

ANÁLITICA PARA LA TOMA DE DECISIONES

CASO PRACTICO APRENDIZAJE SUPERVIZADO



PRESENTADO POR:

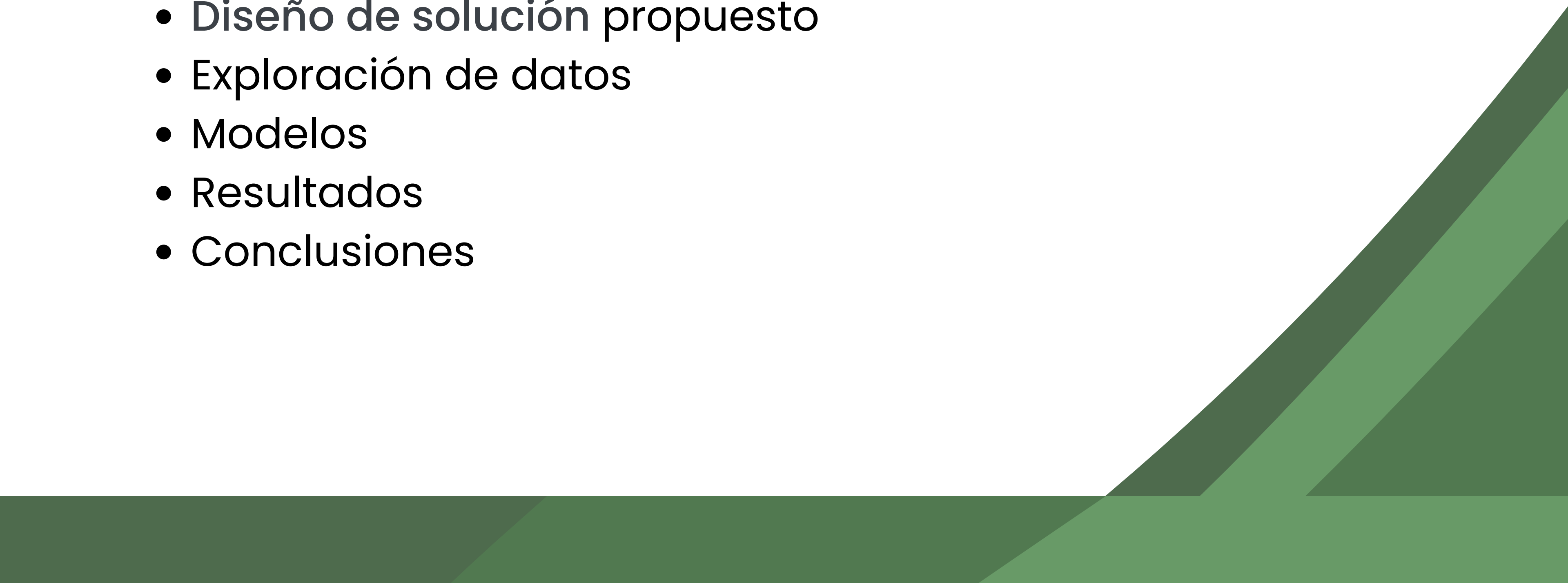
FABIO ANDRES GONZALEZ

NATALIA

SUSANA BARRIENTOS

JAVIER ELOHIM BURGOS

AGENDA

- Diseño de solución propuesto
 - Exploración de datos
 - Modelos
 - Resultados
 - Conclusiones
- 
- The slide features a decorative design on the right side and bottom. On the right, there are several overlapping triangular and quadrilateral shapes in various shades of green, creating a modern, abstract background. A solid dark green horizontal bar runs across the bottom of the slide.




