

Proyecto de Data Science



Abandono Laboral

Análisis del conjunto de datos de Abandono de Trabajo

Alumno: Leonard Quiame

Introducción.

El presente estudio tiene como objetivo analizar el DataSet "AbandonoEmpleados.csv" para explorar posibles relaciones entre variables y comprender mejor los factores que podrían influir en el abandono de empleados en una empresa. Utilizando visualizaciones en Python, se generaron gráficos y se calcularon estadísticas descriptivas para varias variables de interés.

En primer lugar, se planteó la hipótesis de que la cantidad de años que un empleado ha estado en la compañía (anos_compania) podría estar relacionada con su propensión a abandonar el trabajo. Para investigar esto, se creó un histograma de (anos_compañía), que mostró la distribución de la variable. A partir de este histograma, se observó que la mayoría de los empleados tenían menos de 10 años de experiencia en la empresa, lo que podría sugerir una alta rotación laboral. Sin embargo, se necesitarían análisis adicionales para determinar si existe una relación directa entre la antigüedad en la empresa y el abandono de empleados.

Sin embargo, es importante tener en cuenta que estos hallazgos son exploratorios y requieren un análisis más profundo y consideración de otros factores para obtener conclusiones más sólidas.

El estudio sugiere que variables como la antigüedad en la empresa, el nivel de acciones y la capacitación podrían estar relacionadas con el abandono de empleados y la satisfacción laboral.



En esta presentación, exploraremos el desafío del abandono laboral en la organización o empresa. Se utilizará herramientas de visualización en Python para analizar y comprender por qué algunos empleados deciden abandonar sus trabajos, mientras otros permanecen comprometidos. A través de gráficos como histogramas y gráficos de barras, identificaremos tendencias y relaciones clave, como la satisfacción laboral y la formación de los empleados. El objetivo es proporcionar información útil para la toma de decisiones en recursos humanos y mejorar la retención de empleados.

Además se tiene de conocimientos características de este estudio que se ha desarrollado en el notebook que han servido para conocer el desarrollo como son las siguientes:

- ✓ Contexto comercial.
- ✓ Problema comercial.
- ✓ Contexto analítico.

Objetivos.



El gran objetivo del proyecto en el que vas a trabajar es reducir la fuga de empleados de la empresa. Para ello durante esta semana vas a trabajar en 3 grandes cosas:

- ✓ Entender y cuantificar el problema desde el punto de vista de negocio.
- ✓ Desarrollar un sistema automatizado de machine learning que identifique a los empleados que están en mayor riesgo de fuga.
- ✓ Comunicar los resultados de forma exitosa a la dirección.

PREGUNTAS/HIPÓTESIS QUE QUEREMOS RESOLVER MEDIANTE EL ANÁLISIS DE DATOS

❑ Satisfacción Laboral y Abandono.

mayoría de los empleados experimenta una alta satisfacción con sus compañeros de trabajo, lo que podría estar asociado a una menor tasa de abandono laboral.

❑ Relación entre Acciones y Satisfacción.

Existe la hipótesis de que los empleados con un mayor nivel de acciones experimentan una mayor satisfacción con sus compañeros, lo que potencialmente influye en su decisión de permanecer en la empresa.

❑ Experiencia Laboral.

La mayoría de los empleados puede tener una experiencia laboral relativamente corta, lo que podría contribuir a una mayor rotación de empleados.

❑ Formación y Retención.

Se plantea que la cantidad de formaciones recibidas en el último año está correlacionada con una menor propensión al abandono laboral. Los empleados que reciben más formaciones podrían estar más satisfechos y comprometidos.

❑ Influencia de Departamentos y Puestos.

Algunos departamentos o puestos específicos pueden presentar tasas de rotación más altas debido a factores relacionados con las tareas o el entorno laboral.

❑ Satisfacción y Antigüedad.

Los empleados con mayor antigüedad en la empresa pueden experimentar una mayor satisfacción laboral debido a su familiaridad con la organización.

