EMPLOYEE ATTRITION

Victoria Vallejo
L. 61834

01 DATA Y OBJETIVOS

IBM Employee Attrition Data Set.

02 EDA

Exploración de variables.

03 FEATURE ENGINEERING

SMOTE.

04 MODELOS

XGBoost, Random Forest,

05 CONCLUSIÓN

Comparación de modelos.

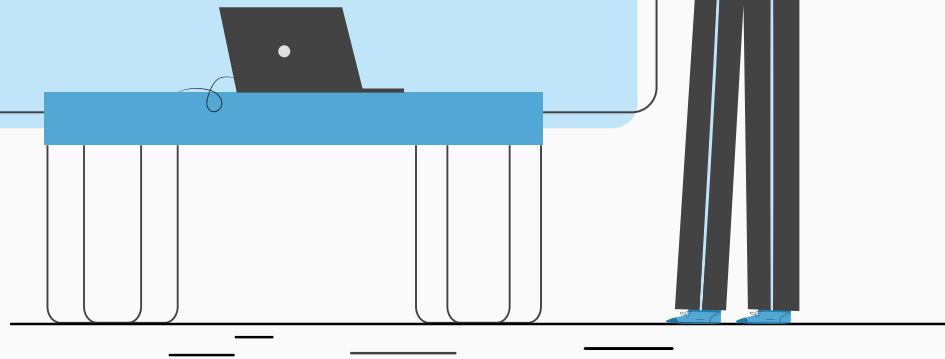
06 PASOS A SEGUIR

Para la retención de talentos.

01.
DATA Y OBJETIVOS

EMPLOYEE ATTRITION

35 VARIABLES
1470 REGISTROS



```
def cant_missings (df):
  cant_missing = df.isnull().sum()
  missings = cant_missing.sum()
  return missings

print(cant_missings(df))
```

```
0
```

La base no presenta Missing Values

VARIABLES

CATEGÓRICAS

Attrition ⟶ TARGET
BusinessTravel
Department
Education
EducationField
EnvironmentSatisfaction
Gender
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
Over18
OverTime
PerformanceRating
RelationshipSatisfaction
StockOptionLevel
WorkLifeBalance

DISCRETAS

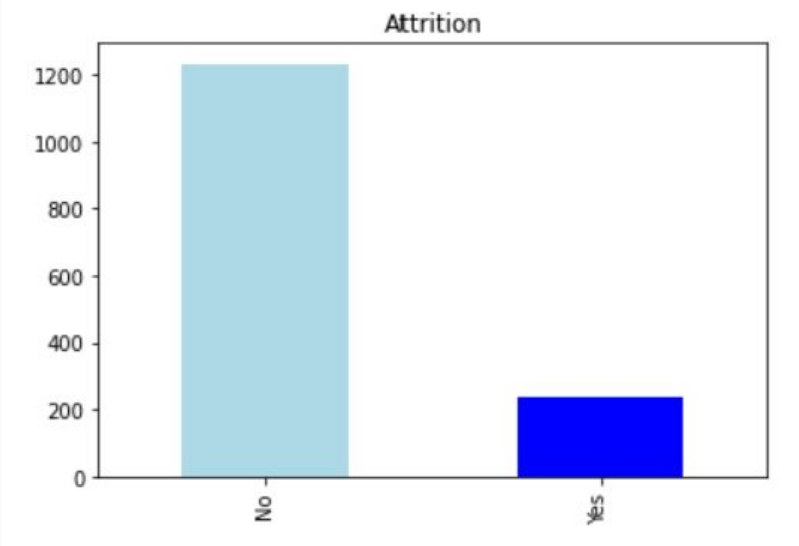
Age
DailyRate
DistanceFromHome
EmployeeCount
EmployeeNumber
HourlyRate
MonthlyIncome
MonthlyRate
NumCompaniesWorked
PercentSalaryHike
StandardHours
TotalWorkingYears
TrainingTimesLastYear
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager

02.
EDA

VARIABLE TARGET: ATTRITION

Porcentaje

No	83.877551
Yes	16.122449



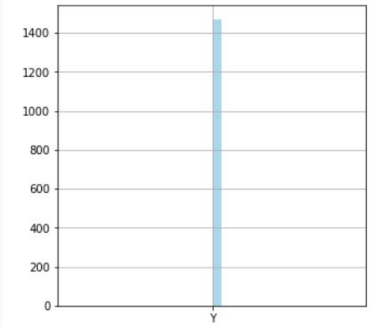
Data NO balanceada!

