



Universidad Internacional de La Rioja
Escuela Superior de Ingeniería y
Tecnología

Máster Universitario en Análisis y Visualización de Datos

Masivos

Análisis Preventivo de la Rotación y
Equidad Laboral en un snapshot
organizacional del rubro TI

Trabajo fin de estudio presentado por:	Adolfo Fredy Huanacuni Apaza Alejandro Mario Bardales Hoyos Igor Catacora Toledo
Tipo de trabajo:	Trabajo Final de Maestría
Director/a:	Abel Coronado
Fecha:	29/09/2025

Resumen

El presente trabajo aborda la alta rotación de personal en el sector tecnológico, no como un evento aleatorio, sino como fenómeno predecible y gestionable. A través de la metodología de People Analytics, el proyecto final busca identificar los factores internos de RRHH que explican y predicen la decisión de un colaborador de abandonar la organización. Los objetivos principales son construir un modelo predictivo de clasificación para desarrollar perfiles de riesgo, establecer un ranking de las variables más influyentes y visualizar los patrones entre la satisfacción, rotación. El análisis no solo predice la rotación con base en factores como el sobretiempo (Overtime) y el bajo ingreso mensual (MonthlyIncome), sino que la auditoría de equidad revela que disparidades salariales estructurales, particularmente en el campo de formación (EducationField), exacerbaban el riesgo de fuga de personal en ciertos segmentos de talento clave.

De esta manera el proyecto valida la tesis de que un enfoque proactivo y basado en los datos es fundamental para transformar a RRHH en un socio estratégico, conectando directamente la equidad con el riesgo operativo.

Además, se concluye, que factores como el desarrollo de la carrera, el liderazgo del supervisor directo (YearsWithCurrManager) y el equilibrio vida-trabajo (WorkLifeBalance) son predictores clave, validando así la tesis de que un enfoque proactivo y basado en datos es fundamental para transformar a RRHH en un socio estratégico. En este escenario la metodología permite identificar los factores que explicarían la rotación del personal, además de tener en cuenta el perfil y performance y preferencias de los empleados, y así anticipar futuras necesidades a los perfiles de alto valor.

Palabras clave: People Analytics, Análisis predictivo, Rotación Laboral, Data Governance.

Abstract

This research addresses high employee turnover in the technology sector, treating it not as a random event but as a predictable and manageable phenomenon. Through the People Analytics methodology, this final project seeks to identify the internal HR factors that explain and predict an employee's decision to leave the organization. The main objectives are to build a predictive classification model to develop risk profiles, establish a ranking of the most influential variables, and visualize patterns between satisfaction and turnover. The analysis not only predicts turnover based on factors like Overtime and low MonthlyIncome, but the equity audit also reveals that structural salary disparities, particularly in EducationField, exacerbate the flight risk of personnel in certain key talent segments.

Thus, the project validates the thesis that a proactive, data-driven approach is fundamental to transforming HR into a strategic partner, directly connecting equity with operational risk. Furthermore, it is concluded that factors such as career development, the leadership of the direct supervisor (YearsWithCurrManager), and work-life balance (WorkLifeBalance) are key predictors. In this scenario, the methodology allows for the identification of factors that would explain staff turnover, in addition to considering the profile, performance, and preferences of employees, thereby anticipating the future needs of high-value profiles.

Keywords: People Analytics, Predictive Analysis, Employee Turnover, Data Governance.