Datos

CARGA DE LOS DATOS.

```
[4] df = pd.read_csv('AbandonoEmpleados.csv', sep = ';', index_col= 'id', na_values='#N/D')
df
```

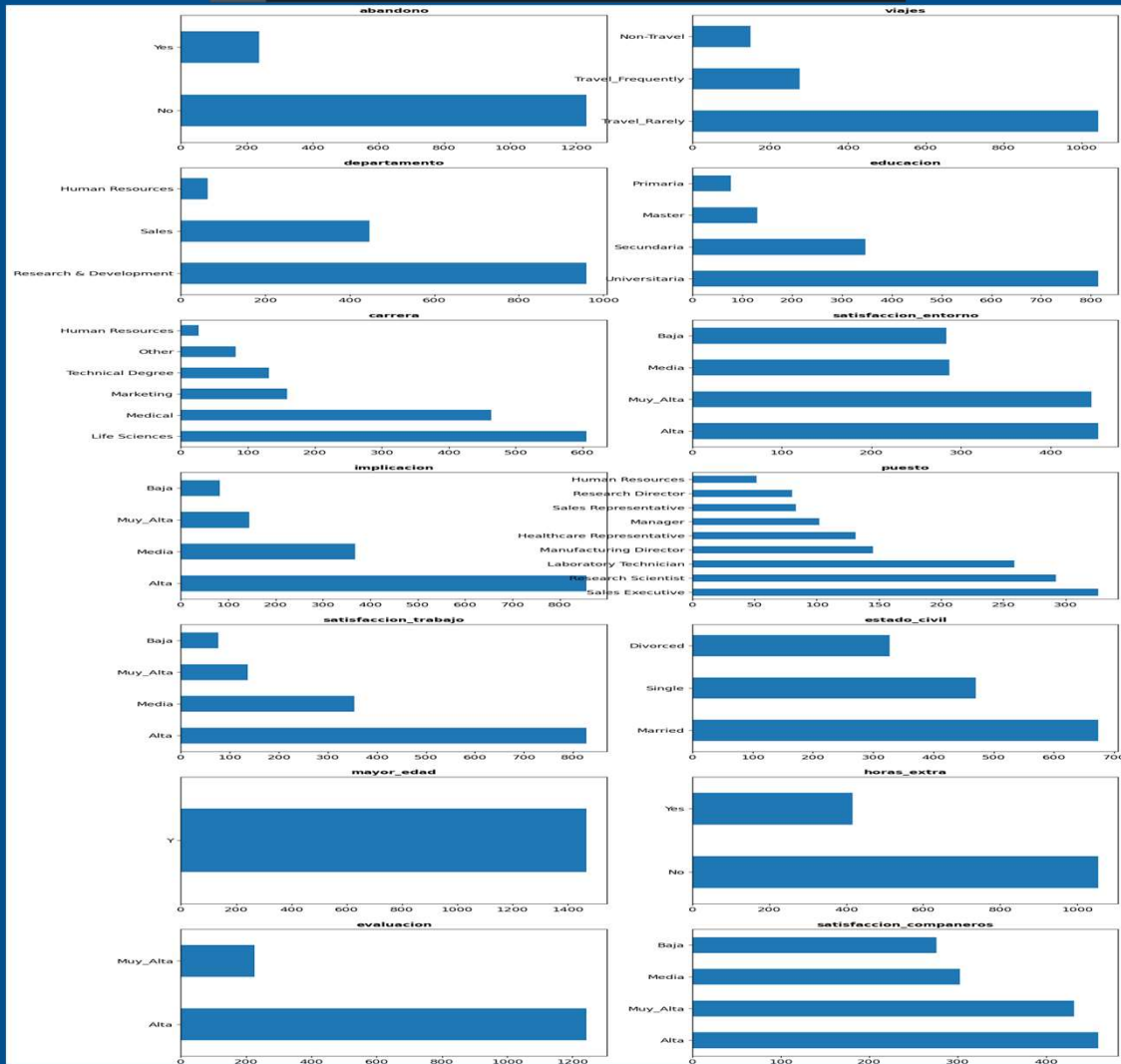
	edad	abandono	viajes	departamento	distancia_casa	educacion	carrera	empleados	satisfaccion_entorno	sexo	...	satisfaccion_companeros	horas_
id													
1	41	Yes	Travel_Rarely	Sales	1	Universitaria	Life Sciences	1	Media	3.0	...	Baja	
2	49	No	Travel_Frequently	Research & Development	8	Secundaria	Life Sciences	1	Alta	2.0	...	Muy_Alta	
4	37	Yes	Travel_Rarely	Research & Development	2	Secundaria	Other	1	Muy_Alta	2.0	...	Media	
5	33	No	Travel_Frequently	Research & Development	3	Universitaria	Life Sciences	1	Muy_Alta	3.0	...	Alta	
7	27	No	Travel_Rarely	Research & Development	2	Universitaria	Medical	1	Baja	3.0	...	Muy_Alta	
...	
2061	36	No	Travel_Frequently	Research & Development	23	Master	Medical	1	Alta	4.0	...	Alta	
2062	39	No	Travel_Rarely	Research & Development	6	Secundaria	Medical	1	Muy_Alta	2.0	...	Baja	
2064	27	No	Travel_Rarely	Research & Development	4	Master	Life Sciences	1	Media	4.0	...	Media	
2065	49	No	Travel_Frequently	Sales	2	Secundaria	Medical	1	Muy_Alta	NaN	...	Muy_Alta	
2068	34	No	Travel_Rarely	Research & Development	8	NaN	Medical	1	Media	4.0	...	Baja	