Distribuciones

Over18

```
df.Over18.unique()
```

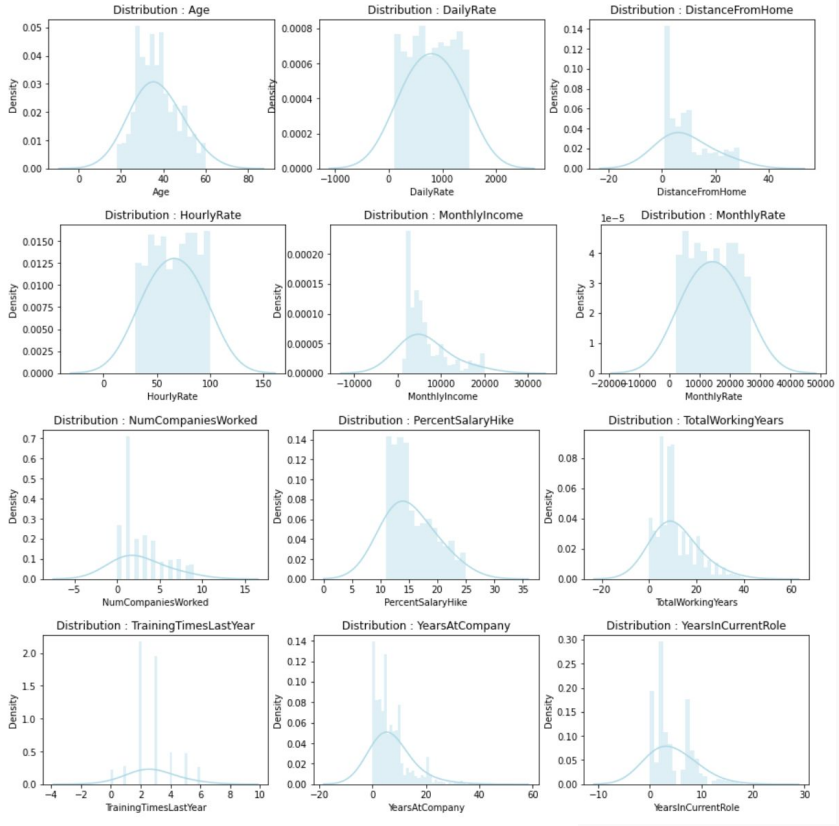
```
array(['Y'], dtype=object)
```



EmployeeNumber

Es el ID

Elimino StandardHours, EmployeeCount, Over18, EmployeeNumber → Quedan 31 variables



OUTLIERS

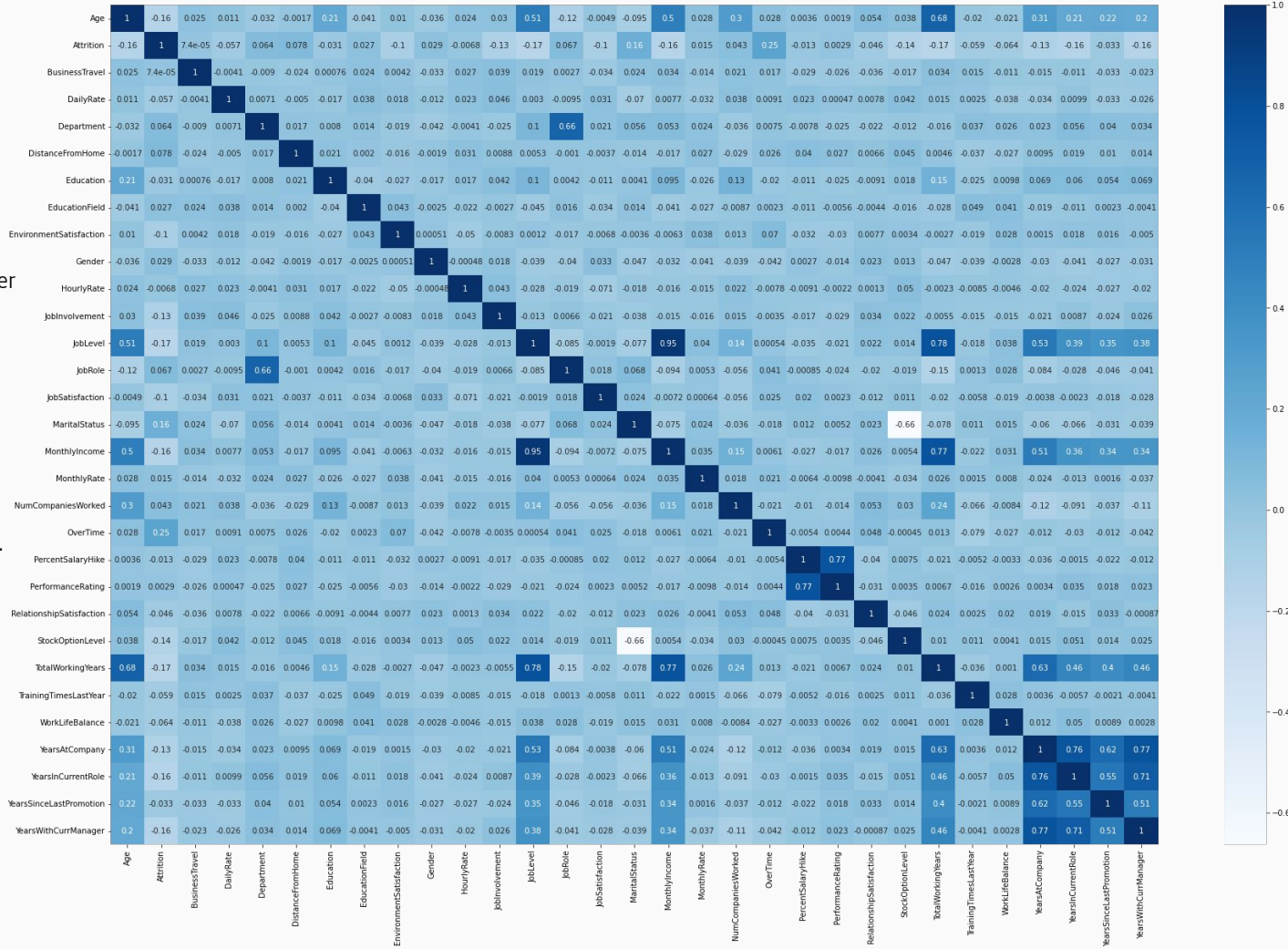
No hay outliers en la base de datos ya que la causa por la cuál varían la grán mayoría de las variables es por cada tipo de trabajo el cuál depende del nivel de trabajo.