Índice de contenidos

1. Descripción del proyecto a alto nivel	10
1.1 Introducción	10
1.2 Justificación del proyecto de Análisis	11
2. Objetivos del análisis de datos	11
2.1 Objetivos Específicos	11
3. Alcance del proyecto	12
4. Descripción de los datos	12
4.1 Fuente de datos interna	12
4.1.1. Enlace de Kaggle:	13
4.1.2. Fuente de datos Externa:	13
4.2 Descripción de las Variables de la Fuente de datos interna	13
4.3 Marco teórico y antecedentes	14
4.4 Teoría de la Integración Laboral	14
4.5 Teoría de la Contingencia	15
4.6 Teorías del Desarrollo Profesional	15
4.7 Modelo Organizacional de Persistencia del Empleado	16
4.8 People Analytics	17
4.8.1.- Definición de People Analytics	17
5. Metodología	19
5.1. Etapa 1: Diagnóstico de Factores de Rotación y Perfiles de Riesgo	19
5.1.1 Hipótesis de Trabajo:	19
5.2 Identificar y Recopilar Datos Relevantes	19
5.2.1. Variables Clave del Dataset "WA_Fn-UseC_-HR-Employee-Attrition.csv":	19
5.3 Limpiar y Preparar los Datos	20
5.3.1. Preprocesamiento de Datos:	20
5.4 Análisis Exploratorio de Datos (EDA) Enfocado en Rotación:	20
6. Análisis Exploratorio de Datos (EDA)	24
6.1 Resumen estadístico de las variables	24

6.1.2 Fuente de datos: WA_Fn-UseC_-HR-Employee-Attrition.csv con 1 https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset .	24
6.1.3 Registros y Variables: 1470 registros y 35 variables.	24
6.1.4 Análisis y tratamiento de calidad de datos	26
6.1.5 Qué Sugiere la corrección para el Análisis y Modelado:	29
6.1.6 Identificación y tratamiento de outliers	30
6.2 Análisis Gráfico	35
6.2.1. Estadísticas Descriptivas de Variables Numéricas	36
6.2.2. Análisis Univariado: Distribución de Variables Individuales	36
6.2.3. Análisis Bivariado: Relación con la Rotación (Attrition):	38
6.2.4. Análisis Multivariado: Correlaciones	39
6.3. Conclusiones del EDA	41
7. Modelado de datos	41
7.1. Objetivos del modelado (predictivo, descriptivo, prescriptivo)	41
7.1.1 Recomendaciones para el Desarrollo de Modelos	41
7.1.2 Técnicas Utilizadas	42
7.1.3 Justificación de la selección del modelo	44
7.1.4. Técnicas de validación del modelo	50
7.1.5. Datos en entrenamiento y prueba	54
7.1.6. Evaluación del rendimiento del modelo con más de una técnica.	54
7.1.7. Interpretación de los resultados del modelo.	56
8. Auditoría de la equidad laboral	59
8.1. Introducción y Objetivos de la Auditoría.	59
8.2. Metodología de Análisis Descriptivo	60
8.3. Hallazgos Detallados de Disparidades por Grupo Demográfico	60
8.3.1 Disparidades por Género (Gender)	60
8.3.2 Disparidades por Estado Civil (MaritalStatus)	61
8.3.3 Disparidades por Campo de Educación (EducationField)	61
8.3.4 Disparidades por Grupo de Edad (AgeGroup)	62
8.4. Resumen del análisis descriptivo de la Auditoría de Equidad.	63
8.5. Implicaciones y Recomendaciones Preliminares para la Equidad Laboral.	
	64
8.6. Análisis de la Tasa de Rotación por Grupos Demográficos	71