1470 rows x 31 columns

ANÁLISIS EXPLORATORIO DE DATOS

EDA CATEGORICAS

```
graficos_eda_categoricos(df.select_dtypes('O'))
```



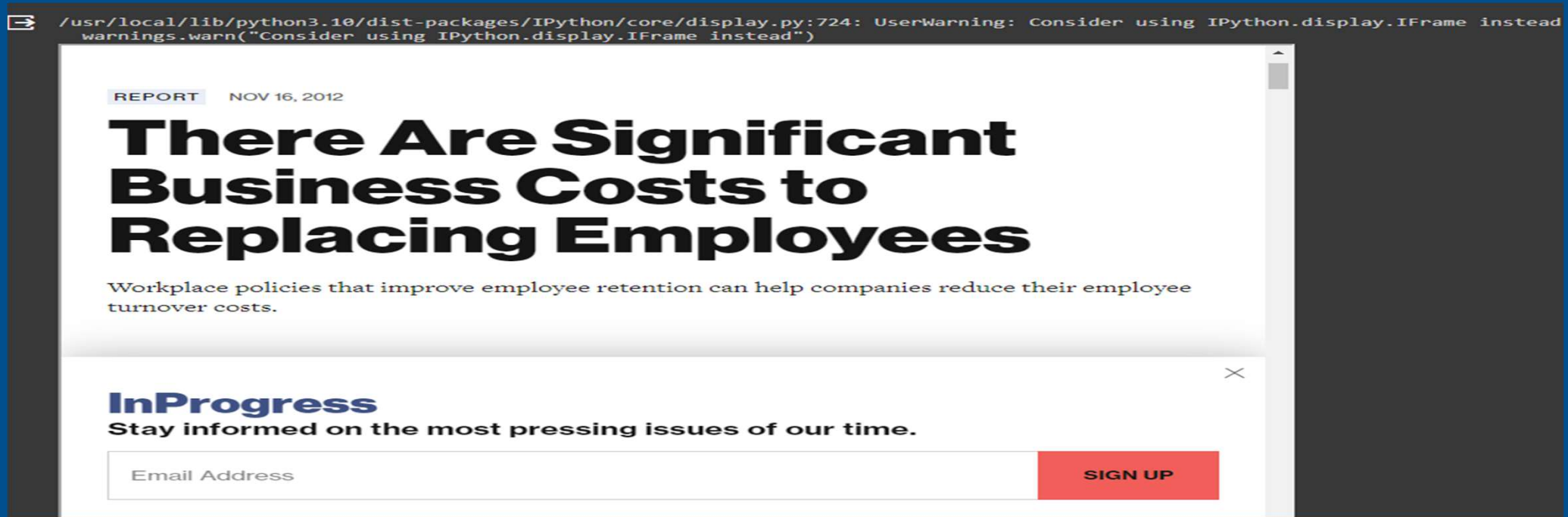
EDA NUMERICAS

```
estadisticos_cont(df.select_dtypes('number'))
```