No se encuentran valores negativos irreales.

CORRELACIONES

0.95: Job Level - Monthly Income
0.78: Job Level - TotalWorkingYears
0.78: Job Level - TotalWorkingHours
0.78: Monthly Income - TotalWorkingYears
0.77: YearsAtCompany - YearsWithCurrManager
0.76: YearsInCurrentRole - YearsAtCompany

Las variables con fuerte relación se encuentran todas relacionadas entre sí: Al tener más años de experiencia en la compañía, sube el puesto de trabajo consecuentemente aumentando el salario mensual como las horas de trabajo a su vez.



MEDIA BY ATTRITION

- La mayor diferencia se observa en MonthlyIncome (Salario), con una diferencia de un 70%.
- Los empleados que se quedan presentan significativos mayores valores de: JobLevel, DailyRate, TotalWorkingYears, YearsAtCompany y YearsInCurrentRole

Media : Empleados no Retenidos

	mean
Age	33.61
DailyRate	750.36
DistanceFromHome	10.63
Education	2.84
EnvironmentSatisfaction	2.46
HourlyRate	65.57
JobInvolvement	2.52
JobLevel	1.64
JobSatisfaction	2.47
MonthlyIncome	4787.09
MonthlyRate	14559.31
NumCompaniesWorked	2.94
PercentSalaryHike	15.10
PerformanceRating	3.16
RelationshipSatisfaction	2.60
StockOptionLevel	0.53
TotalWorkingYears	8.24
TrainingTimesLastYear	2.62
WorkLifeBalance	2.66
YearsAtCompany	5.13
YearsInCurrentRole	2.90
YearsSinceLastPromotion	1.95
YearsWithCurrManager	2.85

Media : Empleados Retenidos

	mean
Age	37.56
DailyRate	812.50
DistanceFromHome	8.92
Education	2.93
EnvironmentSatisfaction	2.77
HourlyRate	65.95
JobInvolvement	2.77
JobLevel	2.15
JobSatisfaction	2.78
MonthlyIncome	6832.74
MonthlyRate	14265.78
NumCompaniesWorked	2.65
PercentSalaryHike	15.23
PerformanceRating	3.15
RelationshipSatisfaction	2.73
StockOptionLevel	0.85
TotalWorkingYears	11.86
TrainingTimesLastYear	2.83
WorkLifeBalance	2.78
YearsAtCompany	7.37
YearsInCurrentRole	4.48
YearsSinceLastPromotion	2.23
YearsWithCurrManager	4.37

