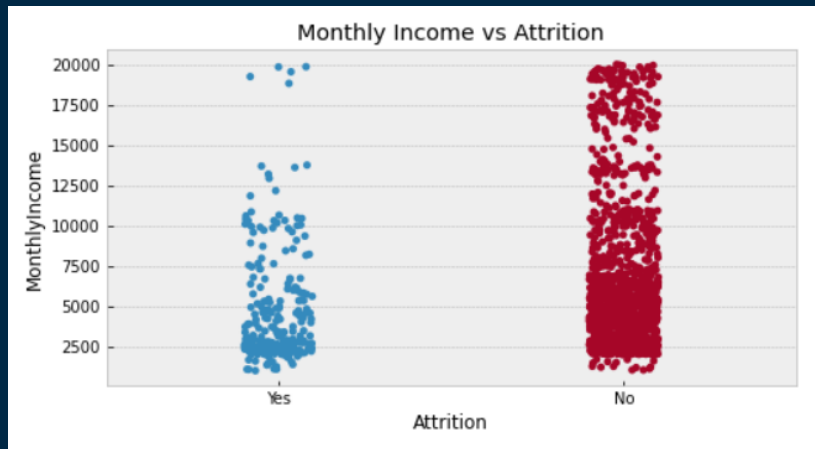
Se observa que la mayor concentración de los empleados se encuentra en Bachelor (3).

También podemos observar que en cada pilar de educación la mayor concentración es en salarios mas bajos siendo pocos quienes cobran mas dinero





# BIVARIADO



Los casos donde se ve el mayor desgaste laboral suele ser en aquellos donde los salarios son mas bajos

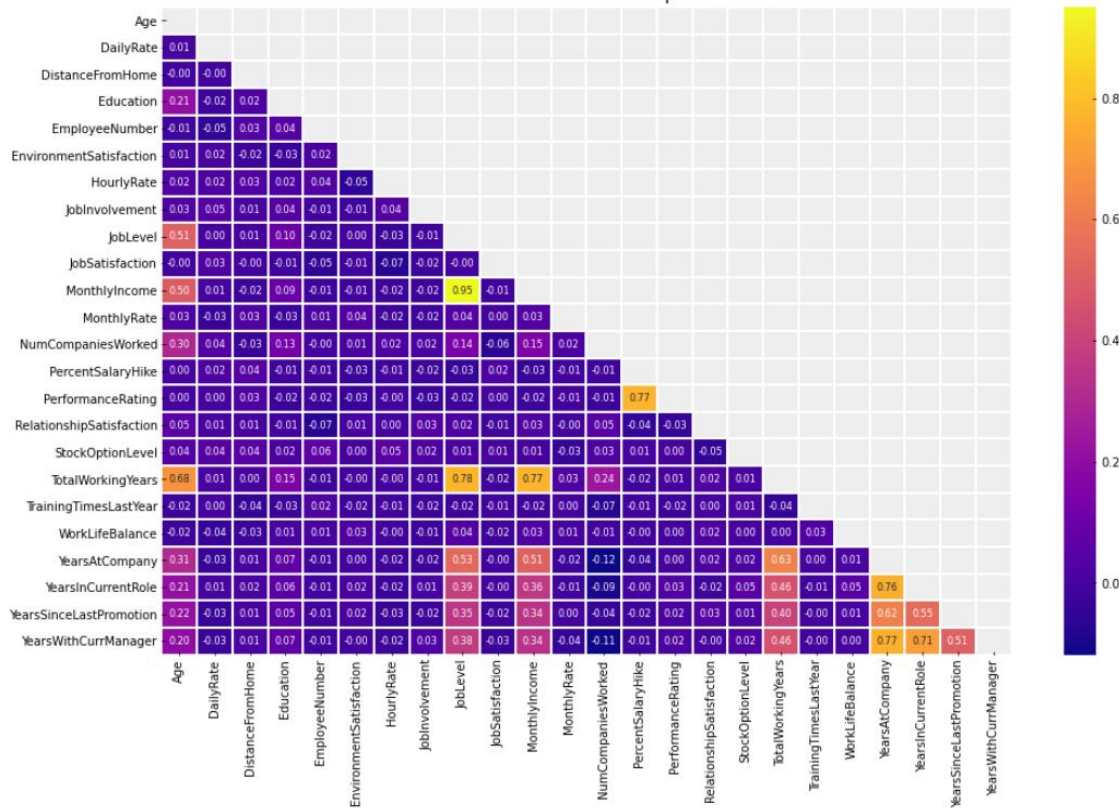
La relación entre el ingreso mensual y el balance vida-trabajo resulta similar a la relación del ingreso con la variable participación laboral. Se observa una aglomeración en los valores de bajos sueldos.