	count	mean	median	std	min	25%	50%	75%	max
edad	1470.0	36.923810	36.0	9.135373	18.0	30.0	36.0	43.0	60.0
distancia_casa	1470.0	9.192517	7.0	8.106864	1.0	2.0	7.0	14.0	29.0
empleados	1470.0	1.000000	1.0	0.000000	1.0	1.0	1.0	1.0	1.0
sexo	1271.0	2.727773	3.0	0.720788	1.0	2.0	3.0	3.0	4.0
nivel_laboral	1470.0	2.063946	2.0	1.106940	1.0	1.0	2.0	3.0	5.0
salario_mes	1470.0	6502.931293	4919.0	4707.956783	1009.0	2911.0	4919.0	8379.0	19999.0
num_empresas_anteriores	1470.0	2.693197	2.0	2.498009	0.0	1.0	2.0	4.0	9.0
incremento_salario_porc	1470.0	15.209524	14.0	3.659938	11.0	12.0	14.0	18.0	25.0
horas_quincena	1470.0	80.000000	80.0	0.000000	80.0	80.0	80.0	80.0	80.0
nivel_acciones	1470.0	0.793878	1.0	0.852077	0.0	0.0	1.0	1.0	3.0
anos_experiencia	1470.0	11.279592	10.0	7.780782	0.0	6.0	10.0	15.0	40.0
num_formaciones_ult_ano	1470.0	2.799320	3.0	1.289271	0.0	2.0	3.0	3.0	6.0
anos_compania	1470.0	7.008163	5.0	6.126525	0.0	3.0	5.0	9.0	40.0
anos_desde_ult_promocion	1470.0	2.187755	1.0	3.222430	0.0	0.0	1.0	3.0	15.0
anos_con_manager_actual	1470.0	4.123129	3.0	3.568136	0.0	2.0	3.0	7.0	17.0

Bases para studio de datos para el estúdio.

```
/usr/local/lib/python3.10/dist-packages/IPython/core/display.py:724: UserWarning: Consider using IPython.display.IFrame instead
warnings.warn("Consider using IPython.display.IFrame instead")
```



REPORT NOV 16, 2012

There Are Significant Business Costs to Replacing Employees

Workplace policies that improve employee retention can help companies reduce their employee turnover costs.

InProgress
Stay informed on the most pressing issues of our time.

Email Address

SIGN UP

¿Cual es el impacto económico de este problema?

Según el estudio "Cost of Turnover" del Center for American Progress:

- * El coste de la fuga de los empleados que ganan menos de 30000 es del 16,1% de su salario.
- * El coste de la fuga de los empleados que ganan entre 30000-50000 es del 19,7% de su salario.
- * El coste de la fuga de los empleados que ganan entre 50000-75000 es del 20,4% de su salario.
- * El coste de la fuga de los empleados que ganan más de 75000 es del 21% de su salario.

Ingeniería de Características

```
ingeniería de características
df['incremento'] = df['incremento_salario_porc'] * 2

from imblearn.over_sampling import RandomOverSampler

oversampler = RandomOverSampler(sampling_strategy='minority')
X_resampled, y_resampled = oversampler.fit_resample(X, y)
```

```
Mejores hiperparámetros: {'C': 0.001, 'penalty': 'l2'}
/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_validation.py:378: FitFailedWarning:
```

```
30 fits failed out of a total of 60.
The score on these train-test partitions for these parameters will be set to nan.
If these failures are not expected, you can try to debug them by setting error_score='raise'.
```

```
Below are more details about the failures:
```

```
-----
30 fits failed with the following error:
```

```
Traceback (most recent call last):
```

```
File "/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_validation.py", line 686, in _fit_and_score
    estimator.fit(X_train, y_train, **fit_params)
```

```
File "/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py", line 1162, in fit
```

```
    solver = _check_solver(self.solver, self.penalty, self.dual)
```

```
File "/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py", line 54, in _check_solver
```

```
    raise ValueError(
```

```
ValueError: Solver lbfgs supports only 'l2' or 'none' penalties, got l1 penalty.
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_search.py:952: UserWarning:
```

```
One or more of the test scores are non-finite: [          nan  0.59821015          nan  0.59804286          nan  0.59803762
          nan  0.59803762          nan  0.59803762          nan  0.59803762]
```

MODELADO

cat_ohe

	viajes_Non-Travel	viajes_Travel_Frequently	viajes_Travel_Rarely	departamento_Human Resources	departamento_Research & Development	departamento_Sales	educacion_Master	educacion_Primary	abandono
0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0
1	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0
2	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0
3	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0
4	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0
...
1465	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0
1466	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0
1467	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0
1468	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0
1469	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0
1470 rows x 44 columns									

Diseño de la modelización.