OBJETIVOS DEL TRABAJO

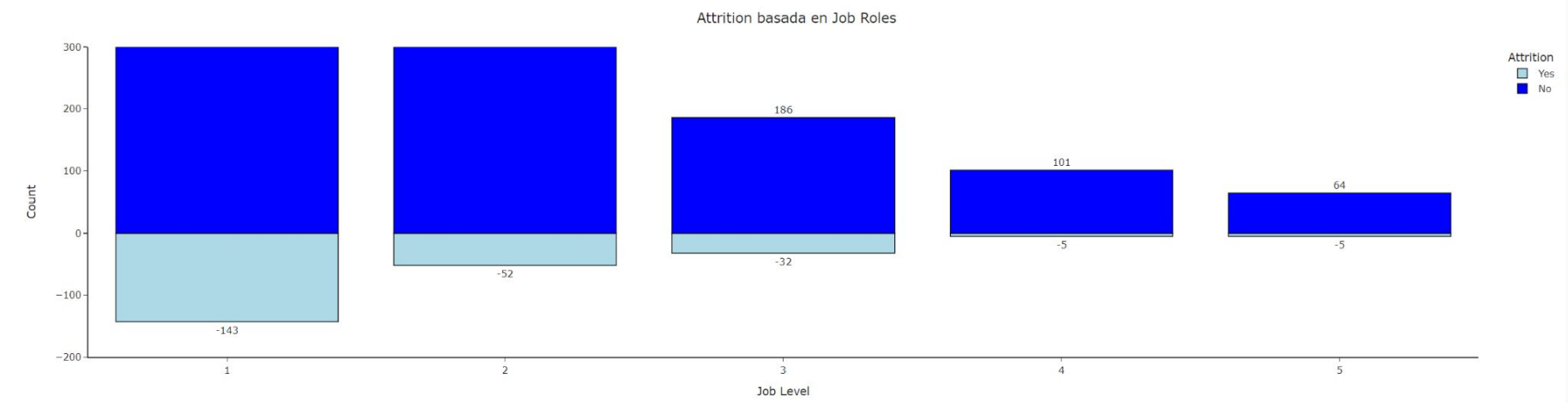
1. Chequear la dependencia de las variables Job Level y Attrition

2. Encontrar aquellas variables que determinan que una persona abandone su puesto

3. Es posible predecir el abandono de un empleado?

4. Proveer un plan estratégico para la retención de talentos.

JOB LEVEL Y ATTRITION



Chi Q Test $\alpha = 0.05$

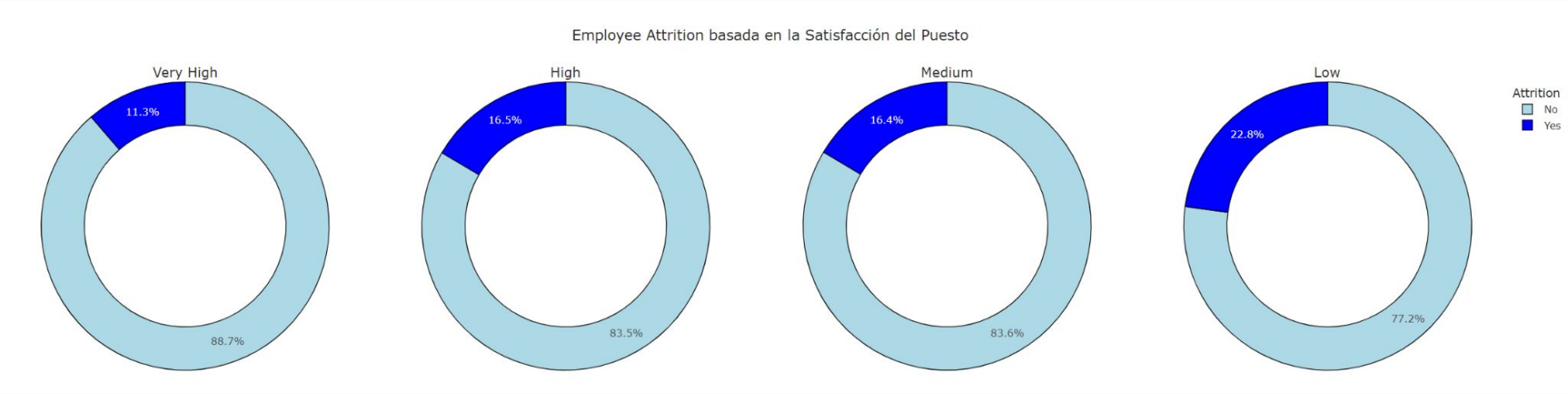
H0: No hay relación significativa entre Job Level y Attrition (independencia)

HA: Hay una relación significativa entre Job Level y Attrition (no independientes)

