Proyecto Profundización: Análisis de datos " Machine learning en RRHH"

Minerva Bobadilla
Latam
22-09-2024

Caso

De análisis

En el dinámico entorno empresarial actual, la retención de talento se ha convertido en un factor crítico para el éxito sostenible de las organizaciones. La capacidad de anticipar y mitigar la pérdida de empleados clave es esencial para mantener la estabilidad y el rendimiento a largo plazo.

En este contexto, este proyecto se enfoca en el análisis de los datos de recursos humanos de una empresa con el objetivo de desarrollar un modelo de machine learning supervisado.

Este modelo tiene como finalidad anticipar y gestionar la rotación del personal, fortaleciendo así la capacidad de la empresa para retener y desarrollar el talento clave.

Machine learning: “Análisis de RRHH”

Para el estudio se cuenta con una base de datos de 4410 trabajadores de una empresa, la cual se proceso con el desarrollo del modelo de machine learning, desde la recopilación y preparación de los datos hasta la implementación y evaluación del modelo en conjunto de datos de RRHH.

Se trabajó con las siguientes variables independientes: 'Age', 'BusinessTravel', 'Department', 'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender', 'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'.

Se trabajó con las siguientes variables dependientes: 'Attrition' (Empleado que abandona la empresa).

Machine learning: “Análisis de RRHH”

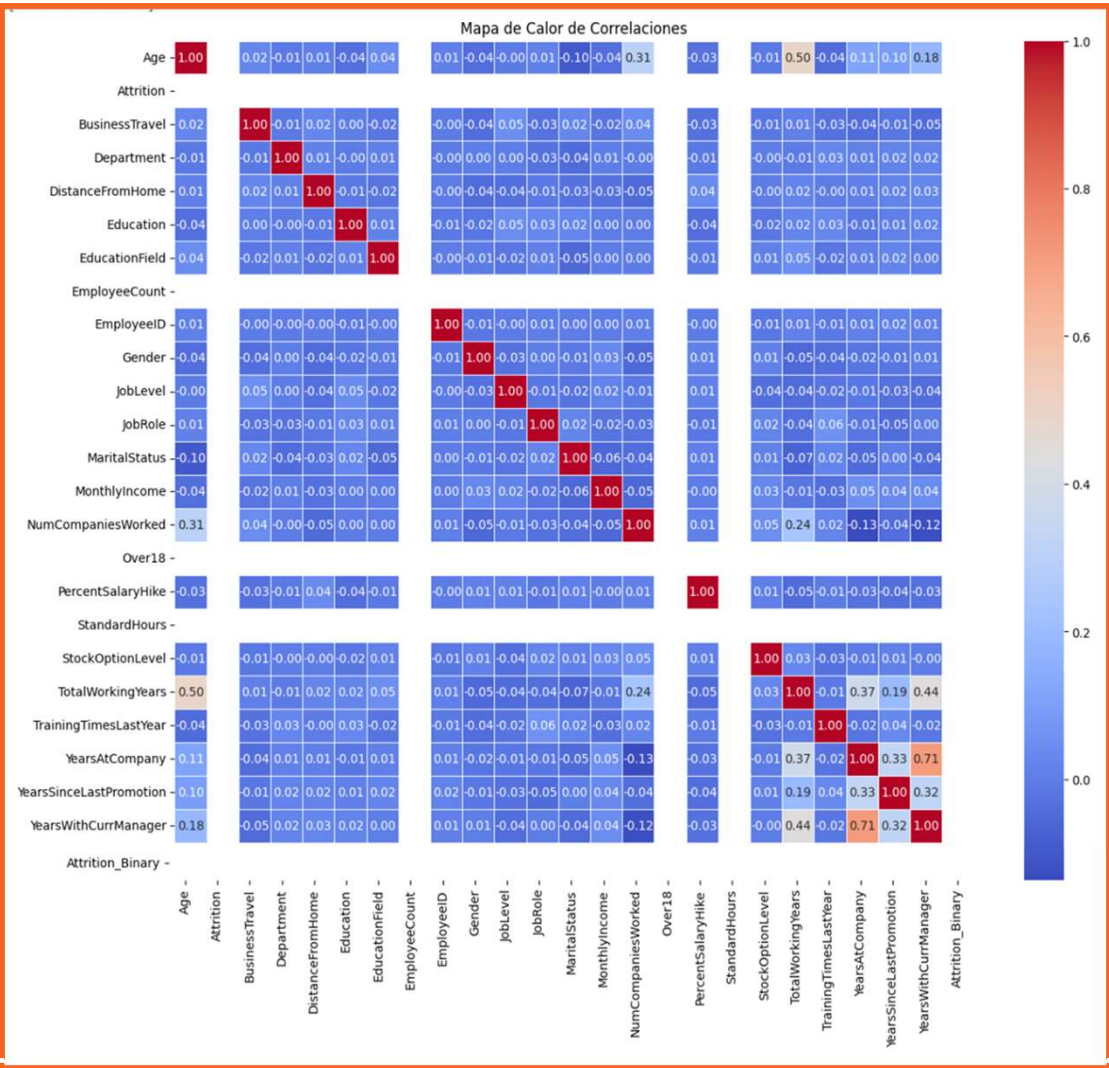


Figura N° 1: Mapa de calor de correlaciones de variables numéricas.

Machine learning: “Análisis de RRHH”

	Age	Age_Range	MonthlyIncome	Income_Range	TotalWorkingYears	\
0	51	46-55	49135	High	1.0	
1	31	26-35	41890	Medium	6.0	
2	32	26-35	49190	High	5.0	
3	38	36-45	83210	Very High	13.0	
4	32	26-35	23420	Low	9.0	

	Experience_Range	DistanceFromHome	Distance_Range	YearsAtCompany	\
0	0-5	6	6-10km	1	
1	6-10	10	11-20km	5	
2	6-10	17	11-20km	5	
3	11-20	2	0-5km	8	
4	6-10	10	11-20km	6	

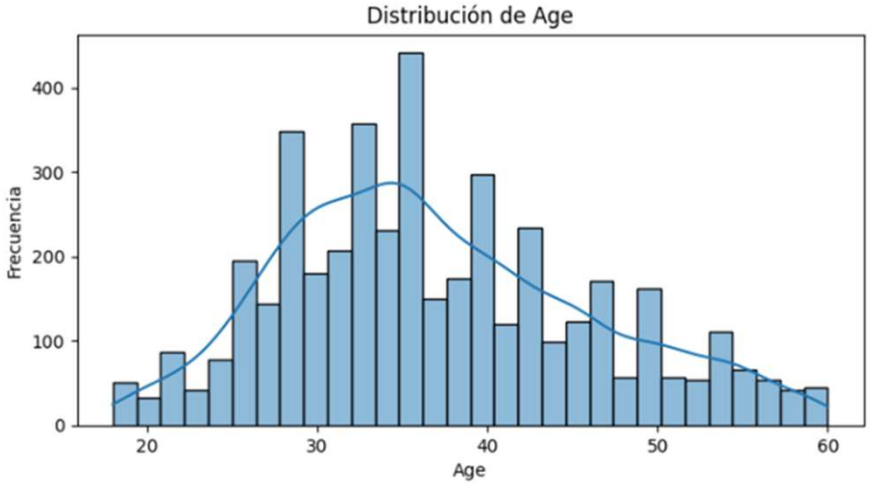
	YearsAtCompany_Range
0	0-5
1	6-10
2	6-10
3	6-10
4	6-10

Tabla N° 1: Creación de nuevas variables con categorías.