Separación predictoras y target:

```
x = df_ml.drop(columns='abandono')
y = df_ml['abandono']
```

Separación train y test.

```
from sklearn.model_selection import train_test_split
```

```
train_x, test_x, train_y, test_y = train_test_split(x, y,
test_size = 0.3)
```

```
# Evaluación
```

```
from sklearn.metrics import roc_auc_score
```

```
roc_auc_score(test_y,pred)
```

```
0.6979343103321305
```

PREDICCIÓN Y VALIDACIÓN SOBRE TEST.

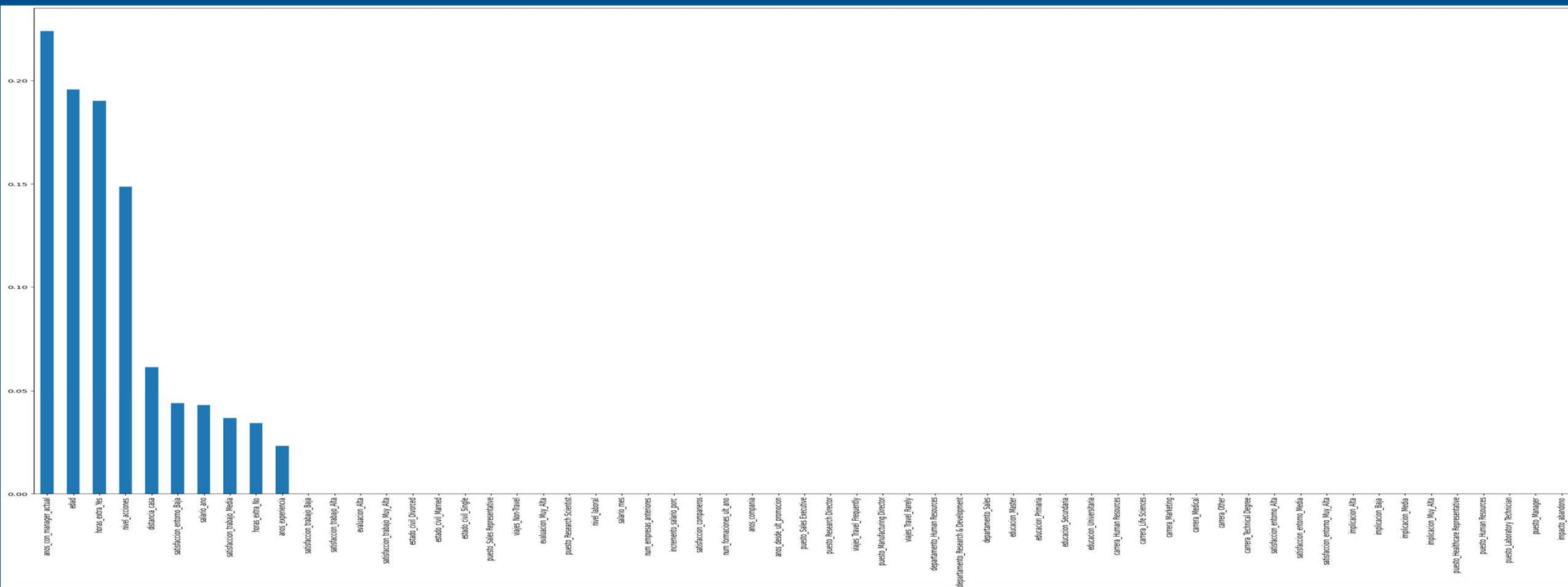
```
# Predicción
```

```
pred = ac.predict_proba(test_x)[: , 1]
pred[:20]
```

```
array([0.1120332 , 0.10071942, 0.1120332 , 0.04557641, 0.11267606,
0.04557641, 0.11267606, 0.1120332 , 0.04557641, 0.04557641, 0.04557641,
0.04557641, 0.1120332 , 0.10071942, 0.04557641, 0.1120332 , 0.20930233,
0.11267606, 0.04557641, 0.04557641])
```