# MULTIVARIADO

Correlation Heatmap



Corr Coef	Variable 1	Variable 2
0,95	MonthlyIncome	JobLevel
0,78	TotalWorkingYears	JobLevel
0,77	TotalWorkingYears	MonthlyIncome
0,77	PerformanceRating	PercentSalaryHike
0,77	YearsWithCurrManager	YearsAtCompany
0,76	YearsInCurrentRole	YearAtCompany
0,71	YearsWithCurrManager	YearsInCurrentRole
0,68	TotalWorkingYears	Age



Estas son las variables con mayor correlación

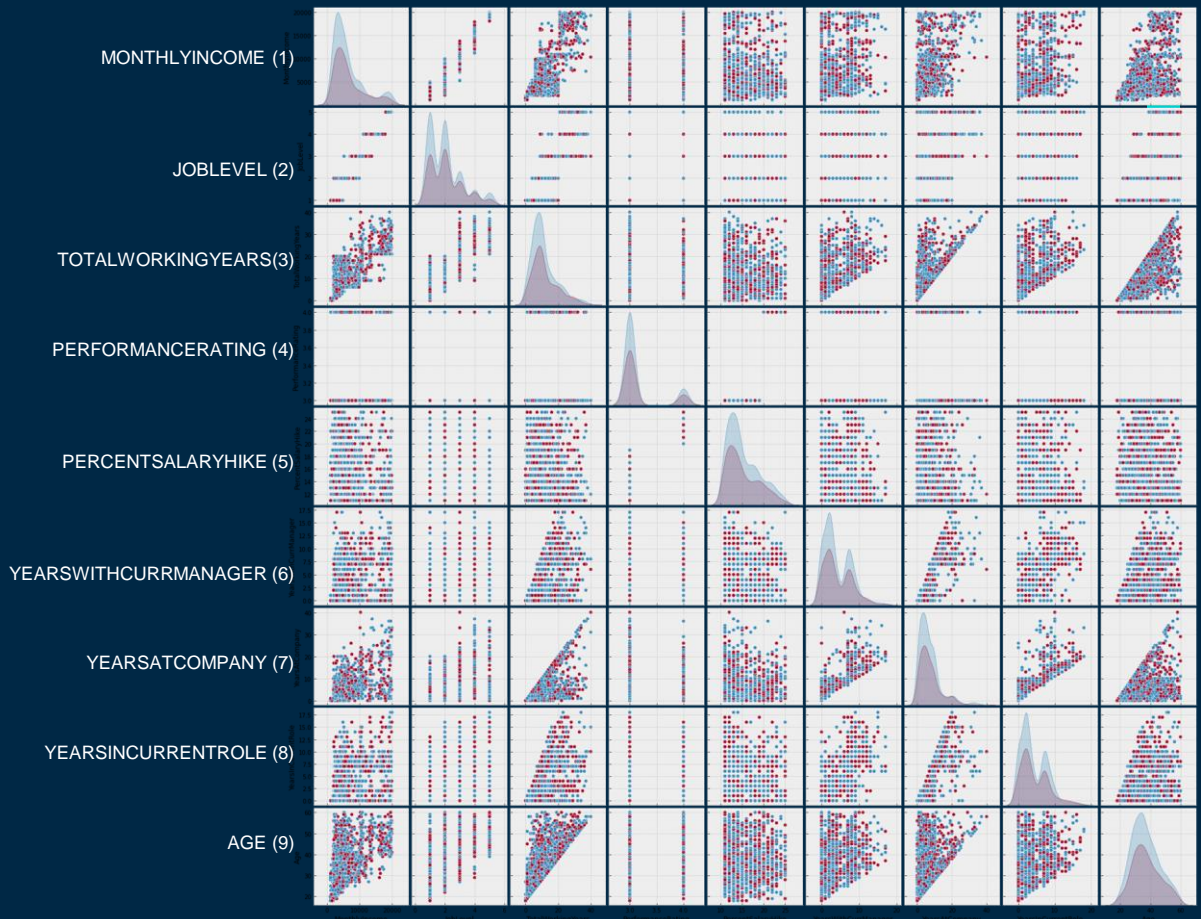




# MULTIVARIADO

Altos coeficientes de correlación no siempre implica una real correlación. Del análisis de las variables con coeficientes de correlación superior a 0.6; observamos que existen reales correlaciones entre las variables `MonthlyIncome` `JobLevel` y `TotalWorkingYears`. Que tiene sentido