INFORME DE ANÁLISIS

NOMBRE DEL PROYECTO:

PROJECT TALENTFLOW

NÚMERO DE GRUPO:

GRUPO 7

INTEGRANTES:

BROWN VARGAS ROMINA ANDREA

PERALTA BAIDAL DARWIN VICENTE

REYES HOLGUIN ROBERTO DANTE

VALLEJO VÁSQUEZ AQUILES ANDERS

VERA CAJAPE NEYVI MAGELES

Introducción

El presente informe detalla el proceso de exploración, limpieza y análisis inicial realizado sobre un dataset compuesto por información de 1470 empleados y 36 variables. El trabajo fue desarrollado en Python utilizando bibliotecas como pandas, numpy y matplotlib.

El objetivo principal consistió en evaluar la estructura, calidad y consistencia de los datos para asegurar que se encuentren en condiciones óptimas antes de aplicar análisis estadísticos o modelos predictivos. Para ello, se revisaron tipos de datos, presencia de valores nulos, duplicados, outliers y columnas irrelevantes. Posteriormente, se aplicaron técnicas de depuración, imputación y selección de variables, con el fin de mejorar la integridad del dataset.

Adicionalmente, se desarrollaron gráficos exploratorios que permitieron examinar la distribución de variables relevantes como la edad, el nivel de ingresos y factores asociados a la permanencia o salida de los empleados. Estos análisis proporcionan una primera visión de los patrones presentes en la fuerza laboral y de los elementos que podrían influir en la decisión de un empleado de continuar o abandonar la organización.

- **Dimensiones del Dataset:**

El dataset contiene 1470 filas y 36 columnas. Esto significa que contamos con información detallada de 1470 empleados y 36 características asociadas a cada uno.

- **Descripción general de columnas:**

Nombre de la columna	Tipo de dato	Contiene nulos
Unnamed: 0	int64	No
Age	int64	No
Attrition	object	No
BusinessTravel	object	No
DailyRate	int64	No
Department	object	No
DistanceFromHome	int64	No
Education	int64	No
EducationField	object	No
EmployeeCount	int64	No
EmployeeNumber	int64	No
EnvironmentSatisfaction	int64	No
Gender	object	No
HourlyRate	int64	No
JobInvolvement	int64	No
JobLevel	int64	No
JobRole	object	No
JobSatisfaction	int64	No
MaritalStatus	object	No
MonthlyIncome	int64	No
MonthlyRate	int64	No

NumCompaniesWorked	int64	No
Over18	object	No
OverTime	object	Sí
PercentSalaryHike	int64	No
PerformanceRating	int64	No
RelationshipSatisfaction	float64	Sí
StandardHours	int64	No
StockOptionLevel	int64	No
TotalWorkingYears	int64	No
TrainingTimesLastYear	int64	No
WorkLifeBalance	int64	No
YearsAtCompany	int64	No
YearsInCurrentRole	int64	No
YearsSinceLastPromotion	int64	No
YearsWithCurrManager	int64	No

- **Tipos de datos presentes:**

Numéricos (int64, float64): 27 columnas. Incluyen variables como Age, MonthlyIncome, YearsAtCompany, DailyRate, etc.

Categóricas (object): 9 columnas. Entre ellas se encuentran Attrition, Gender, Department, JobRole y BusinessTravel.

Columnas posiblemente irrelevantes:

Unnamed: 0 es un índice adicional sin valor analítico.

EmployeeCount y StandardHours: A simple vista parece tener valores constantes.

Over18: podría no aportar información útil, ya que todos los registros probablemente son mayores de edad.

Memoria utilizada: El DataFrame ocupa aprox 414 KB, un tamaño manejable para análisis de datos en memoria.

- **Análisis de duplicados:**

No se encontraron registros duplicados en el dataset.

- **Análisis de Valores Nulos:**

Se encontraron valores nulos (vacíos) en las siguientes columnas:

OverTime: 30 valores nulos
RelationshipSatisfaction: 30 valores nulos

- **Detección de Outliers**

Los outliers son valores atípicos que se encuentran fuera del rango normal. Se detectaron usando el método IQR (rango intercuartílico).

MonthlyIncome: 114 outliers detectados

TotalWorkingYears: 63 outliers detectados

YearsAtCompany 104 outliers detectados

En general, los valores atípicos detectados reflejan la diversidad y estructura real del personal, no inconsistencias en los datos.

- **Observaciones relevantes:**

Mejoras aplicadas durante el procesamiento

Con el objetivo de optimizar la calidad del dataset y asegurar su consistencia para los análisis posteriores, se aplicaron las siguientes mejoras técnicas:

Eliminación de columnas sin valor analítico

Se identificaron y eliminaron las columnas Unnamed:0, EmployeeCount, StandardHours y Over18, debido a que contienen valores constantes o no aportan información relevante para el análisis.