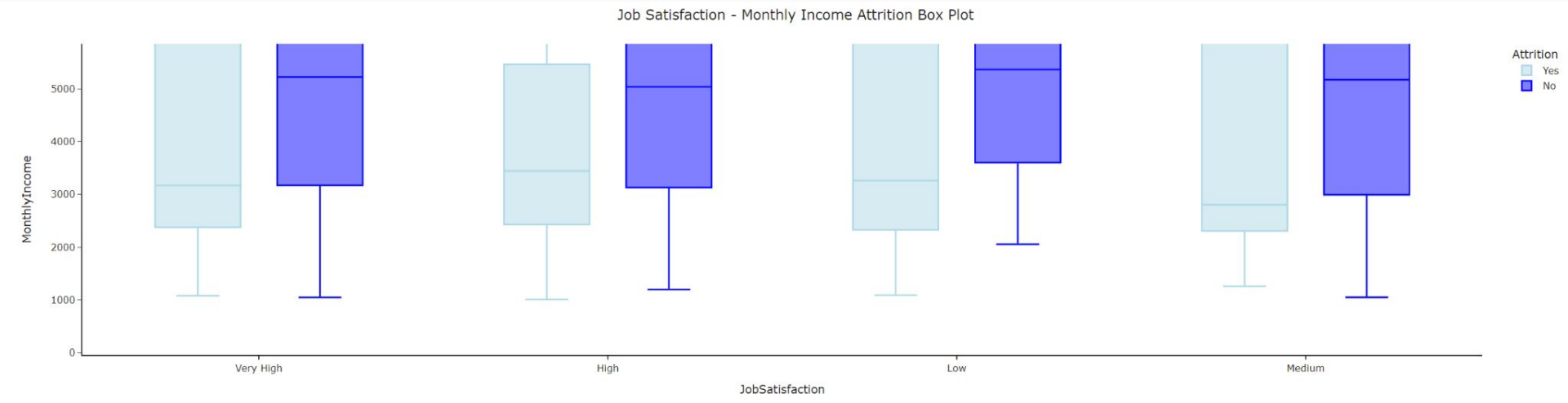
```
Chi2_score: 72.52901310667391, Degrees of freedom: 4, p-value: 6.634684715458909e-15
```

Rechazo H0, p-value < 0.5 por lo tanto las variables están significativamente relacionadas. La variable Job Level va a ser importante al momento de realizar los modelos

A su vez, se observa una relación entre Job Level, Monthly Income y Attrition



Es posible afirmar que a mayor Job Satisfaction, hay un menor porcentaje de Attrition.


Job Satisfaction - Monthly Income Attrition Box Plot
MonthlyIncome
0
1000
2000
3000
4000
5000
Very High
High
Low
Medium
JobSatisfaction
Attrition
Yes
No

Luego de los 6 años bajo el mando de un manager se observa que una gran cantidad abandonan el empleo.

03.
FEATURE ENGINEERING

División de variables en Categóricas y Discretas.
Label Encoder para las 17 variables Categóricas

```
Transformación encoders:

100%|██████████| 17/17 [00:00<00:00, 368.38it/s]Attrition  :  [1 0]  =  ['Yes' 'No']
BusinessTravel  :  [2 1 0]  =  ['Travel_Rarely' 'Travel_Frequently' 'Non-Travel']
Department  :  [2 1 0]  =  ['Sales' 'Research & Development' 'Human Resources']
EducationField  :  [1 4 3 2 5 0]  =  ['Life Sciences' 'Other' 'Medical' 'Marketing' 'Technical Degree'
 'Human Resources']
Gender  :  [0 1]  =  ['Female' 'Male']
JobRole  :  [7 6 2 4 0 3 8 5 1]  =  ['Sales Executive' 'Research Scientist' 'Laboratory Technician'
 'Manufacturing Director' 'Healthcare Representative' 'Manager'
 'Sales Representative' 'Research Director' 'Human Resources']
MaritalStatus  :  [2 1 0]  =  ['Single' 'Married' 'Divorced']
OverTime  :  [1 0]  =  ['Yes' 'No']
```