cuando pensamos que el ingreso mensual suele ser mayor cuanto mayor es el nivel de responsabilidad en el trabajo y para acceder a estos altos puestos de seniority también se requiere haber trabajado una cierta cantidad de años; cuantos más años de trabajo haya tenido una persona más chance tendrá de haber accedido a puestos de mayor responsabilidad y por ende de mayor sueldo.

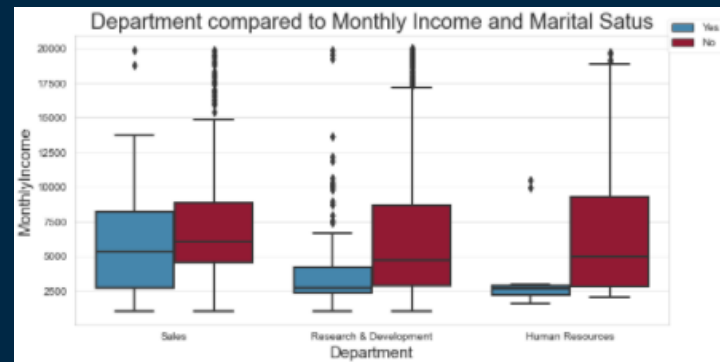
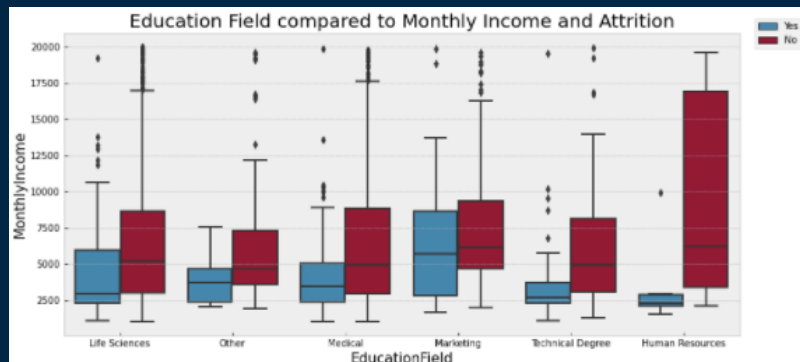
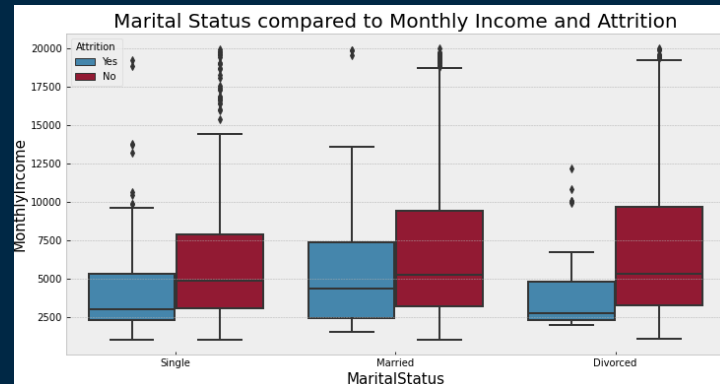
El resto de las variables muestran una falsa correlación por que cuanto mayor edad tienen los empleados mayor es la participación en la segunda variable analizada. Las variables sin implicancia temporal no muestran correlación más allá de lo mencionado.