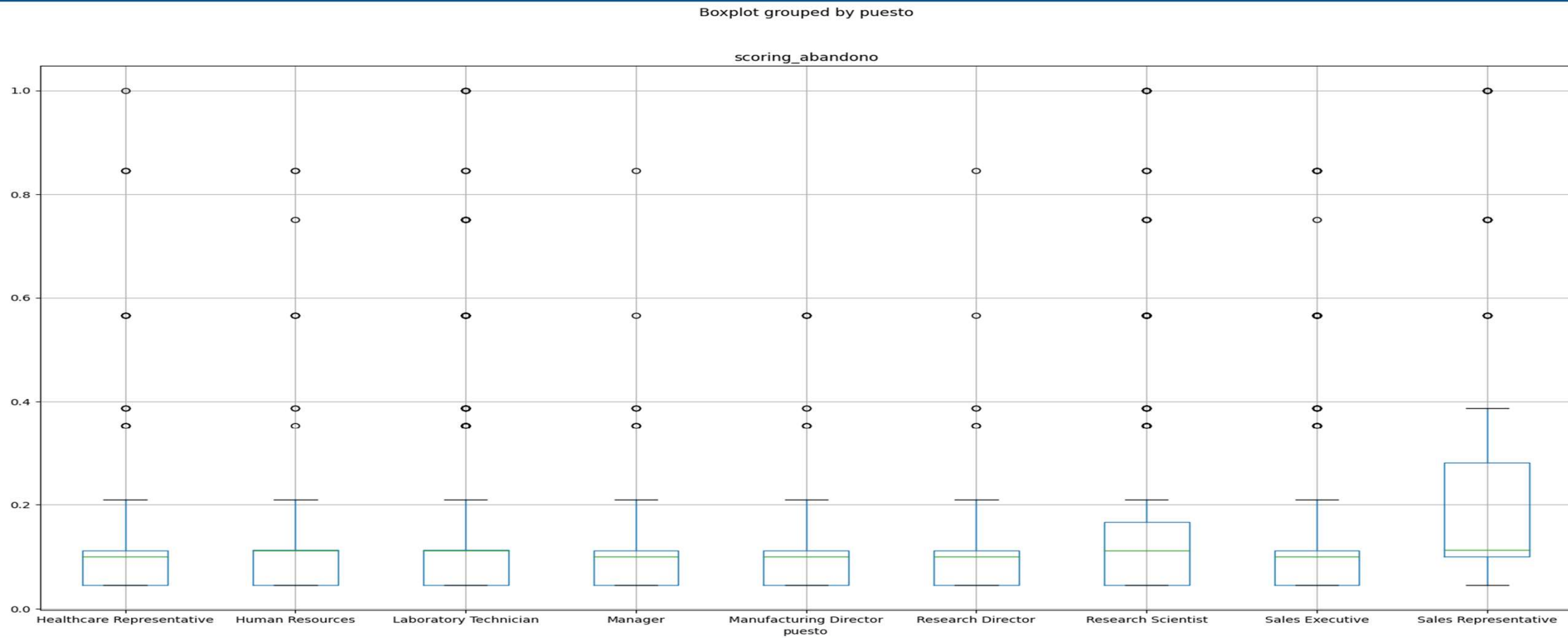
Para este entrenamiento se tiene una evaluación de 0.697 factible para trabajo y se pudo desarrollar un árbol de decisiones para visualizar las variables que pueden estar interviniendo en el abandono laboral.

Importancia de variables al aplicar modelo-árbol de decisiones.



Las variables mas resaltante que nos da el modelo son: Años_con_manager_actual.
edad.
horas_extra_Yes

Scoring por abandono.



Conociendo ya visualización de probabilidad de puestos de trabajos que puedan abandonar

Resumen estadísticos en Notebook.