LABEL ENCODER

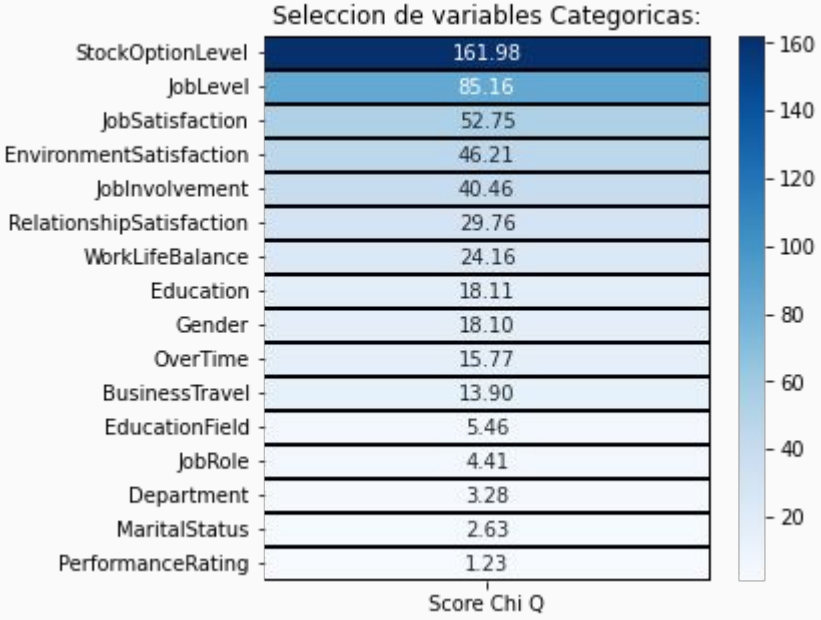
SMOTE

Técnica estadística de sobremuestreo de minorías sintéticas para aumentar el número de casos de un conjunto de datos de forma equilibrada.
SMOTE toma todo el conjunto de datos como una entrada, pero solo aumenta el porcentaje de los casos minoritarios.
P = 0.80

```
Counter({1: 986, 0: 1233})
```

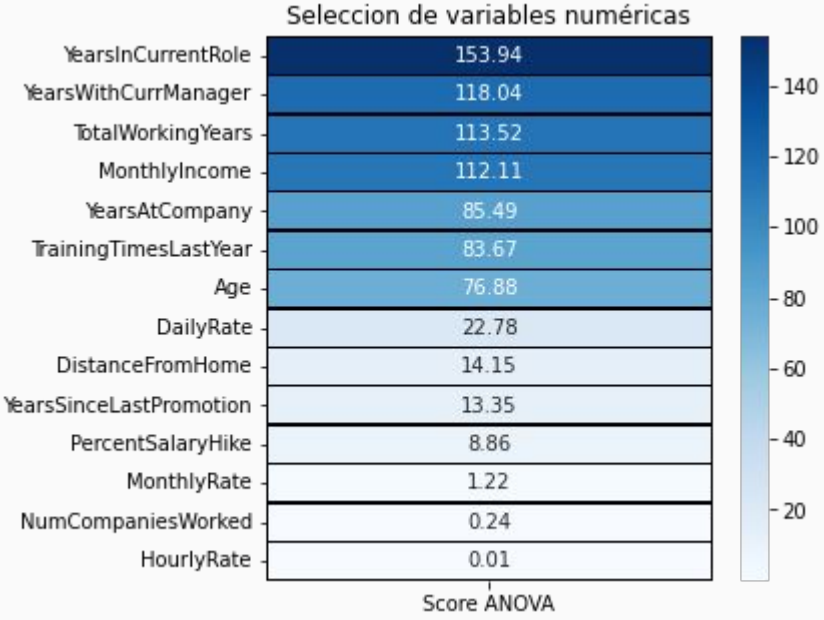
SMOTE	Previo	Post
Attrition YES	237	986
Attrition NO	1233	1233

CHI Q SCORE



Elimino: Performance Rating, JobRole, Marital Status, EducationField, Department, BusinessTravel y Gender

F - ANOVA



Elimino: NumCompaniesWorked, HourlyRate, MonthlyRate, PercentSalaryHike, YearsSinceLastPromotion, DistanceFromHome, DailyRate

VARIABLES A UTILIZAR PARA MODELOS

Age
Education
EnvironmentSatisfaction
JobInvolvement
JobLevel
JobSatisfaction
MonthlyIncome
OverTime
RelationshipSatisfaction
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsWithCurrManager

04.
MODELOS

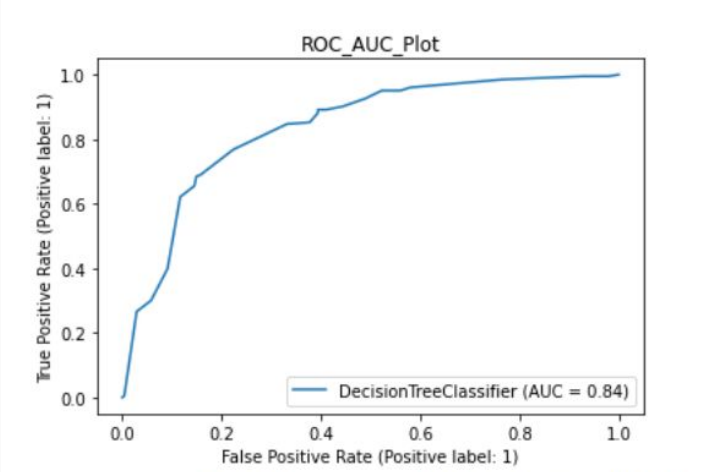
TRAIN Y TEST 80 - 20

```
x_train, x_test, y_train, y_test = train_test_split(Col, Lcateg, test_size = 0.20, random_state = 2)
```