# MULTIVARIADO

Evaluando distintas variables respecto al Monthly Income y haciendo la apertura para empleados con y sin desgaste laboral, podemos ver que por lo general los empleados con desgaste (color azul) se agrupan en los menores rangos salariales; por lo cual esta variable pareciera influir en la posibilidad de que los empleados renuncien.

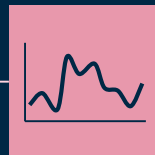


# Aplicación de Algoritmos de ML

05

# Algoritmos empleados

ÁRBOL DE  
DECISIÓN

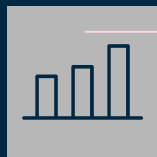


RANDOM FOREST

REGRESIÓN  
LOGÍSTICA



XGBOOST



KNN



# SUMMARY

A continuación se presenta un resumen con los principales resultados obtenidos para cada modelo aplicado al dataset con una división de 70/30 para train/test:

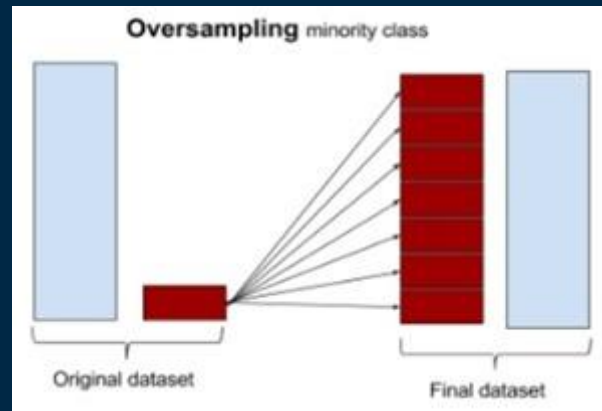
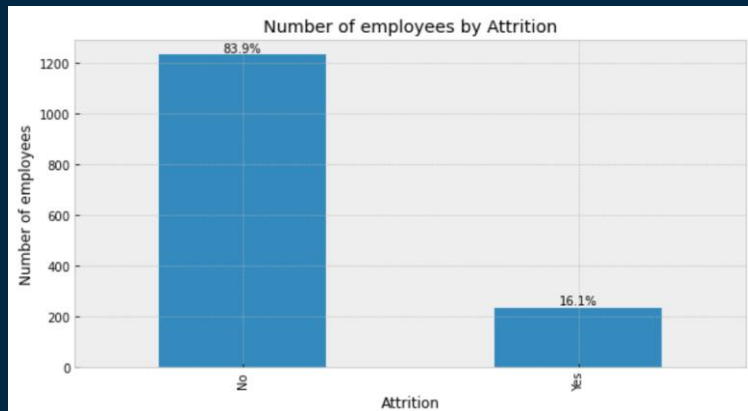
Métrica \ Modelo	arbol_1	arbol_2	forest_1	forest_2	LogReg	knn	XGboost
Accuracy	0.861678	0.809524	0.863946	0.759637	0.843537	0.836735	0.836735
Precision	0.500000	0.350650	1.000000	0.297300	0.277780	0.176470	0.351350
Recall	0.114750	0.442620	0.016390	0.016390	0.081970	0.049180	0.213110
ROC_curve	0.691390	0.696270	0.700630	0.707940	0.686500	0.510760	0.710960

\*En el apéndice se puede ver el detalle de modelos evaluados

# Oversampling

The background is a solid dark blue. It features several abstract geometric elements: a teal square in the upper left, a teal square in the lower left, a teal square in the lower right, and a teal square in the middle right. There are also orange squares in the upper left, upper right, and middle right. A pink square is located in the upper right. A vertical white line is on the far left, a vertical white line is on the far right, and a vertical white line is in the middle left. A horizontal white line is in the middle right. A horizontal orange line is at the bottom, starting from the left and ending with a teal square. A horizontal pink line is at the bottom, starting from the left and ending with a teal square.