Machine learning: “Análisis de RRHH”

```
Medidas de tendencia central para 'Age':
Media: 36.923809523809524
Mediana: 36.0
Moda: 35

Medidas de dispersión para 'Age':
Desviación estándar: 9.133301271011144
Varianza: 83.41719210705376
Rango: 42
```




```
Medidas de tendencia central para 'DistanceFromHome':
Media: 9.19251700680272
Mediana: 7.0
Moda: 2

Medidas de dispersión para 'DistanceFromHome':
Desviación estándar: 8.105025518905281
Varianza: 65.69143866210581
Rango: 28
```

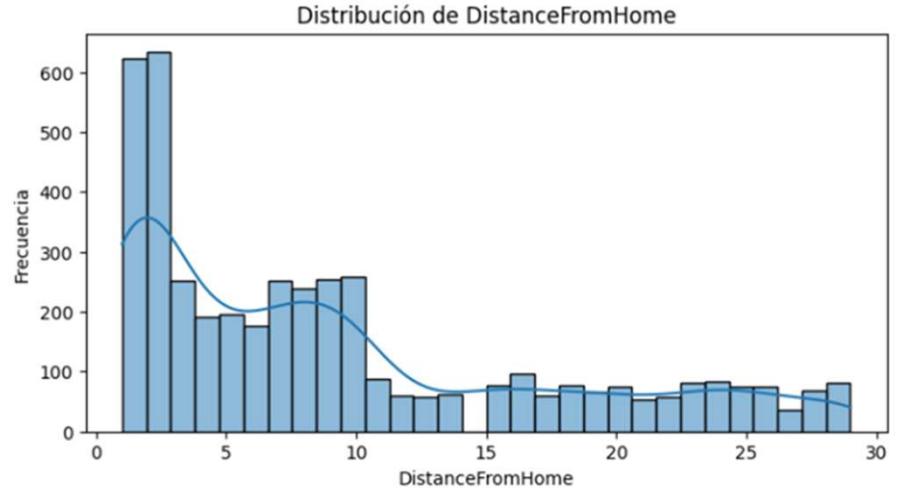
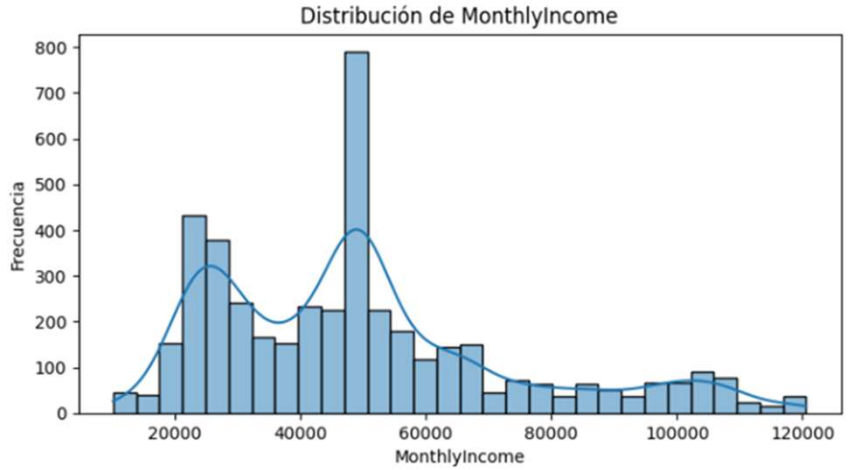


Figura N° 2 y 3: Medidas de tendencia central.

Machine learning: “Análisis de RRHH”

```
Medidas de tendencia central para 'MonthlyIncome':
Media: 49620.82653061225
Mediana: 49107.5
Moda: 49190

Medidas de dispersión para 'MonthlyIncome':
Desviación estándar: 24157.936569576497
Varianza: 583605899.2996815
Rango: 110520
```




```
Medidas de tendencia central para 'PercentSalaryHike':
Media: 15.209523809523809
Mediana: 14.0
Moda: 11

Medidas de dispersión para 'PercentSalaryHike':
Desviación estándar: 3.659107516298374
Varianza: 13.389067815831256
Rango: 14
```

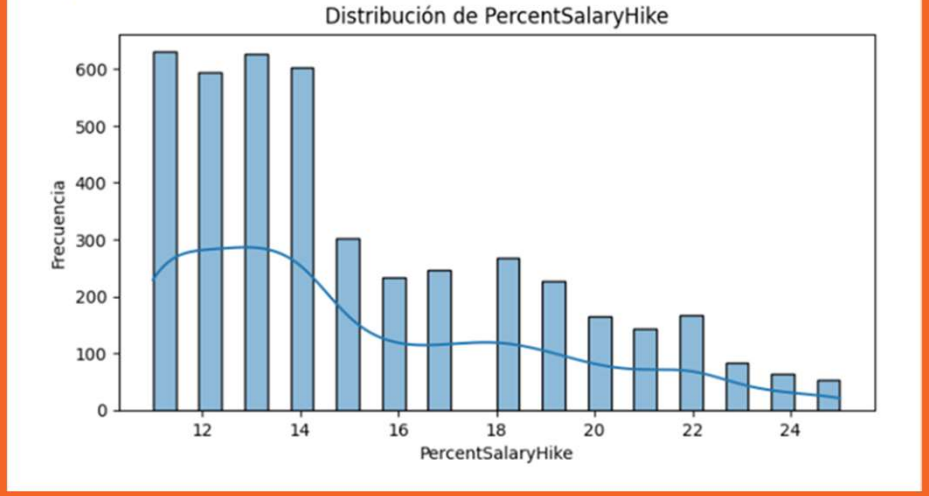
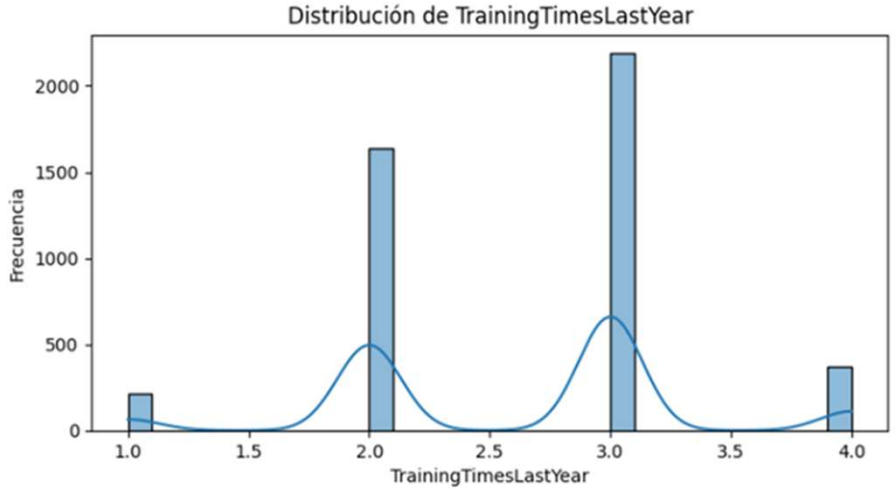


Figura N° 4 y 5: Medidas de tendencia central.