Tratamiento de valores nulos

Se detectó que las columnas OverTime y RelationshipSatisfaction contenían valores nulos dado que ambas son variables categóricas/ordinales y el porcentaje de nulos era bajo, se decidió imputarlos utilizando la moda (el valor más frecuente), conservando así la coherencia interna del dataset.

Esta estrategia permitió mantener la integridad de los datos sin eliminar registros y sin introducir sesgos significativos, al tiempo que se preserva la naturaleza categórica de las variables.

- **Calidad General de los Datos:**

Los datos presentan una estructura limpia y bien organizada, con un número consistente de registros. Se verificó que los tipos de datos fueran coherentes con la naturaleza de cada variable. Además, se corrigieron inconsistencias menores en nombres de columnas y formatos, mejorando la legibilidad del dataset.

Las columnas eliminadas **(**Unnamed: 0**,** EmployeeCount**,** StandardHours y Over18**)** no aportaban información relevante para el análisis.
Al excluirlas, se redujo el volumen de datos innecesarios, manteniendo solo variables significativas para futuros modelos de *People Analytics*.

Se comprobó que el dataset no contenía filas duplicadas, lo cual refuerza la fiabilidad de los registros individuales y evita sobre ponderar observaciones.

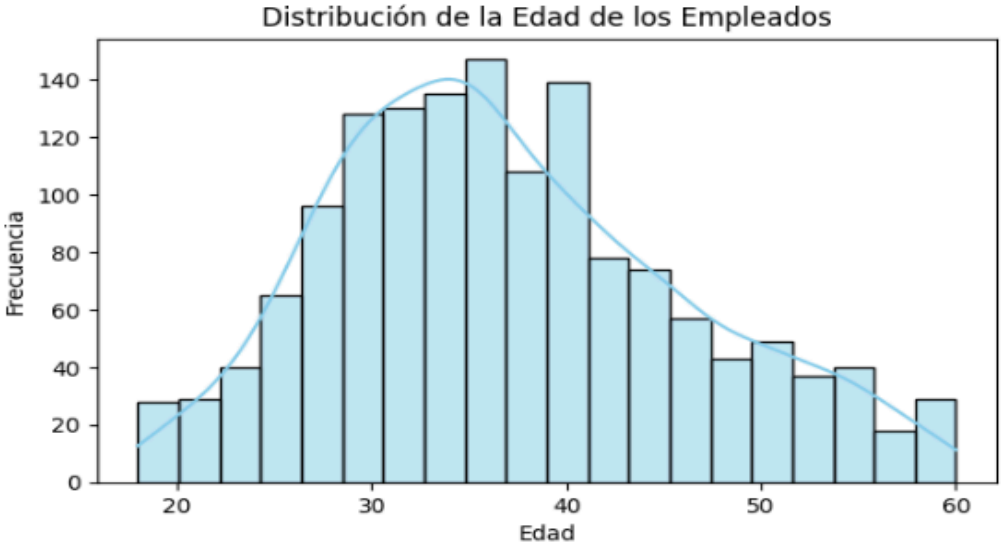
- **Posible impacto en futuros análisis**

Mayor confiabilidad en los resultados:

Al haber eliminado columnas sin relevancia analítica y haber tratado adecuadamente los valores nulos, los resultados estadísticos y modelos predictivos que se desarrollen a partir de este dataset serán más precisos y menos propensos a errores o sesgos.

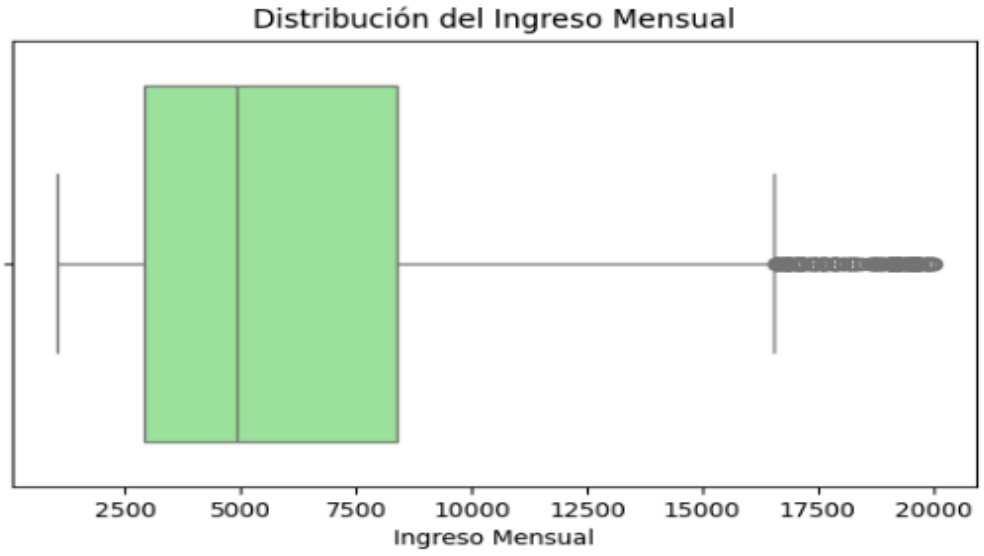
- **Gráficos de distribución**:

Distribución de la edad de los empleados



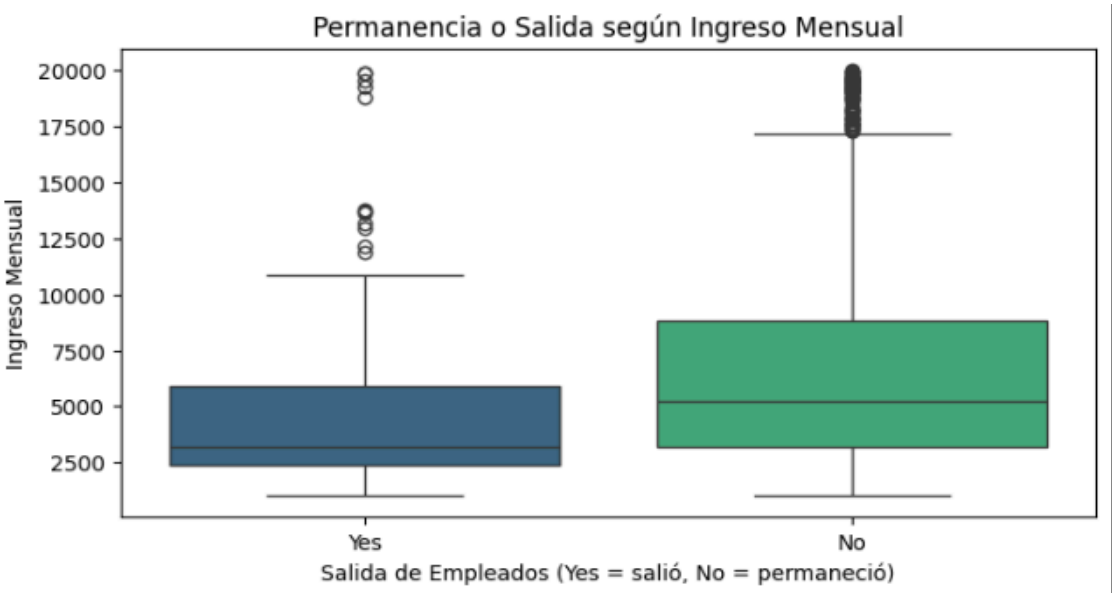
La variable Age presenta una distribución aproximadamente normal, concentrada entre los 30 y 40 años. Esto indica que la mayoría de empleados se encuentra en una etapa laboral activa y con experiencia, mientras que hay menos trabajadores jóvenes o mayores.

Distribución del ingreso mensual



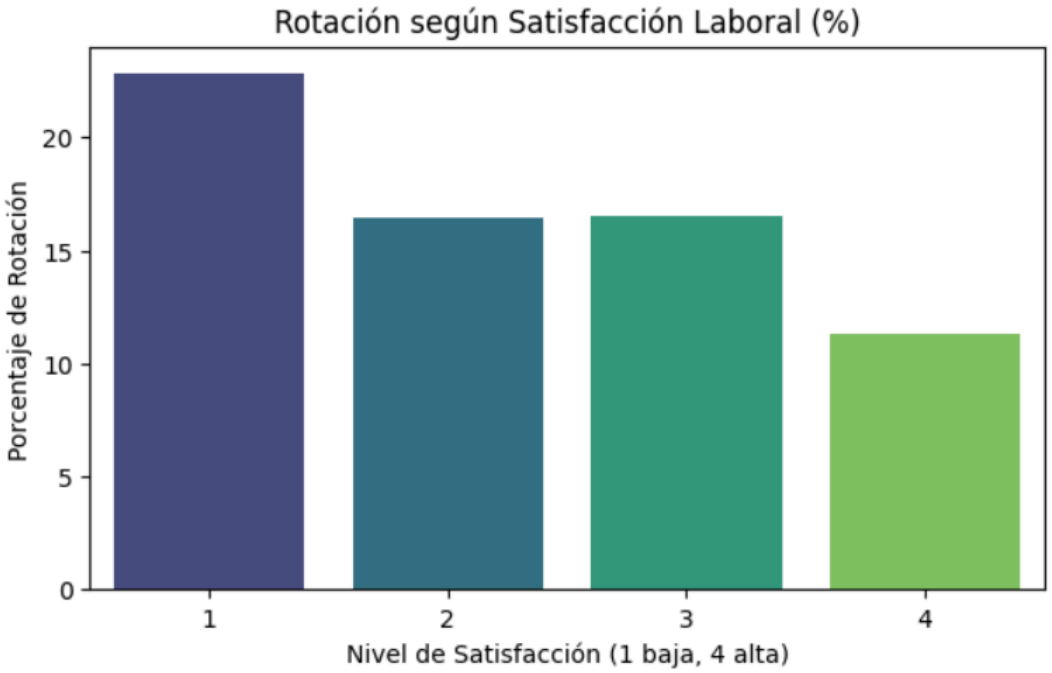
El boxplot muestra un sesgo positivo, con la mayoría de empleados percibiendo ingresos bajos o medios y unos pocos con sueldos significativamente altos. Los valores extremos reflejan cargos jerárquicos o técnicos especializados dentro de la organización.