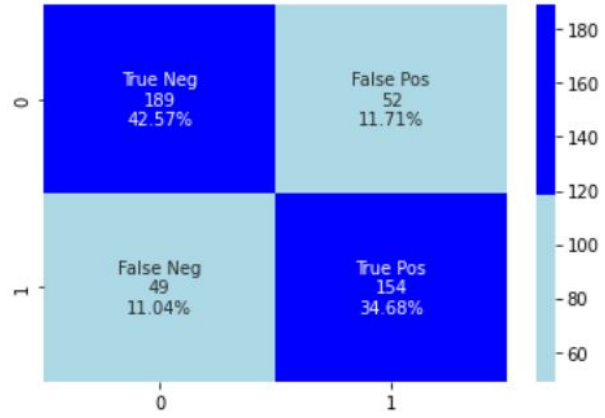
DECISION TREE CLASSIFIER

classifier_dt = DecisionTreeClassifier(random_state = 100,max_depth = 5,min_samples_leaf = 2)

```
Cross Validation Score :  81.14%
ROC_AUC Score :  77.14%
```



	precision	recall	f1-score	support
0	0.79	0.78	0.79	241
1	0.75	0.76	0.75	203
accuracy			0.77	444
macro avg	0.77	0.77	0.77	444
weighted avg	0.77	0.77	0.77	444

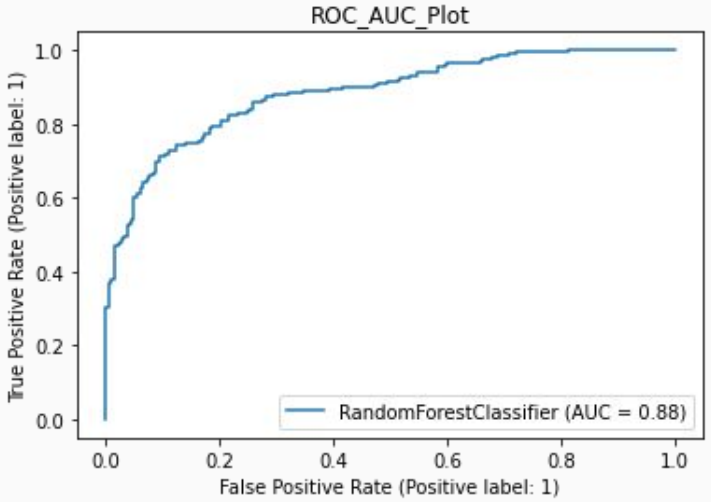


AUC 77.14% < 80% casi posee poder predictivo, pero le falta.
Cross validation bueno de 81.14%
Accuracy 77,25%

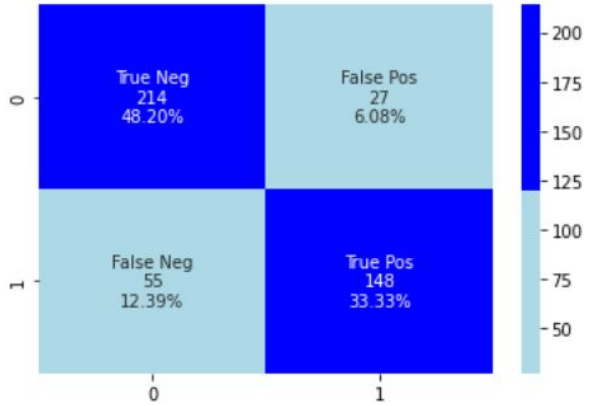
RANDOM FOREST

classifier_rf = RandomForestClassifier(max_depth = 4,random_state = 0)

```
Cross Validation Score :  86.49%
ROC_AUC Score :  80.85%
```



	precision	recall	f1-score	support
0	0.80	0.89	0.84	241
1	0.85	0.73	0.78	203
accuracy			0.82	444
macro avg	0.82	0.81	0.81	444
weighted avg	0.82	0.82	0.81	444

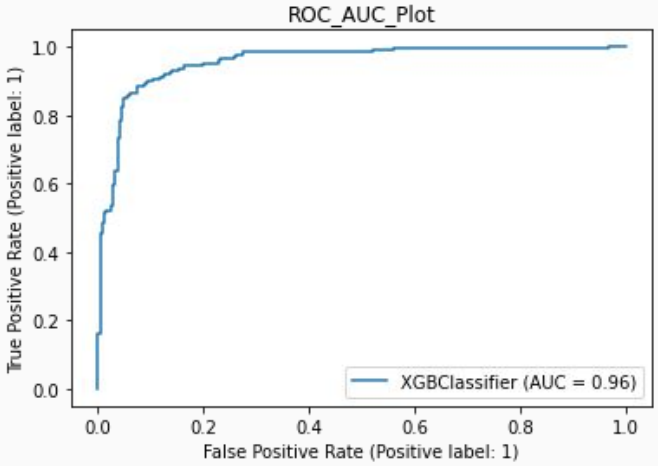


AUC 80.85% > 80% posee poder predictivo.
Cross validation bueno de 86.49%
Accuracy 81.53%