Machine learning: “Análisis de RRHH”

```
Medidas de tendencia central para 'TrainingTimesLastYear':
Media: 2.614965986394558
Mediana: 3.0
Moda: 3

Medidas de dispersión para 'TrainingTimesLastYear':
Desviación estándar: 0.707701893321818
Varianza: 0.5008419698112858
Rango: 3
```




```
Medidas de tendencia central para 'YearsAtCompany':
Media: 5.419727891156462
Mediana: 5.0
Moda: 5

Medidas de dispersión para 'YearsAtCompany':
Desviación estándar: 3.4646935703294837
Varianza: 12.004101536282464
Rango: 15
```

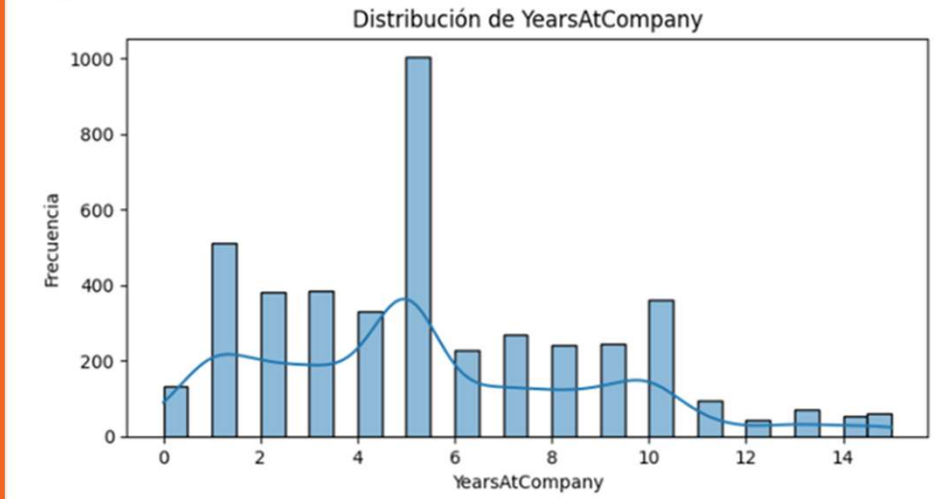
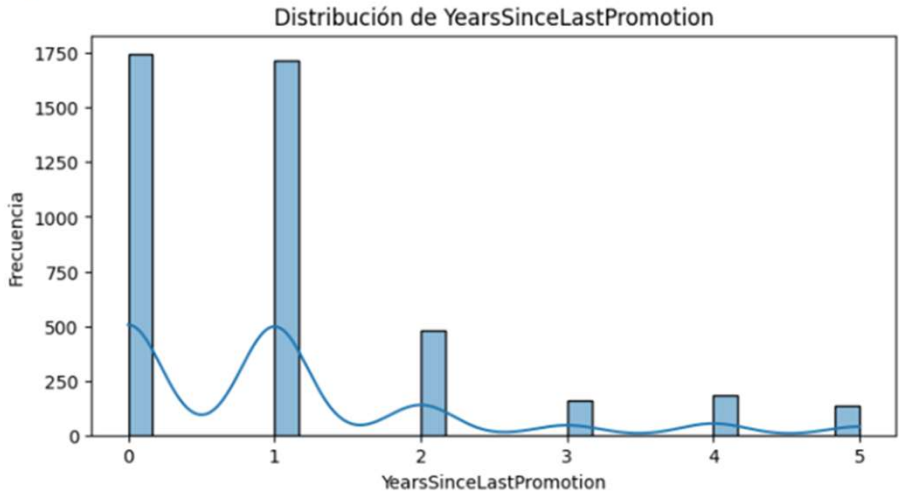


Figura N° 6 y 7: Medidas de tendencia central.

Machine learning: “Análisis de RRHH”

```
Medidas de tendencia central para 'YearsSinceLastPromotion':
Media: 1.030612244897959
Mediana: 1.0
Moda: 0

Medidas de dispersión para 'YearsSinceLastPromotion':
Desviación estándar: 1.2278306104363181
Varianza: 1.5075680079244216
Rango: 5
```




```
Medidas de tendencia central para 'YearsWithCurrManager':
Media: 3.9979591836734696
Mediana: 3.0
Moda: 2

Medidas de dispersión para 'YearsWithCurrManager':
Desviación estándar: 3.3671859090929974
Varianza: 11.337940946394436
Rango: 14
```

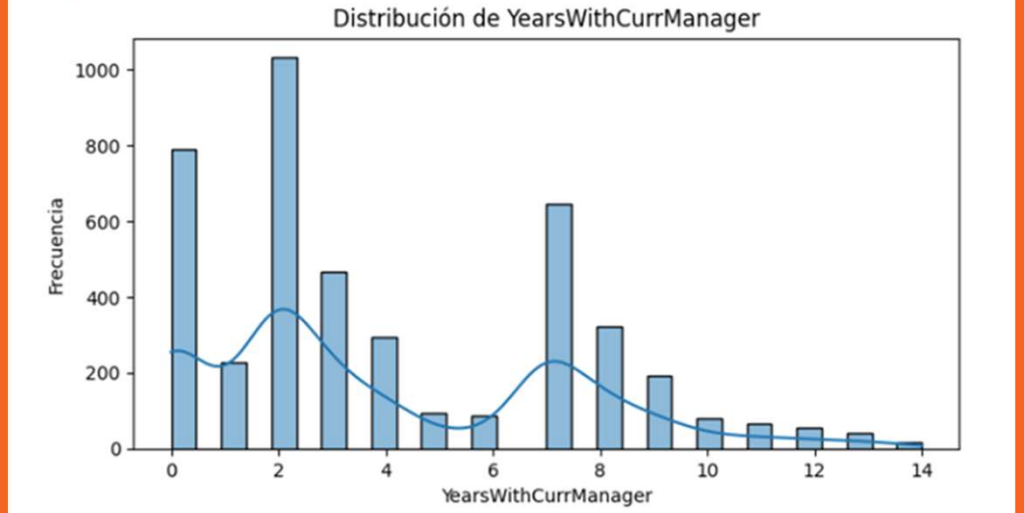


Figura N° 8 y 9: Medidas de tendencia central.