Permanencia o salida según Ingreso Mensual



El gráfico evidencia que el nivel de ingreso influye en la permanencia de los empleados: los salarios bajos se asocian a una mayor probabilidad de salir, mientras que los ingresos medios o altos se relacionan con permanecer en la empresa.

Rotación según el nivel de Satisfacción Laboral



A medida que el nivel de satisfacción aumenta, la tasa de salida disminuye. Los empleados con satisfacción baja (nivel 1) presentan una rotación cercana al 22%, mientras que aquellos con satisfacción alta (nivel 4) apenas superan el 10%. Este patrón sugiere que la satisfacción laboral es un factor clave que influye directamente en la permanencia del personal dentro de la organización.

- **Propuesta de análisis:**

Opción 1:Análisis de Factores Demográficos que Influyen en la Renuncia

Objetivo:

Evaluar cómo características demográficas como la edad, el género y el estado civil están asociadas a la probabilidad de que un empleado renuncie a la organización.

Contexto:

Tras el proceso de limpieza y preparación del dataset, se propone analizar qué perfiles poblacionales presentan mayor tendencia a la rotación laboral. Comprender estos patrones permitirá identificar grupos con mayores riesgos de salida y apoyar la toma de decisiones en estrategias de retención.

Variables a analizar:

- Attrition (variable objetivo)
- Age
- Gender

- MaritalStatus
- YearsAtCompany

Preguntas clave:

- ¿Los empleados más jóvenes presentan una tasa de renuncia más alta?
- ¿Existen diferencias en la rotación entre hombres y mujeres?
- ¿Los empleados solteros o divorciados renuncian más que los casados?
- ¿La permanencia previa en la empresa (YearsAtCompany) modifica la relación entre perfil demográfico y renuncia?

Opción 2: Análisis de Factores Económicos y de Trayectoria que Predicen la Renuncia

Objetivo:

Identificar qué elementos relacionados con ingreso, antigüedad y desarrollo profesional influyen en que un empleado decida renunciar.

Contexto:

Con los datos depurados y organizados, se plantea investigar variables que reflejan trayectoria laboral, crecimiento y compensación económica, para comprender cómo estas afectan la permanencia.

Variables sugeridas:

- Attrition (variable objetivo)
- MonthlyIncome
- YearsAtCompany

- TotalWorkingYears
- JobLevel
- PercentSalaryHike

Preguntas de análisis:

- ¿Los empleados con menor nivel de ingresos presentan mayor tasa de renuncia?
- ¿Existe un período crítico de antigüedad donde la renuncia es más frecuente?
- ¿La falta de crecimiento laboral (pocos ascensos o niveles bajos) se asocia a mayor rotación?
- ¿Los incrementos salariales influyen en la decisión de permanecer o dejar la empresa?

Resultados Finales del Análisis

El proceso de limpieza y exploración del dataset permitió obtener una base de datos organizada, consistente y apta para análisis más avanzados. Se eliminaron columnas sin valor analítico, se trataron los valores nulos y se verificó la ausencia de registros duplicados. La detección de outliers reflejó la diversidad natural del personal sin comprometer la calidad del conjunto de datos.

Los análisis gráficos mostraron patrones claros en variables como la edad y el ingreso, y permitieron identificar factores que influyen en la permanencia o salida de los empleados, como el nivel salarial y la satisfacción laboral.

En conjunto, los procedimientos aplicados optimizaron el dataset y establecieron una base sólida para futuros análisis demográficos, económicos o predictivos, garantizando que se desarrollen sobre información confiable y representativa de la organización.