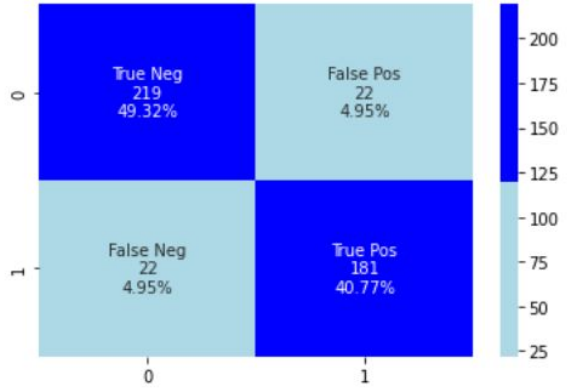
XGBOOST

classifier_xgb = XGBClassifier(learning_rate= 0.01,max_depth = 3,n_estimators = 3000)

```
Cross Validation Score :  92.41%
ROC_AUC Score :  90.02%
```



	precision	recall	f1-score	support
0	0.91	0.91	0.91	241
1	0.89	0.89	0.89	203
accuracy			0.90	444
macro avg	0.90	0.90	0.90	444
weighted avg	0.90	0.90	0.90	444



AUC 90.02% > 80% posee alto poder predictivo.
Cross validation muy bueno de 92.41%
Accuracy 90.09%

05.
CONCLUSIÓN

COMPARACIÓN DE MODELOS

	DECISION TREE CLASSIFIER	RANDOM FOREST	**XGBOOST**
Accuracy	77.25%	81.53%	90.09%
Cross Validation	81.14%	86.49%	92.41%
Score ROC AUC	77.14%	80.85%	90.02%

El modelo óptimo para la predicción de Attrition es XGBOOST.

06.
PASOS
A SEGUIR

PLAN ESTRATÉGICO en base a las variables más fuertes respecto a Attrition

Monthly Income: las personas con salarios más altos tienen menos probabilidades de abandonar la empresa. Se podría recopilar información sobre los sueldos de referencia de la industria en el mercado actual para determinar si la empresa ofrece salarios competitivos. De ser menores a los actuales, se deberían aumentar los sueldos para disminuir el abandono del empleo.

Job Levels: Como fue observado, los empleados en menores Job Levels suelen abandonar más el trabajo y en ellos es en lo que más se invierte tiempo principalmente al entrenarlos. Podría crearse un plan de acuerdo a cada rango etáreo al tener distintos intereses con el fin de aumentar el Job Satisfaction, por ejemplo para empleados jóvenes entre 25 a 35 años dar la posibilidad de home office 3 veces a la semana.

TotalWorkingYears: los empleados con más experiencia tienen menos probabilidades de irse. Los empleados que tienen entre 5 y 8 años de experiencia deben identificarse como los que potencialmente tienen un mayor riesgo de irse.

YearsWithCurrManager: una gran cantidad de personas que se van se van 6 años después que sus Gerentes Actuales. Al usar los detalles del Gerente de línea para cada empleado, se puede determinar qué Gerente tuvo la mayor cantidad de renuncias de empleados durante el último año. Se pueden usar varias métricas para determinar si se deben tomar medidas con un gerente de línea:

- número de empleados a cargo de gerentes que muestran altas tasas de rotación: indicaría que es posible que sea necesario revisar la estructura de la organización para mejorar la eficiencia con menos empleados bajo un mismo gerente.
- número de años que el gerente de línea estuvo en una posición particular: puede indicar que los empleados necesitan capacitación gerencial o que se les asigne un mentor.
- patrones en los empleados que renunciaron: puede indicar patrones recurrentes en los empleados que se van, se pueden tomar medidas.

PLAN ESTRATÉGICO en base a las variables más fuertes respecto a Attrition

“Plan de Retención" estratégico para cada grupo etáreo y nivel de puesto.
Además de los pasos sugeridos para cada variable específica mencionada, se pueden iniciar reuniones presenciales entre un representante de recursos humanos y los empleados para discutir las condiciones de trabajo. También se podría realizar una reunión con el Gerente de línea de tales empleados para discutir el ambiente de trabajo dentro del equipo y si se pueden tomar medidas para mejorarlo.

GRACIAS !