Machine learning: “Análisis de RRHH”

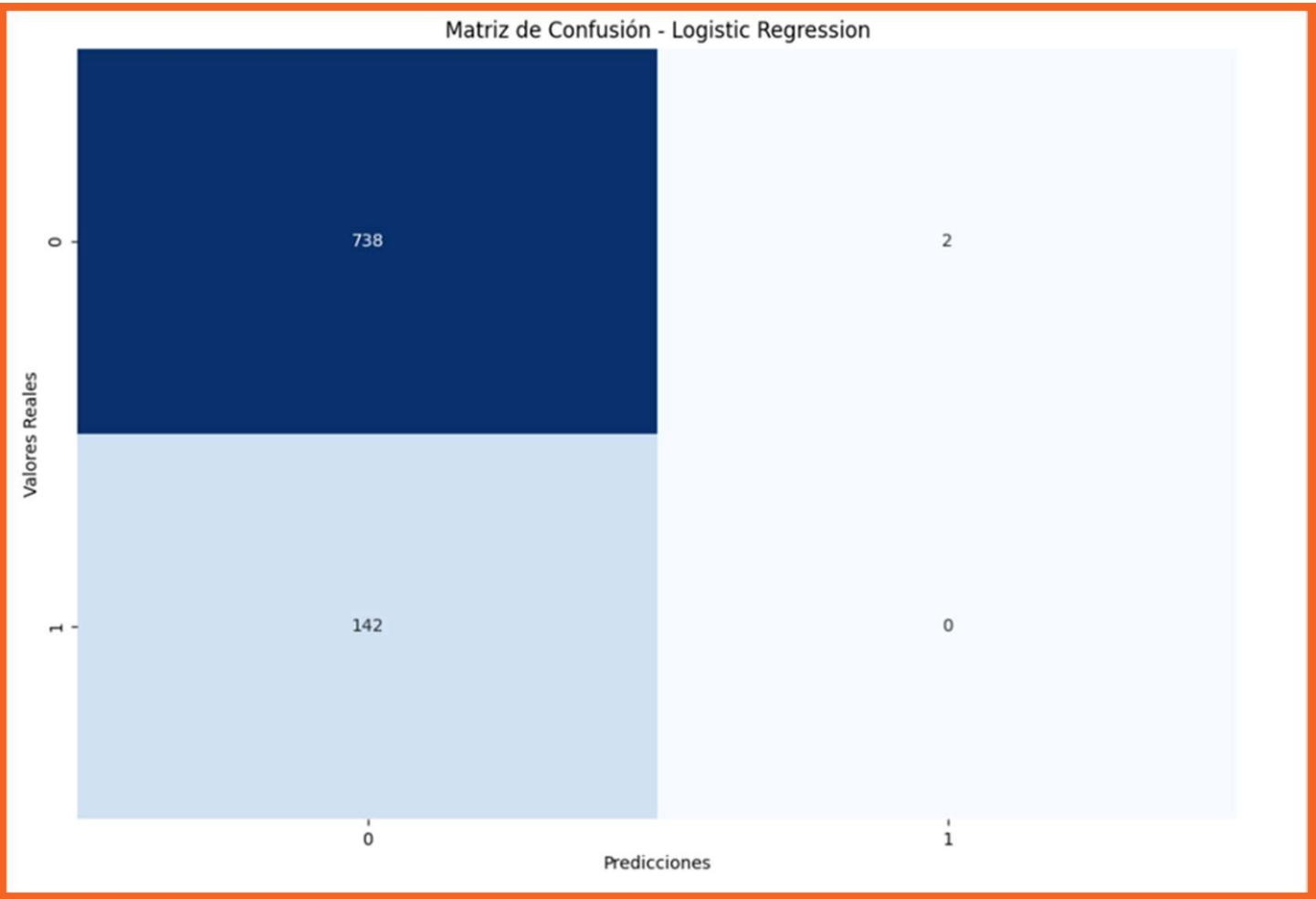


Figura N° 10: Matriz de confusión para modelo de Regresión Logística.

Machine learning: “Análisis de RRHH”

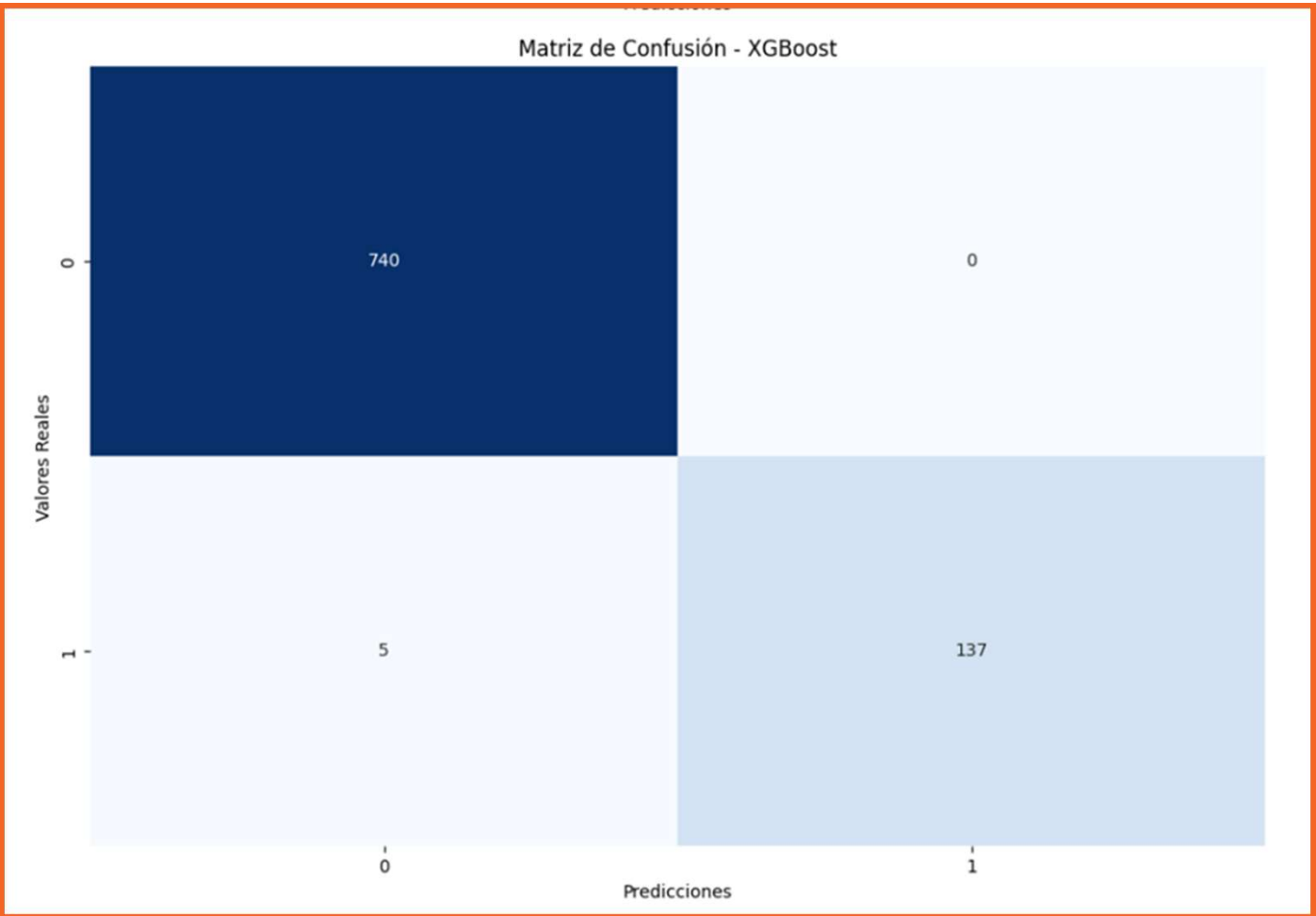


Figura N° 11: Matriz de confusión para modelo de Regresión XG-Bosst.