DISEÑO DE SOLUCIÓN PROPUESTO

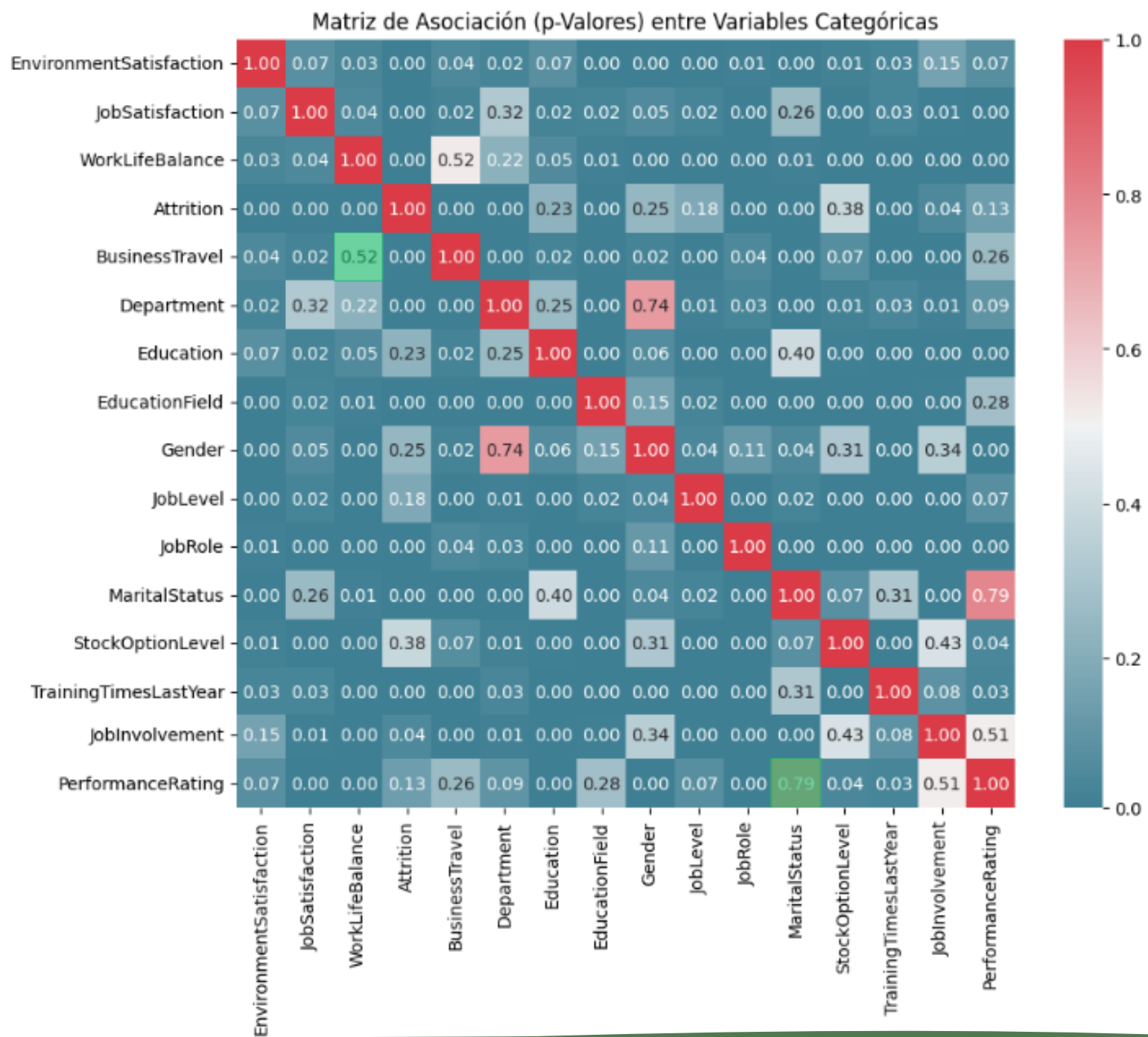
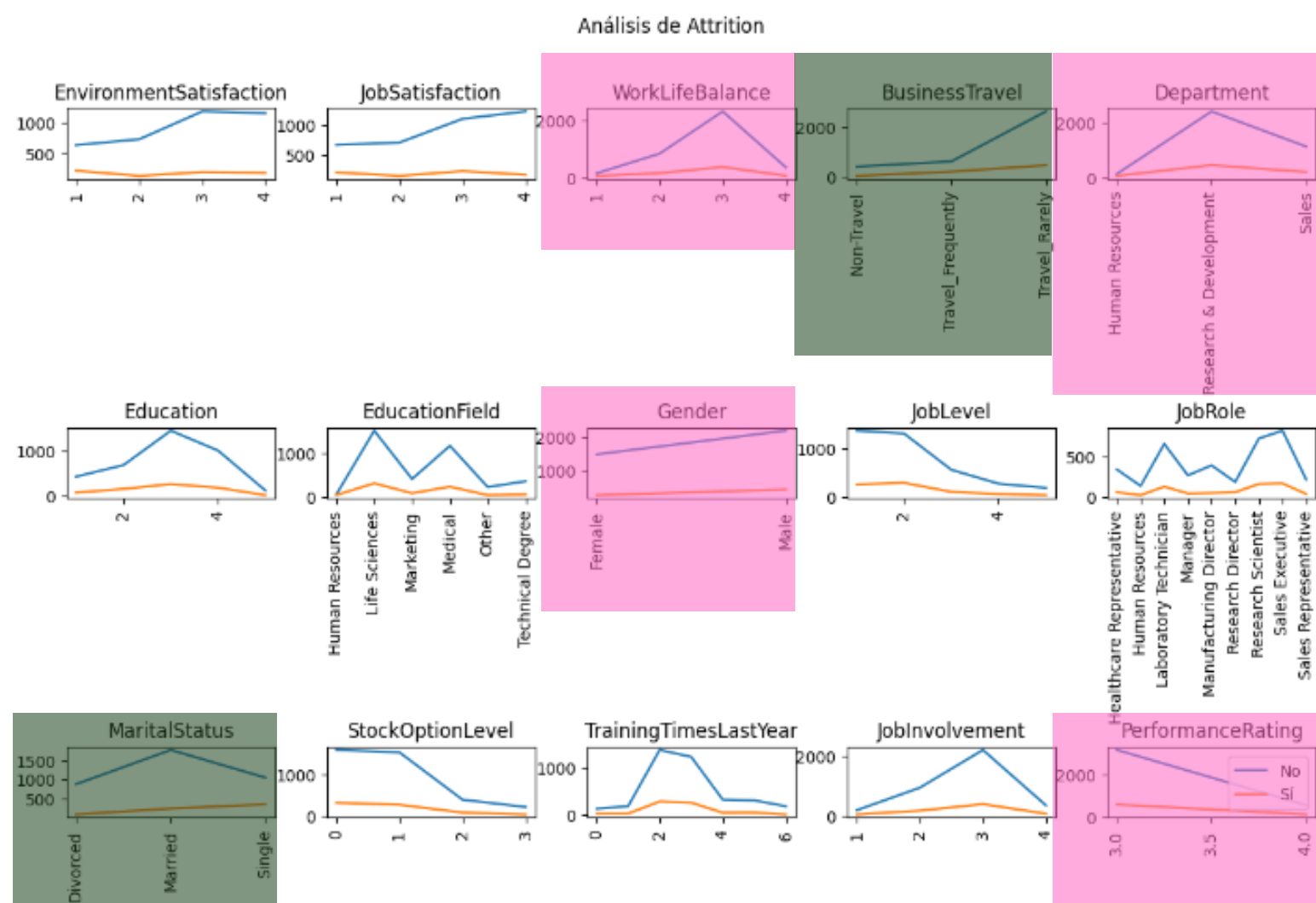
EN PRIMER INSTANCIA PLANTEAMOS EL USO DE UN MODELO DE REGRESIÓN LOGÍSTICA, PUESTO QUE DICHO MODELO ES ADECUADO PARA PROBLEMAS DE CLASIFICACIÓN BINARIA, COMO ES EL CASO DE ESTE, DONDE SE TRATA DE PREDECIR SI UN EMPLEADO ABANDONARÁ SU EMPLEO O NO.