8.6.1 Introducción y Objetivo del Análisis.	71
8.6.2 Metodología de Análisis.	72
8.6.3 Hallazgos Clave por Grupo Demográfico	72
8.6.4 Interpretación de los Intervalos de Confianza (IC 95%)	76
8.6.5 Priorización y Recomendaciones Estratégicas	77
8.6.6 Monitoreo Continuo y Validación	79
8.7. Análisis inferencial	79
8.7.1 Introducción y Objetivo del Análisis inferencial.	79
8.7.2 Metodología y Especificación de los Modelos OLS	79
8.7.3 Análisis de Resultados de los Modelos OLS.	82
8.7.4 Conclusiones y Discusión	88
8.7.5 Próximos Pasos y Consideraciones Adicionales	89
8.8. Factores demográficos y de control que influyen en los años desde la última promoción.	90
8.9. Regresión binomial negativa (NB2) con enlace log y SE robustos.	90
8.9.1 Especificaciones	90
8.9.2 Tratamientos de variables	90
8.9.3 Diccionario del Modelo (Etiquetas de negocio)	90
8.9.4. Análisis de resultados,	91
<i>Tabla 2 métricas de ajuste.</i>	91
8.9.5. Principales Efectos (Etiquetas de negocio)	91
8.9.6. Gráficos de diagnóstico y sensibilidad	91
9. Análisis de equidad de género en relación con interacciones	92
9.1 Objetivos	92
9.2 Hipótesis	92
9.3. Base de Datos Usada	92
9.4. Modelos	93
9.5. Modelo 1 Base	93
9.6. Modelo 2 capital humano	93
9.7. Modelo 3 roles niveles	93
9.8. Análisis de resultados	95
9.8.1. Resultados Efecto Principal de Género (C(Gender)[T.Male])	100
9.8.2. Efecto del Campo de Educación (C(EducationField))	101
9.8.3. Efecto de la Edad (Age)	101
9.8.4. Efectos de Interacción	101

9.9. Conclusiones	102
9.10. Recomendaciones	103
9.11. Modelo de Despliegue entorno local	103
10 Prototipos	103
10.1. Factores en la rotación de empleados: Indicadores principales	103
10.2. Factores en la rotación de empleados: gráficos	104
10.2.1. Rotación por departamento	104
10.2.2. Rotación por sobretiempo (Over Time)	104
10.2.3. Rotación por edad (Boxplot)	104
10.2.4. Rotación y frecuencia de viaje de negocios	104
10.2.5. Rotación por cargo	105
10.2.6. Rotación por ingreso (Boxplot)	105
10.3. Empleados en riesgo de rotación: Gráficos	105
10.3.1. Riesgo por nivel (Gráfico de dona)	105
10.3.2. Riesgo alto por departamento (Gráfico de barras).	106
10.3.3. Empleados en riesgo (Tabla detallada)	106
10.3.4. Rotación por rol (Gráfico de barras horizontales)	106
11 Discusión General	107
11.1 Conclusiones Generales: Del Qué al Porqué de la Fuga de Talento	108
11.2 El Epicentro del Riesgo: La Intersección Demográfica Crítica	108
11.3 Recomendaciones Estratégicas para la retención y Equidad del talento.	
	109
11.4 Trabajo y Consideraciones Futuras.	110
12. Referencia o Bibliografías:	110
13. Anexo 1	112

Índice de figuras

Figura 1. Diagrama Causa-Efecto del menú de estilos.....	11
Figura 2. Descripción de variables de la fuente de datos.....	13
Figura 3. Flujos de la Metodología.	17
Figura 4. Gráfico dispersión Ingresos mensuales vs. tasa mensual.....	21
Figura 5. Outliers con el método híbrido	24
Figura 6. Comparativo de Outliers con el método híbrido y el método IQR	26
Figura 7. Estadísticas Descriptivas de Variables Numéricas.	28
Figura 8. Histogramas de Variables Numéricas.	29
Figura 9. Diagrama de caja de variables numéricas.	32
Figura 10. Variables Categóricas vs. Attrition.	34
Figura 11. Correlaciones entre las variables numéricas.	36
Figura 12. Variables incluidas en el modelo predictivo.	40
Figura 13. F2 Vs Umbral.	41
Figura 14. Matriz de confusión del modelo en el conjunto de prueba.	42
Figura 15. Curva ROC con el punto 0.30.	44
Figura 16. Coeficientes de la Regresión Logística.	46
Figura 17. Importancia Random Forest.	48
Figura 18. Importancia Gradient Boosting.	49
Figura 19. Métricas de rendimiento en el conjunto de datos de prueba.	50
Figura 20. Curva ROC. (Elaboración propia)	51
Figura 21. Gráficos auditoria (parte 1)	64
Figura 22. Gráficos auditoria (parte 2)	65
Figura 23. Gráficos auditoria (parte 3)	66
Figura 24. Gráficos auditoria (parte 4)	67