Machine learning: “Análisis de RRHH”

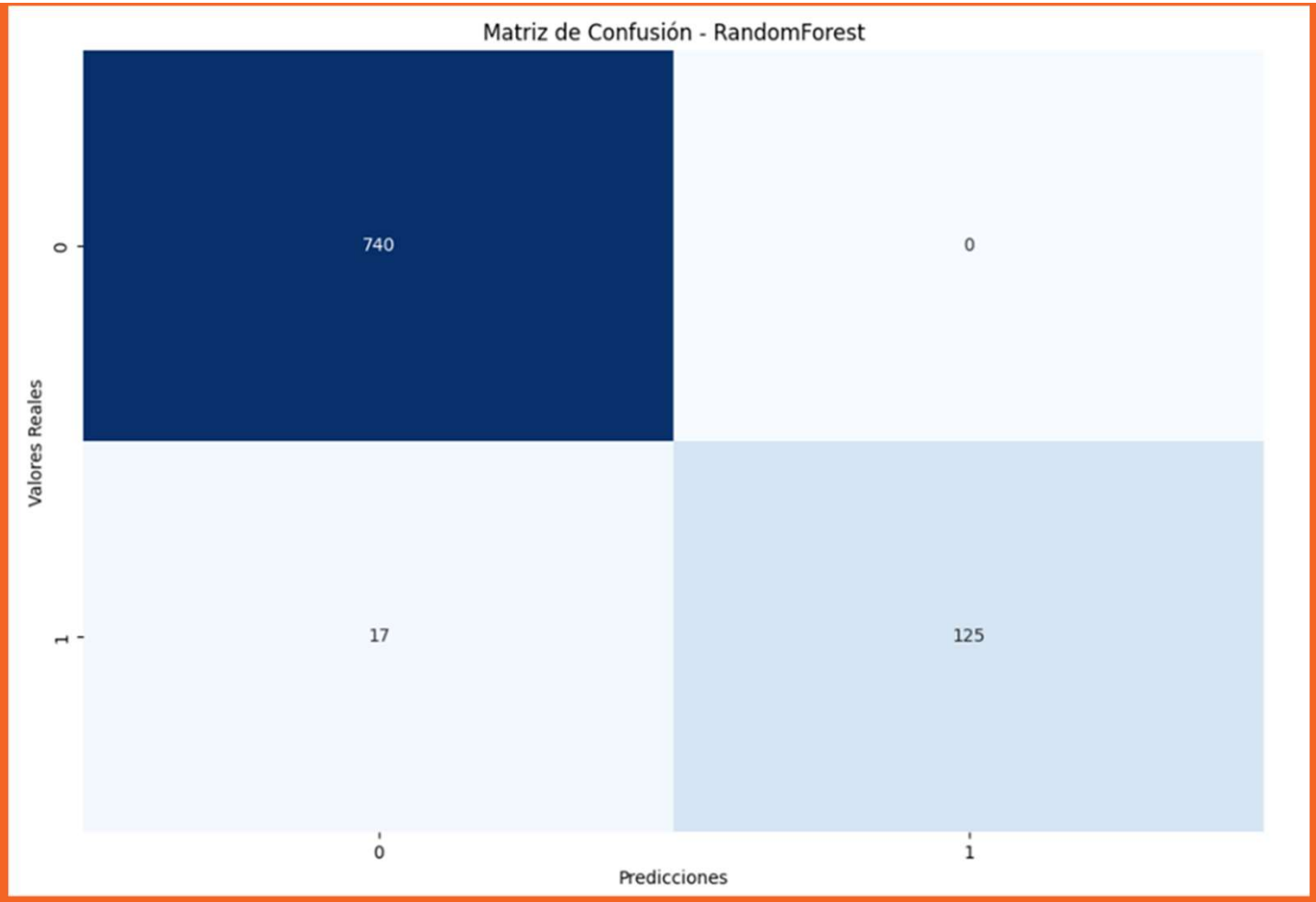


Figura N° 12: Matriz de confusión para modelo de Regresión RandomForest.

Machine learning: “Análisis de RRHH”