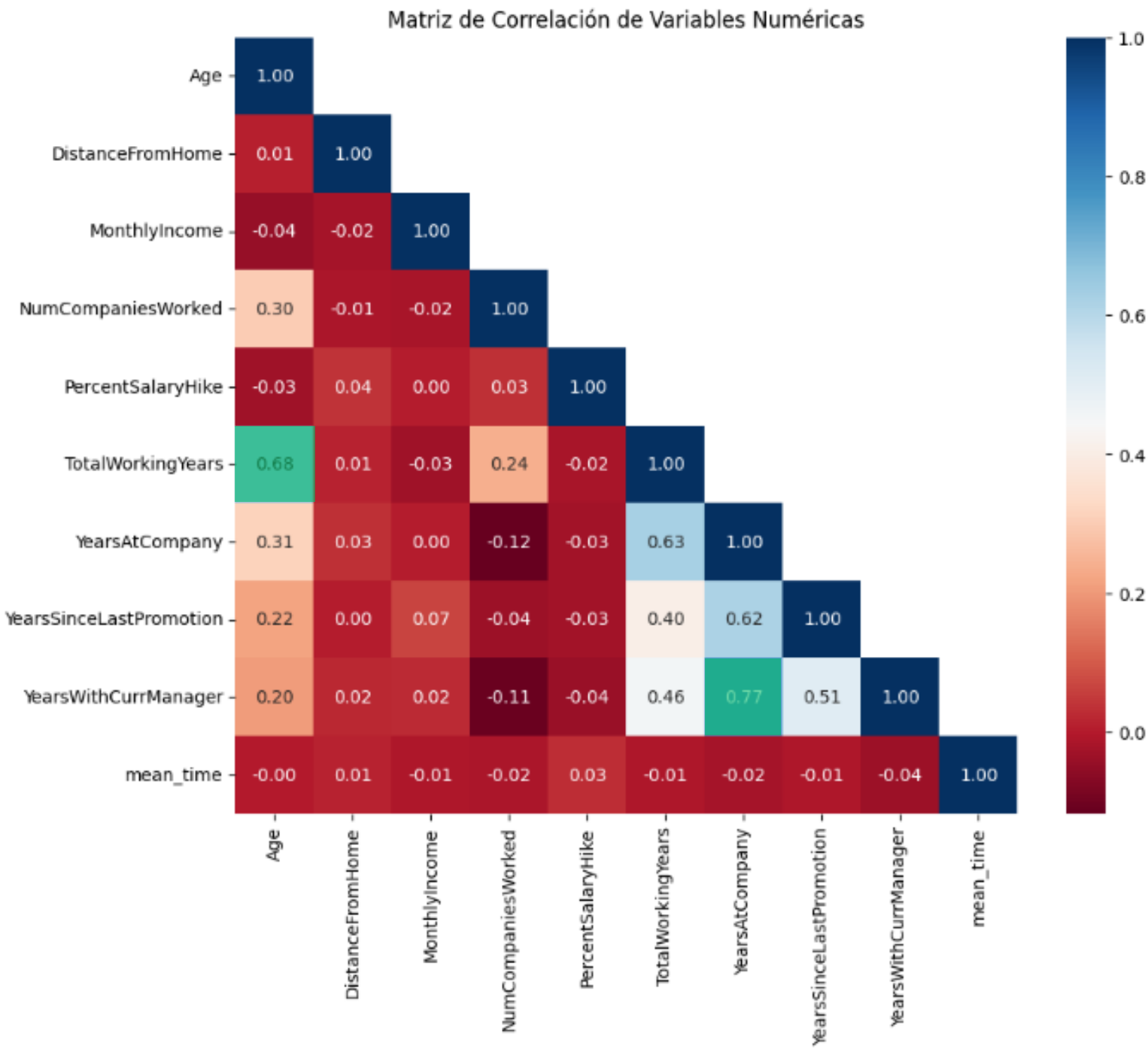
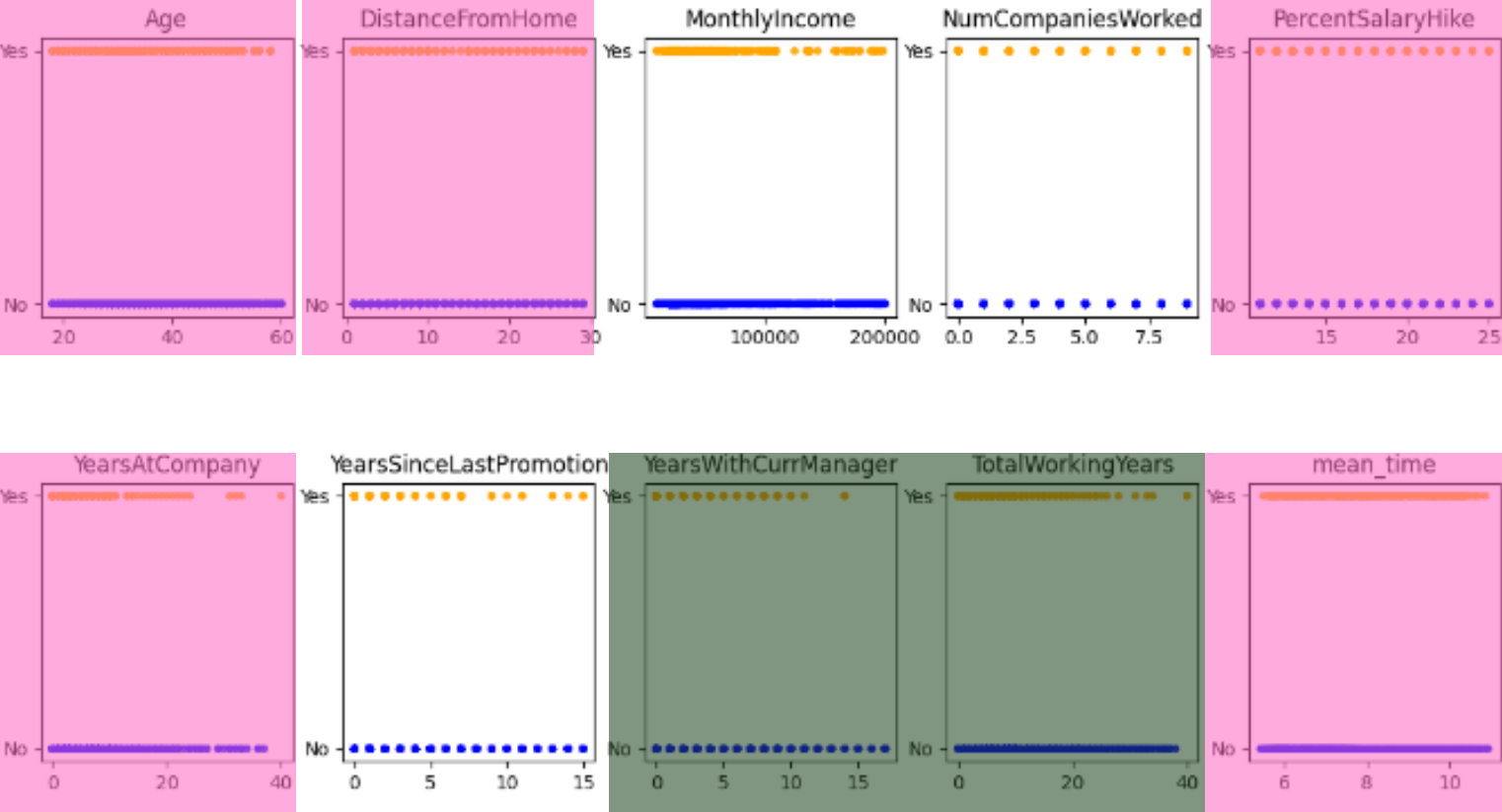
SIN EMBARGO TAMBIÉN PLANTEAMOS LA ALTERNATIVA DE USAR ÁRBOLES DE DECISIÓN, YA QUE SON UN MODELO DE APRENDIZAJE SUPERVISADO QUE PUEDE SER UTILIZADO PARA PROBLEMAS DE CLASIFICACIÓN Y REGRESIÓN, CON LA DESVENTAJA QUE PUEDEN SER MENOS PRECISOS QUE LOS MODELOS DE REGRESIÓN LOGÍSTICA