Figura 25. Gráficos auditoria (parte 4)	68
Figura 26. Gráficos auditoria (parte 4)	69
Figura 27. Tasa de rotación por AgeGroup (parte 4)	71
Figura 28. Gráficos Demográfico N° de personas vs EducationField.....	71
Figura 29. Tasa de rotación por MaritalStatus.....	72
Figura 30. Tasa de rotación por EducationField	73
Figura 31. Tasa de rotación por Gender	74
Figura 32. Gráfico de coeficientes.....	80
Figura 33. Gráfico de valores (parte 1)	81
Figura 34. Gráfico de valores (parte 2)	81
Figura 35. Efectos Género.....	82

" Análisis Preventivo de la Rotación y Equidad Laboral en un Snapshot Organizacional del rubro TI".

1. Descripción del proyecto a alto nivel

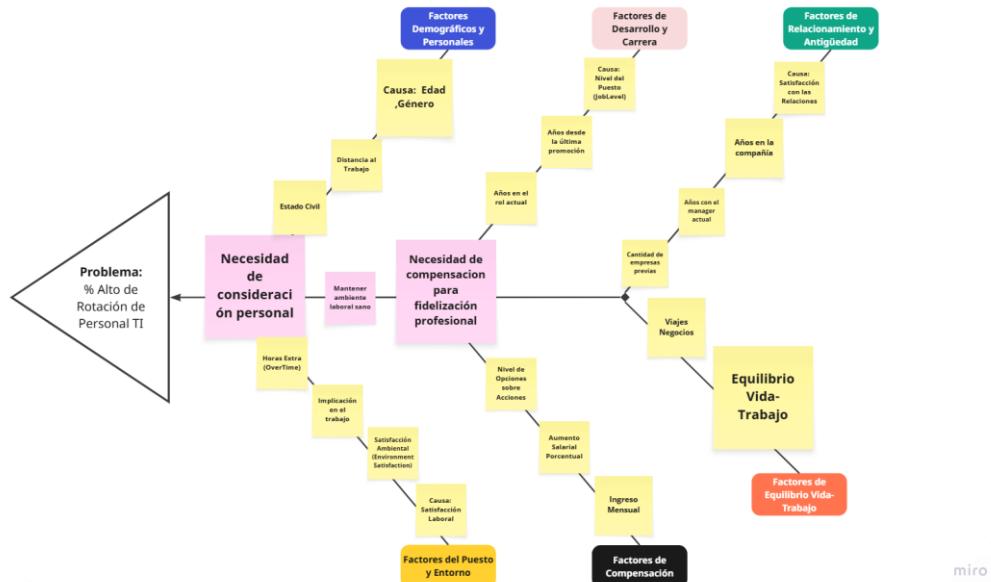
1.1 Introducción

La satisfacción laboral-profesional, en sus múltiples dimensiones, es una cuestión excelente para abordarse; adicional a las carreras, proyectos y organizaciones que se ajusten a intereses, e impactan directamente en el bienestar del personal.

Además, las organizaciones que ponen enfoque a sus colaboradores en sus procesos y objetivos empresariales no solamente generan organizaciones con mayor abundancia colectiva, sino también forman organizaciones de gestión de datos, sabiendo que estas últimas, deberían suficientemente flexibles para trabajar de manera efectiva en este entorno de evolución.

Es por ello por lo que se hace interesante que las organizaciones cuenten con modelos de pronóstico de la explicación profesional, capaces de anteponer si un nuevo asistente contará con satisfacción laboral alta o desestimada. En base a este pronóstico de solución podría permitirse a las organizaciones gestionar decisiones anticipadas en aquellos colaboradores con una alta probabilidad de baja satisfacción y evitar desistimiento laboral.