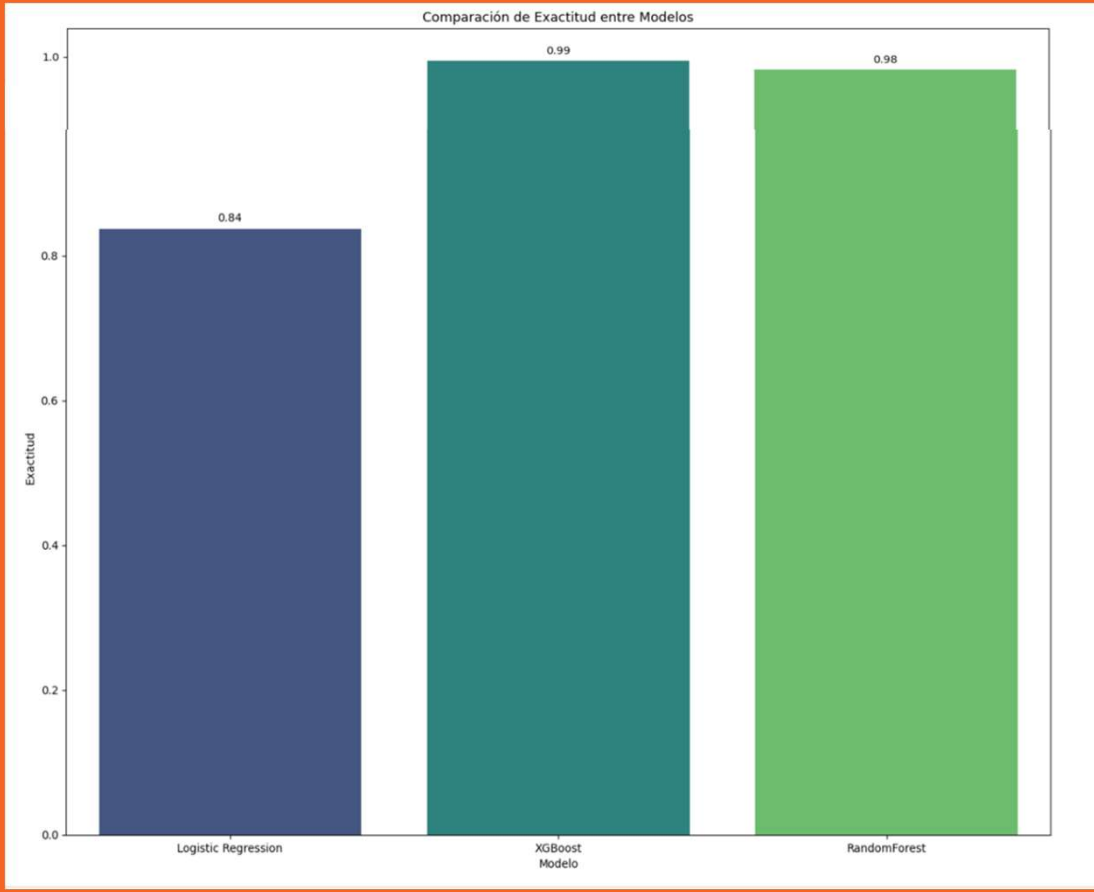


Figura N° 13: Comparación de exactitud entre modelos para Regresiones Logistica, XG-Bosst Y RandomForest.

Tabla N° 2: Resultados de modelos machine learning supervisado

Resultados de los modelos:

	Modelo	Exactitud (Accuracy)
0	Logistic Regression	0.836735
1	XGBoost	0.994331
2	RandomForest	0.980726

Machine learning: “Análisis de RRHH”

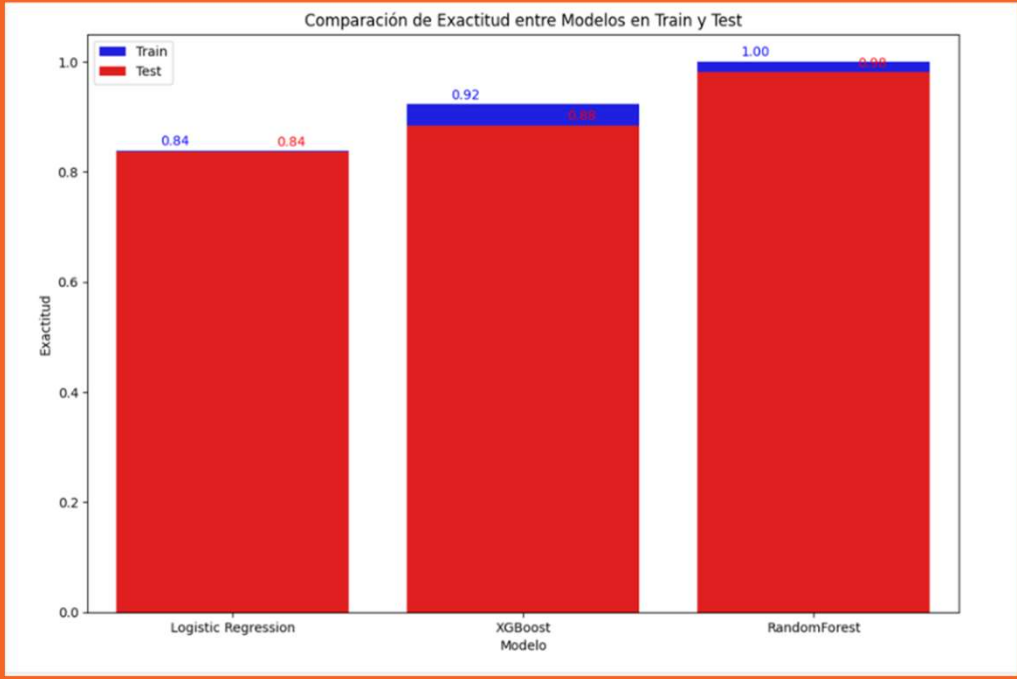


Figura N° 14: Verificación de los modelos XG-Boost y RandomForest para indicar si están sobreajustados.

Tabla N° 3: Resultados de modelos train vs test machine learning supervisado.

Resultados de los modelos (Train vs Test):

	Modelo	Exactitud en Train	Exactitud en Test
0	Logistic Regression	0.837868	0.836735
1	XGBoost	0.922902	0.884354
2	RandomForest	1.000000	0.980726

Machine learning: “Análisis de RRHH”