Prueba Chi-cuadrado para Variables Categóricas:

```
from scipy import stats

anova_result =
stats.f_oneway(df['edad'][df['departamento'] ==
'Sales'],
               df['edad'][df['departa
mento'] == 'Research & Development'],
               df['edad'][df['departa
mento'] == 'Human Resources'])
print("Valor F:", anova_result.statistic)
print("Valor p:", anova_result.pvalue)

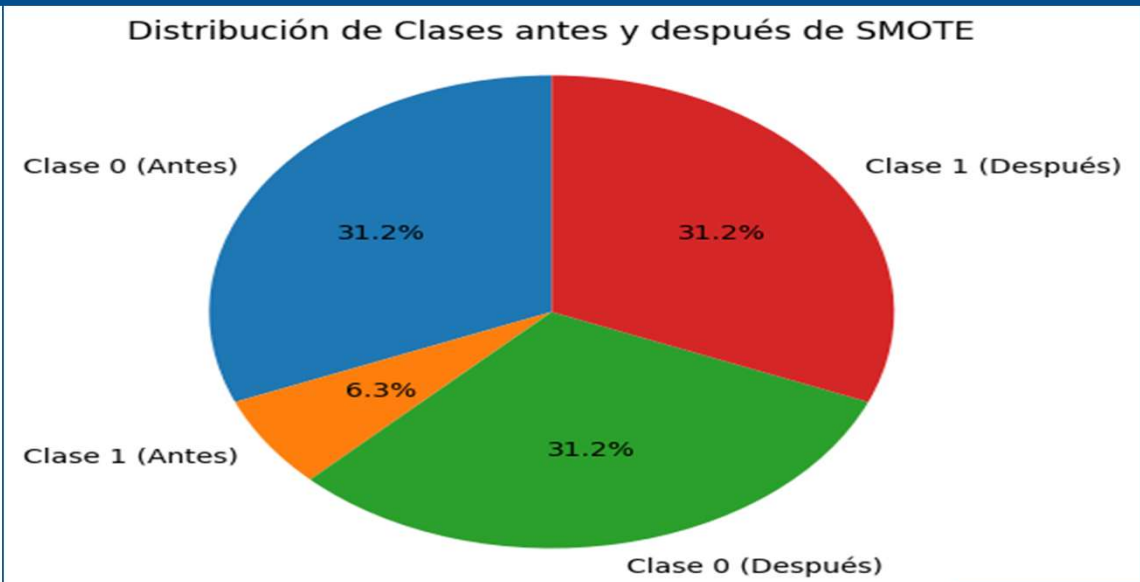
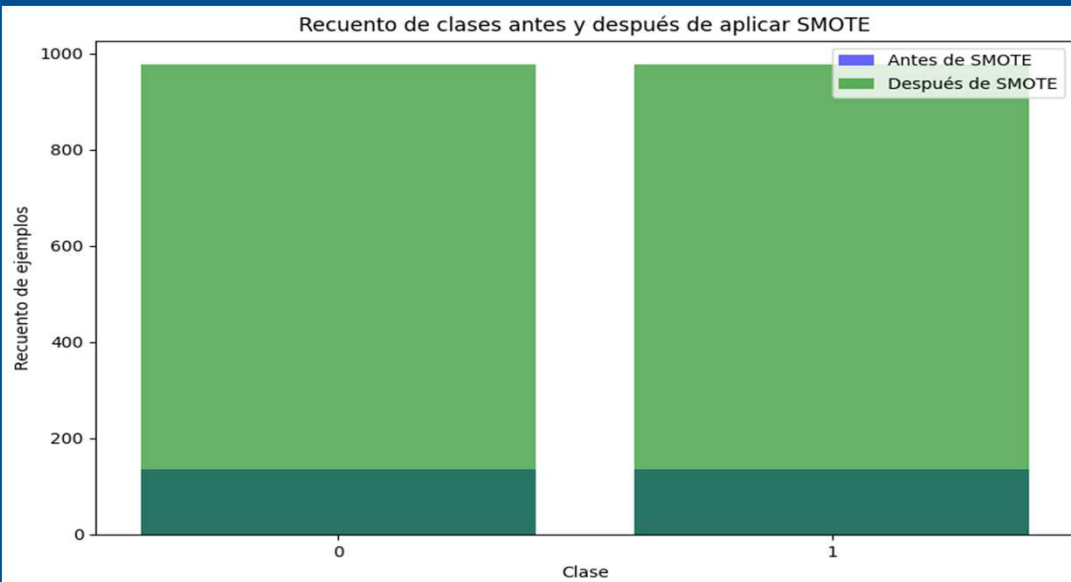
Valor F: 0.7655031924976903 Valor p:
0.46528552999349515

from scipy.stats import chi2_contingency

contingency_table = pd.crosstab(df['sexo'],
df['abandono'])
chi2, p, dof, expected =
chi2_contingency(contingency_table)
print("Valor Chi-cuadrado:", chi2)
print("Valor p:", p)
Valor Chi-cuadrado: 26.729303959810345 Valor p:
6.709037698658819e-06
```