Figura 1 Diagrama Causa-Efecto (Ishikawa)



Fuente: Elaboración propia

1.2 Justificación del proyecto de Análisis

La presente investigación se justifica en base a 3 ejes críticos para la gestión de toda empresa actual: alta competitividad de talento, impacto financiero de la rotación, y necesidad de evolucionar a RRHH a una entidad proactiva y basada en datos.

La implementación del resultado de análisis de regresión como un sistema de recomendación o alerta temprana, recomendaría acciones preventivas a la alta gerencia y RRHH a identificar, diseñar e implementar intervenciones focalizadas y personalizadas antes que la insatisfacción se consolide y se evite la renuncia del personal. Esta investigación es pertinente y necesaria, porque aborda un problema de alto impacto económico, genera una propuesta de solución metodológica.

2. Objetivos del análisis de datos

2.1 Objetivos Específicos

- Determinar un modelo sistémico, prediciendo el riesgo o incertidumbre de fuga de personal de alto valor en organizaciones.
- Construir perfiles de empleados con mayor y menor riesgo de rotación.
- Visualizar patrones entre satisfacción, compromiso y rotación.

- Identificar si existen sesgos o disparidades en las condiciones laborales dentro de una organización.

3. Alcance del proyecto

Este proyecto se centrará en realizar un análisis diagnóstico y correlacional de la rotación y la equidad laboral dentro de una organización, utilizando una base de datos transversal (snapshot en un punto del tiempo) provista. El estudio abarcará tres etapas principales:

Etapa 1: Diagnóstico de Factores de Rotación y Perfiles de Riesgo: Se identificarán los factores internos más influyentes asociados a la rotación de empleados en el momento del snapshot. Esto incluirá la construcción de perfiles detallados (personas) de los empleados con mayor y menor riesgo de rotación, y la visualización de las interrelaciones entre las variables de satisfacción, compromiso y rotación.

Etapa 2: Auditoría de Equidad Laboral: Se detectarán y analizarán posibles disparidades significativas en la compensación y la progresión de carrera entre diferentes grupos demográficos (género, estado civil, campo de educación) dentro del dataset, así como las diferencias en la tasa de rotación entre estos grupos, en el momento del snapshot.

Etapa 3: Contextualización Externa y Benchmarking: Se comparará la situación interna de la empresa (salarios, rotación) con el contexto macroeconómico y los benchmarks de la industria relevantes para el mismo período del snapshot. Esto permitirá identificar cómo las condiciones externas o la competitividad de la empresa pueden influir en la rotación observada.

En el presente trabajo de investigación se plantea abarcar hasta la Etapa 2 como el objetivo principal y alcance esperado, y la Etapa 3 como análisis complementarios que se desarrollarán en la medida que el tiempo lo permita.

4. Descripción de los datos

4.1 Fuente de datos interna

WA_Fn-UseC_-HR-Employee-Attrition.csv.

4.1.1. Enlace de Kaggle:

<https://www.kaggle.com/datasets/chandramaniwagh/employee-attrition>

4.1.2. Fuente de datos Externa:

Bancos Centrales Nacionales, Institutos Nacionales de Estadística, Organizaciones Internacionales: Fondo Monetario Internacional (World Economic Outlook Database). Banco Mundial (World Bank) (World Development Indicators - WDI). Organización Internacional del Trabajo (OIT).