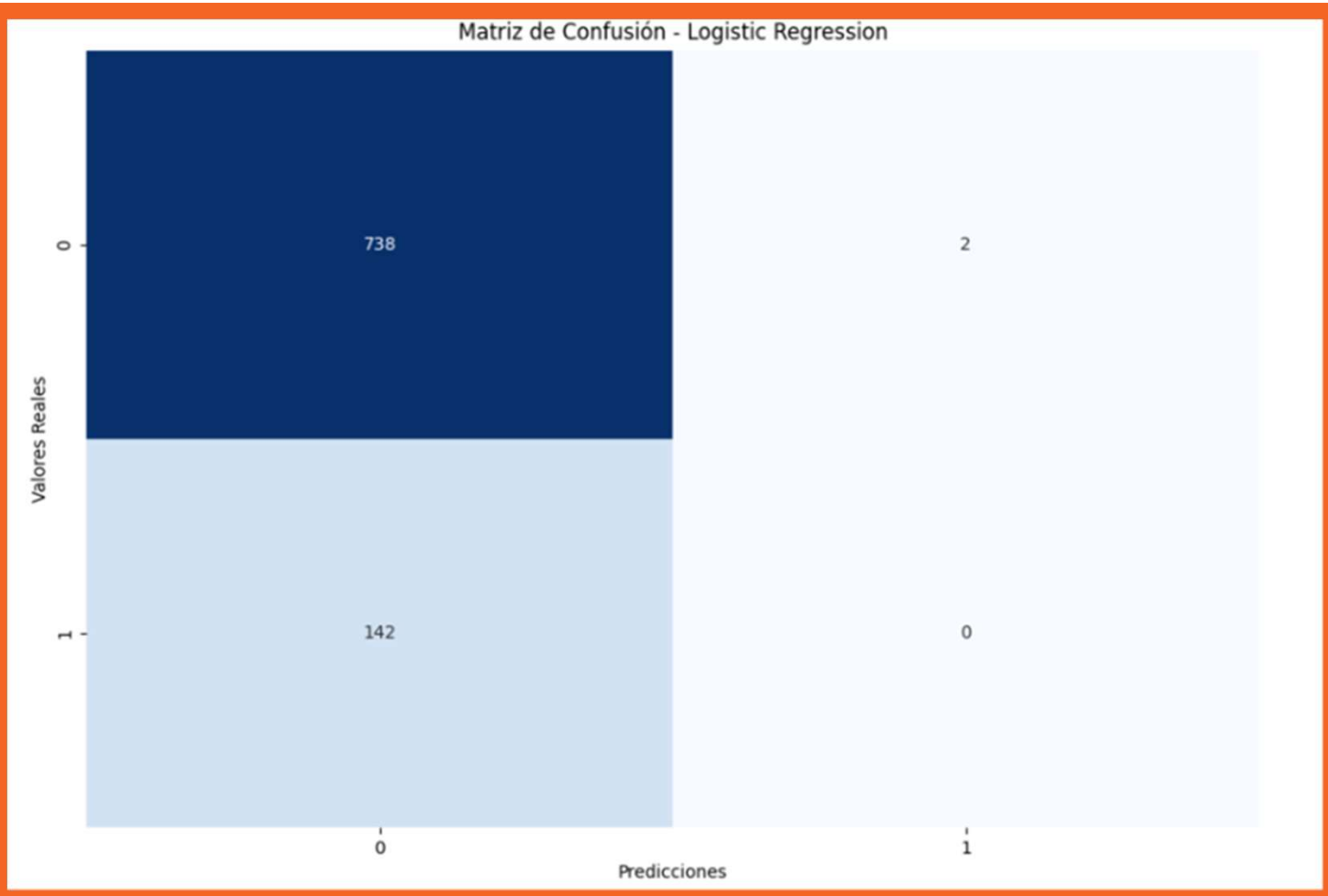


Figura N° 15: Matriz de confusión para modelo de Regresión Logística con validación cruzada.

Machine learning: “Análisis de RRHH”

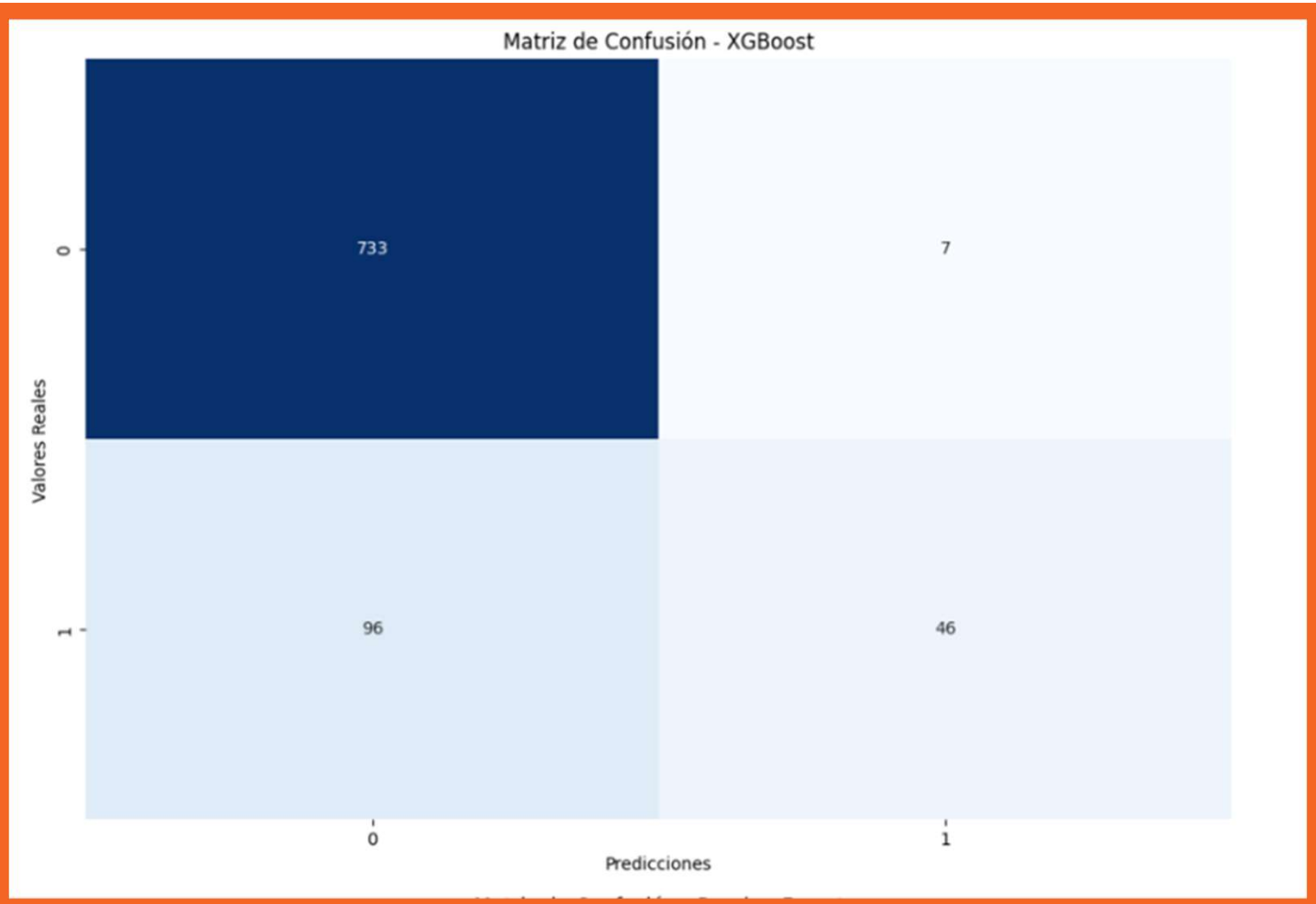


Figura N° 16: Matriz de confusión para modelo de Regresión XG-Boosst validación cruzada.

Machine learning: “Análisis de RRHH”