EXPLORACIÓN DE LOS DATOS

1. JUNTAR LAS BASES
 2. ELIMINACIÓN DE VARIABLES CON UN ÚNICO VALOR
 3. TRATAMIENTO DE DATOS DE FALTANTES
 4. ANÁLISIS EXPLORATORIO
 5. CAMBIOS EN LA VARIABLE OBJETIVO
- 

EXPLORACION DE DATOS





A large green circular graphic is positioned on the left side of the slide, partially cut off by the edge.

MODELOS; SIN SELECCIÓN DE VARIABLES

1. **MODELO DE REGRESIÓN RANDOM FOREST REGRESSOR**

MODELOS; CON SELECCIÓN DE VARIABLES

SELECCIÓN DE VARIABLES:

K-Best con ANOVA

	Age	TotalWorkingYears	YearsWithCurrManager	mean_time	MaritalStatus_Single
0	51	1.0	0	6.8702	0
1	31	6.0	4	7.3160	1
2	32	5.0	3	6.8161	0
3	38	13.0	5	6.7892	0
4	32	9.0	4	7.8776	1



MODELOS; CON SELECCIÓN DE VARIABLES

1. **MODELO LOGISTIC REGRESSION**
2. **MODELO RANDOMFOREST CLASSIFIER**
3. **MODELO GRADIENTBOOSTING CLASSIFIER**

RESULTADOS

Metricas regresion logistica sin selección de variables

LogisticRegression	RandomForestRegressor
Accuracy (Train): 84.89%	Accuracy (Train): 94.51%
MSE entrenamiento: 0.15	MSE entrenamiento: 0.0074282312925170076
MAE entrenamiento: 0.15	MAE entrenamiento: 0.048582766439909296
R2 entrenamiento: 0.12	R2 entrenamiento: 0.9451635230062987
Accuracy (Test): 14.220607806319164%	Accuracy (Test): 56.38320441037127%
MSE prueba: 0.11520812670373928	MSE prueba: 0.05858061224489795
MAE prueba: 0.24830583345103788	MAE prueba: 0.14264172335600908
R2 prueba: 0.14220607806319163	R2 prueba: 0.5638320441037127

RandomForestRegressor supera significativamente a LogisticRegression en términos de rendimiento en el conjunto de entrenamiento y mantiene un mejor rendimiento en el conjunto de prueba en términos de exactitud, MSE, MAE y R2.