# EVALUACIÓN CON OVERSAMPLING



Debido a que el dataset utilizado es desbalanceado, contando con 83,9% de la muestra con Attrition=NO, y 16,1% de la muestra con Attrition=YES; se consideró la opción de hacer un oversampling; llevando la cantidad de casos con respuesta *Yes* a igualar la cantidad de casos con respuesta *No*.

# EVALUACIÓN CON OVERSAMPLING

Luego se probaron todos los modelos nuevamente, pero en términos generales no se obtuvo mejora sustancial de los resultados:

División 70/30 dataset original

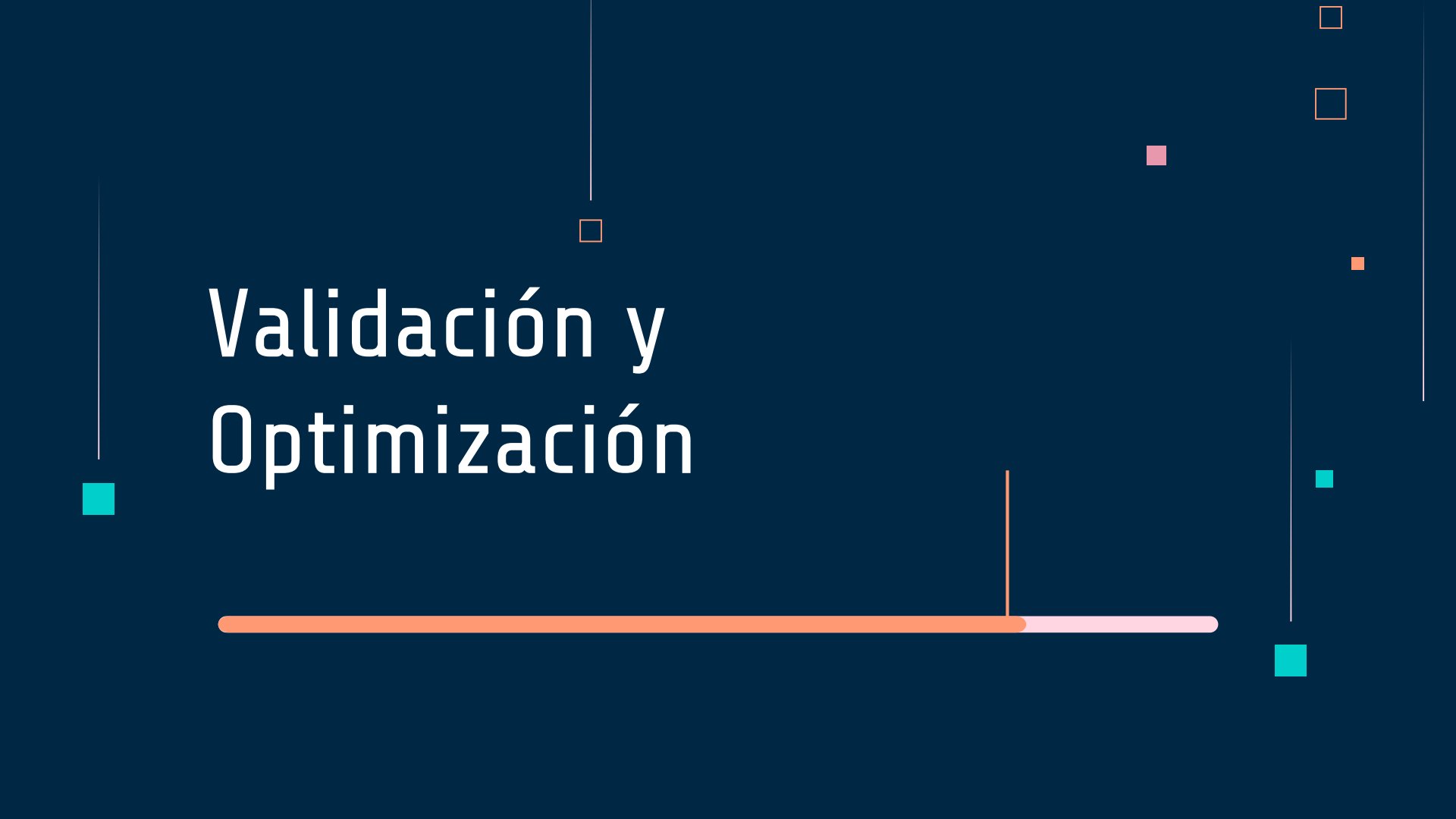
Métrica \ Modelo	arbol_1	arbol_2	forest_1	forest_2	LogReg	knn	XGboost
Accuracy	0.861678	0.809524	0.863946	0.759637	0.843537	0.836735	0.836735
Precision	0.500000	0.350650	1.000000	0.297300	0.277780	0.176470	0.351350
Recall	0.114750	0.442620	0.016390	0.016390	0.081970	0.049180	0.213110
ROC_curve	0.691390	0.696270	0.700630	0.707940	0.686500	0.510760	0.710960