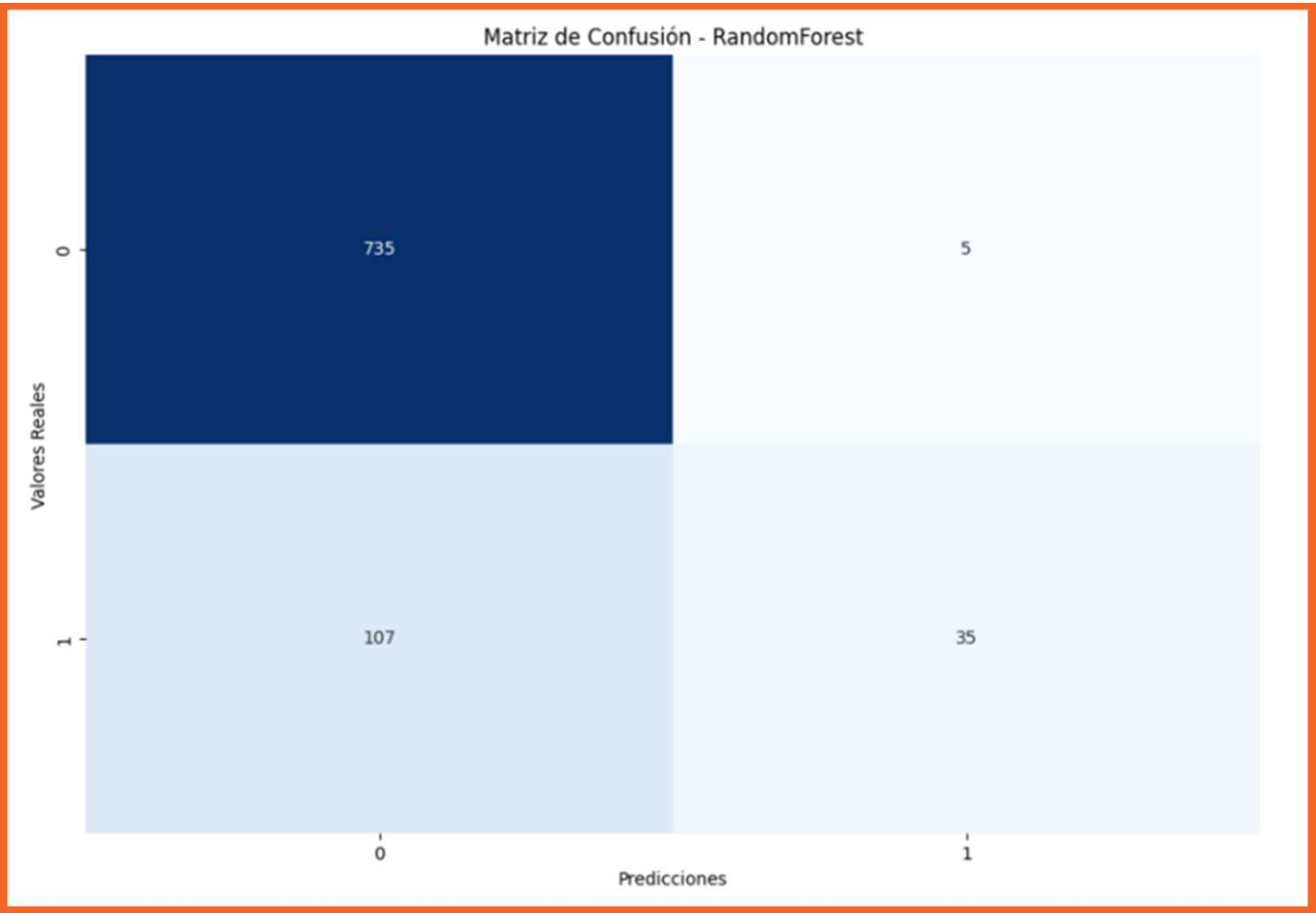


Figura N° 17: Matriz de confusión para modelo de Regresión RandomForest con validación cruzada.

Machine learning: “Análisis de RRHH”

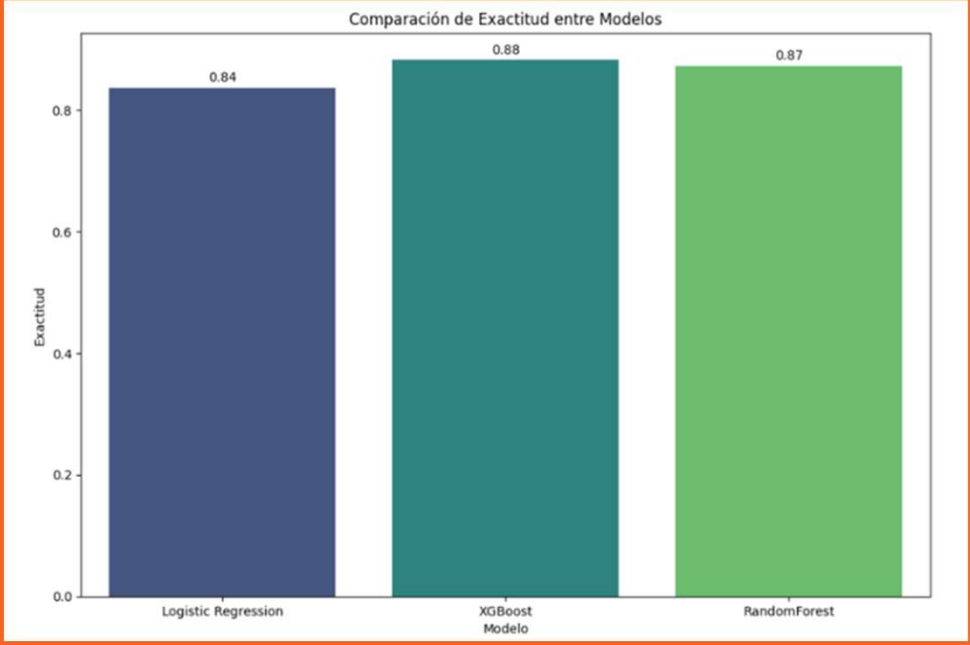


Figura N° 18: Comparación de exactitud entre modelos para Regresiones logistica, XG-Bosst Y RandomForest, validación cruzada con técnica de cross-validation.

Tabla N° 4: Resultados de exactitud para Regresiones logistica, XG-Bosst y RandomForest, validación cruzada con técnica de cross-validation.

```
Resultados de los modelos:
                Modelo  Exactitud (Accuracy)
0  Logistic Regression              0.836735
1              XGBoost              0.883220
2         RandomForest              0.873016

Exactitud promedio con cross-validation (Logistic Regression): 0.839
Exactitud promedio con cross-validation (XGBoost): 0.889
Exactitud promedio con cross-validation (Random Forest): 0.870
```

Conclusiones

Durante el desarrollo del proyecto se utilizó el proceso de desarrollo del modelo de machine learning, desde la recopilación y preparación de los datos hasta la implementación y evaluación del modelo.

El análisis inicial de base de datos fue Dividir la base en train y test, luego se entrenaron 3 modelos de regresión (Logistica, XG-Bosst Y RandomForest), en dos oportunidades, para comparar la exactitud de cada algoritmo en la base de test. Cuyos resultados fueron los siguientes:

1) En el primera iteración: XGBoost y Random Forest tienen una precisión muy alta en este conjunto de datos, lo que sugiere que pueden estar sobreajustados a los datos de entrenamiento. Regresión Logística, siendo un modelo más simple, ofrece una precisión decente y es menos propenso a sobreajustar. Sin embargo, no captura tanta complejidad en los datos como XGBoost o Random Forest. Como la precisión fue más alta que en el conjunto de prueba, es necesario hacer una validación cruzada.
2) Segunda iteración : XGBoost es el mejor modelo hasta ahora (88.32%), y pequeños ajustes en los hiperparámetros podrían mejorar aún más su rendimiento. Logistic Regression es un modelo más sencillo, pero su desempeño es consistente y confiable. Random Forest es competitivo, pero no supera a XGBoost en este caso, aunque tiene un buen equilibrio y generaliza bien.

Recomendaciones

Seguir ajustando hiperparámetros en XGBoost y Random Forest para intentar mejorar la generalización. Considerar reducir ligeramente la complejidad del modelo para asegurarte de que no esté memorizando partes del conjunto de entrenamiento.

Muchas gracias