Informe de Clasificación para Regresión Logística:				
	precision	recall	f1-score	support
0	0.87	1.00	0.93	255
1	0.00	0.00	0.00	39
accuracy			0.87	294
macro avg	0.43	0.50	0.46	294
weighted avg	0.75	0.87	0.81	294
AUC-ROC para Bosque Aleatorio: 0.5327300150829563				
Informe de Clasificación para Bosque Aleatorio:				
	precision	recall	f1-score	support
0	0.88	0.94	0.91	255
1	0.24	0.13	0.17	39
accuracy			0.83	294
macro avg	0.56	0.53	0.54	294
weighted avg	0.79	0.83	0.81	294

SMOTE.



VISUALIZACIONES Y RESÚMENES NUMÉRICOS GENERADOS RECOMENDACIONES BASADOS EN LOS INSIGHTS OBSERVADOS.

Satisfacción con los compañeros de trabajo.

Pregunta: ¿cuál es la distribución de la satisfacción con los compañeros de trabajo entre los empleados?

Hipótesis: la mayoría de los empleados están satisfechos con sus compañeros de trabajo, lo que podría estar relacionado con una menor tasa de abandono laboral.

Hallazgo clave: se observó una distribución positiva de la satisfacción, lo que sugiere un ambiente laboral positivo.

Relación entre el nivel de acciones y la satisfacción con compañeros de trabajo.

Pregunta: ¿existe alguna relación entre el nivel de acciones y la satisfacción con los compañeros de trabajo?

Hipótesis: los empleados con un nivel de acciones más alto podrían tener una mayor satisfacción con sus compañeros, lo que podría influir en su decisión de quedarse en la empresa.

Hallazgo clave: se encontró una correlación positiva entre el nivel de acciones y la satisfacción de los empleados.

Años de experiencia laboral.

Pregunta: ¿cuál es la distribución de años de experiencia laboral de los empleados en la empresa?

Hipótesis: la mayoría de los empleados podrían tener una experiencia laboral relativamente corta, relacionada con una mayor tasa de rotación.

Hallazgo clave: la distribución de la experiencia laboral es variada, lo que sugiere una retención de empleados razonable.

Formaciones y propensión al abandono.

Pregunta: ¿la cantidad de formaciones recibidas se correlaciona con la propensión al abandono laboral?

Hipótesis: más formaciones están relacionadas con una menor propensión al abandono laboral.

Hallazgo clave: no se encontró una correlación clara entre la cantidad de formaciones y la propensión al abandono.

Influencia del departamento y puesto en el abandono.

Pregunta: ¿el departamento y el puesto de los empleados influyen en su decisión de abandonar la empresa?

Hipótesis: algunos departamentos o puestos pueden tener una mayor tasa de rotación debido a factores relacionados con el trabajo o el entorno laboral.

Hallazgo clave: se identificaron diferencias significativas entre departamentos y puestos que podrían requerir atención adicional.

Satisfacción y antigüedad.

Pregunta: ¿existe una relación entre la satisfacción laboral y la antigüedad de los empleados en la empresa?

Hipótesis: empleados con mayor antigüedad tienen mayor satisfacción laboral debido a la familiaridad con la empresa.

Hallazgo clave: se observó una relación positiva entre la antigüedad y la satisfacción laboral.

CONCLUSIÓN.

Estos análisis proporcionan información valiosa para comprender la dinámica de abandono laboral en la empresa, Los hallazgos respaldan la importancia de mantener un ambiente laboral positivo y retener empleados con experiencia.

BIBLIOTECAS UTILIZADAS.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
from IPython.display import Image
from scipy.stats import chi2_contingency
!pip install scikit-learn imbalanced-learn
!pip install requests beautifulsoup4
import requests
from bs4 import BeautifulSoup
!pip install requests
!pip install pandas
!pip install lxml
%matplotlib inline
```