Dataset Balanceado (Oversampling)

Métrica \ Modelo	arbol_1b	arbol_2b	forest_1b	forest_2b	LogRegb	knn_b	XGboost_b
Accuracy	0.646259	0.811791	NaN	NaN	0.673469	NaN	0.786848
Precision	0.176870	0.333330	NaN	NaN	0.221480	NaN	0.296300
Recall	0.426230	0.360660	NaN	NaN	0.540980	NaN	0.393440
ROC_curve	0.617470	0.641780	NaN	NaN	0.641780	NaN	0.620530

En general, al probar algunos modelos con el dataset balanceado, no se obtiene una mejoría respecto al dataset sin balancear; por lo cual es recomendable realizar la optimización de hiperparámetros

# Validación y Optimización



# RESULTADOS FINALES

Dataset  
con Original

Métrica \ Modelo	arbol_1	arbol_2	forest_1	forest_2	LogReg	knn	XGboost
Accuracy	0.861678	0.809524	0.863946	0.759637	0.843537	0.836735	0.836735
Precision	0.500000	0.350650	1.000000	0.297300	0.277780	0.176470	0.351350
Recall	0.114750	0.442620	0.016390	0.016390	0.081970	0.049180	0.213110
ROC_curve	0.691390	0.696270	0.700630	0.707940	0.686500	0.510760	0.710960

Dataset  
Oversampling

Métrica \ Modelo	arbol_1b	arbol_2b	forest_1b	forest_2b	LogRegb	knn_b	XGboost_b
Accuracy	0.646259	0.811791	NaN	NaN	0.673469	NaN	0.786848
Precision	0.176870	0.333330	NaN	NaN	0.221480	NaN	0.296300
Recall	0.426230	0.360660	NaN	NaN	0.540980	NaN	0.393440
ROC_curve	0.617470	0.641780	NaN	NaN	0.641780	NaN	0.620530

Dataset con Stratified  
Kfold y GridSearchCV

Métrica \ Modelo	arbol_optimizado	Regresión Logística optimizada	XGboost optimizado
Accuracy	0.82337	0.866848	0.847826
Precision	0.22222	0.666670	0.600000
Recall	0.03333	0.366670	0.200000
ROC_curve	0.66523	0.796750	0.795830

Se adoptó el modelo de **regresión logística** luego de realizar la **optimización de hiperparámetros**, para predecir que empleados sufrirán Attrition, ya que es el que logra los mejores indicadores

# CONCLUSIONES

06

# CONCLUSIONES

- Se realizó el EDA sobre el dataset obteniendo una descripción general de las variables que lo componen.
- Con el dataset se pudo estudiar el riesgo que tienen los empleados de abandonar la empresa por desgaste laboral (Attrition).
- La variable objetivo, desgaste laboral (Attrition), se encuentra desbalanceada. El 16,1% indica tener desgaste laboral.
- El modelo adoptado de Regresión Logística, con los óptimos parámetros, permite predecir qué empleados pudieran estar sufriendo desgaste laboral, y tomar acciones para mejorar el ambiente.
- ● El modelo predice con una exactitud del 86,68%, una precisión del 66,67% y sensibilidad de 36,67%. ■



# THANKS