RESULTADOS

Metrics confusion_matrix

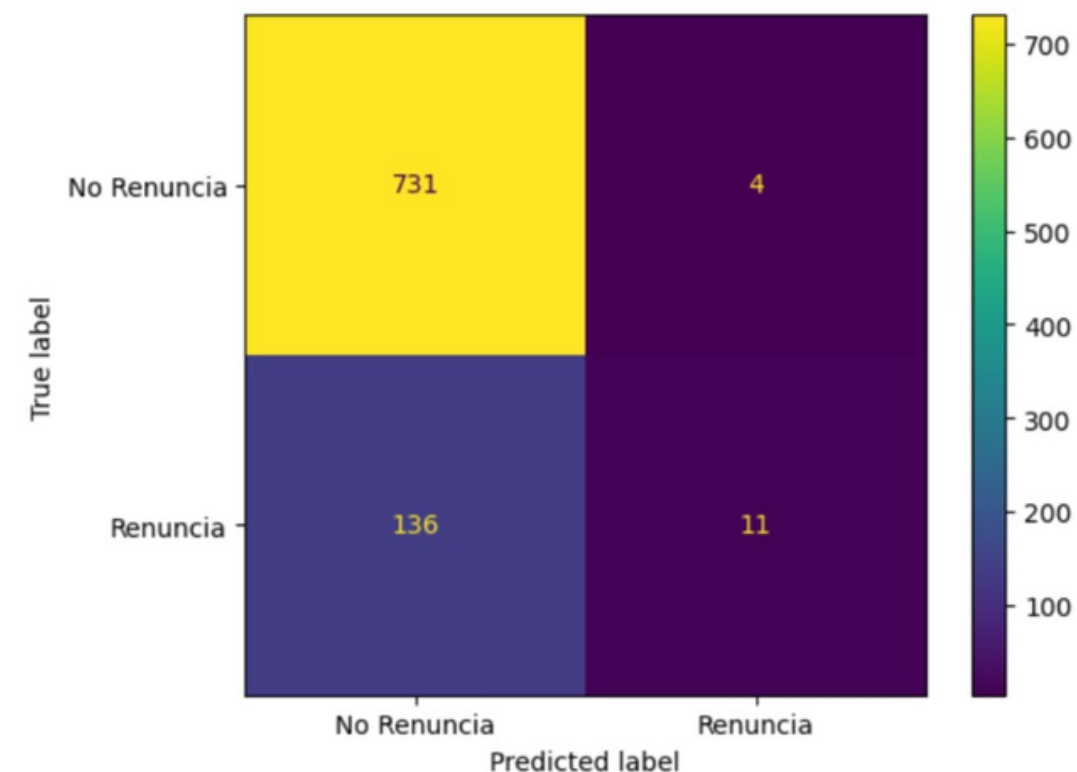
Precision: 0.7333333333333333

Recall: 0.07482993197278912

Especificidad:

0.9945578231292517

F1 score: 0.13580246913580246



731 Verdaderos Negativos (personas que dijeron que no renunciarían y no lo hicieron).

136 Verdaderos Positivos (personas que dijeron que renunciarían y lo hicieron).

11 Falsos Positivos (personas que dijeron que no renunciarían pero renunciaron).

4 Falsos Negativos (ninguna persona dijo que renunciaría).

CONCLUSIONES

- Se identificaron las variables más relevantes para predecir la rotación de empleados. Estas variables pueden incluir aspectos como la satisfacción laboral, el nivel educativo, el salario mensual y la cantidad de años trabajados en la empresa. Comprender estas variables es crucial para tomar medidas preventivas.
- Al realizar la comparación entre los modelos de entrenamiento y validación con y sin selección de variables, se evidencia que el modelo que se apoya de la selección de variables cuenta con mejores métricas. Esto indica que la selección de variables ha ayudado al modelo a mejorar su capacidad predictiva y a reducir la complejidad al enfocarse en las características más relevantes para el problema en cuestión.
- Se evaluaron las métricas de rendimiento de los modelos, como la exactitud, el MSE (Mean Squared Error), el MAE (Mean Absolute Error) y el R2 (Coeficiente de Determinación). Estas métricas proporcionan información sobre qué tan bien se desempeñan los modelos en términos de predicción.
- Es importante destacar que el modelo RandomForestClassifier superó significativamente a la Regresión Logística en términos de rendimiento en el conjunto de entrenamiento y en el conjunto de prueba. Esto sugiere que el uso de árboles de decisión puede ser más efectivo en este problema de clasificación.
- La selección de variables es un paso crucial en la construcción de modelos de machine learning. En este estudio, se demostró que al seleccionar cuidadosamente las variables más relevantes, se puede mejorar significativamente el rendimiento de los modelos de clasificación.

GRACIAS!