4.2 Descripción de las Variables de la Fuente de datos interna

El dataset contiene 1470 observaciones y 35 variables. Si desea más detalle de las variables en mención ver Anexo 1. A continuación, se detalla la descripción de las variables relevantes para el análisis de rotación de empleados:

Figura 2. Descripción de variables de la fuente de datos

Nombre tecnico	Nombre amigable	Descripcion	Tipo de dato
Age	Edad	Edad del empleado en años	INTEGER
Attrition	Rotacion	Si el empleado ha abandonado la empresa (Yes = Sí, No = No)	BOOLEAN
BusinessTravel	ViajesNegocios	Frecuencia de viajes de negocios (Non-Travel = No viaja; Travel_Frequently = Viaja frecuentemente; Travel_Rarely = Rara vez viaja)	VARCHAR
DailyRate	TasaDiaria	Tarifa diaria del empleado	INTEGER
Department	Departamento	Departamento del empleado (Sales, Research & Development, Human Resources)	VARCHAR
DistanceFromHome	DistanciaCasa	Distancia desde casa al trabajo en millas	INTEGER
Education	NivelEducacion	Nivel educativo (1 = Inferior a preparatoria, 2 = Preparatoria, 3 = Licenciatura, 4 = Maestría, 5 = Doctorado)	INTEGER
EducationField	AreaEducacion	Área de estudio (Life Sciences, Medical, Marketing, Technical Degree, Human Resources, Other)	VARCHAR
EmployeeCount	ConteoEmpleados	Contador de empleados (valor constante 1)	INTEGER
EmployeeNumber	NumeroEmpleado	Identificador único del empleado	INTEGER
EnvironmentSatisfaction	SatisfaccionEntorno	Satisfacción con el entorno laboral (1 = Bajo, 2 = Medio, 3 = Alto, 4 = Muy alto)	INTEGER
Gender	Genero	Género del empleado (Male = Hombre, Female = Mujer)	VARCHAR
HourlyRate	TarifaHora	Tarifa por hora	INTEGER
JobInvolvement	CompromisoTrabajo	Grado de involucramiento en el trabajo (1 = Bajo, 2 = Medio, 3 = Alto, 4 = Muy alto)	INTEGER

Nombre tecnico	Nombre amigable	Descripcion	Tipo de dato
JobLevel	NivelTrabajo	Nivel del puesto (1 al 5)	INTEGER
JobRole	RolTrabajo	Rol del empleado (Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director, Human Resources)	VARCHAR
JobSatisfaction	SatisfaccionTrabajo	Satisfacción con el trabajo (1 = Bajo, 2 = Medio, 3 = Alto, 4 = Muy alto)	INTEGER
MaritalStatus	EstadoCivil	Estado civil (Single = Soltero, Married = Casado, Divorced = Divorciado)	VARCHAR
MonthlyIncome	IngresoMensual	Ingreso mensual	INTEGER
MonthlyRate	TasaMensual	Tarifa mensual	INTEGER
NumCompaniesWorked	NumEmpresasTrabajadas	Número de empresas en las que ha trabajado	INTEGER
Over18	Mayor18	Indicador de mayor de 18 años (valor constante Y)	BOOLEAN
Overtime	HorasExtra	Si trabaja horas extra (Yes = Sí, No = No)	BOOLEAN
PercentSalaryHike	AumentoPorcentualSalario	Porcentaje de aumento salarial	INTEGER
PerformanceRating	CalificacionRendimiento	Calificación de rendimiento (3 = Bueno, 4 = Excelente)	INTEGER
RelationshipSatisfaction	SatisfaccionRelacion	Satisfacción con las relaciones (1 = Bajo, 2 = Medio, 3 = Alto, 4 = Muy alto)	INTEGER
StandardHours	HorasEstandar	Horas estándar de trabajo (valor constante 80)	INTEGER
StockOptionLevel	NivelOpcionesAccion	Nivel de opciones sobre acciones (0 a 3)	INTEGER
TotalWorkingYears	AñosExperienciaTotal	Años totales de experiencia laboral	INTEGER
TrainingTimesLastYear	CapacitacionesUltimoAño	Número de capacitaciones en el último año	INTEGER
WorkLifeBalance	EquilibrioVidaTrabajo	Balance vida-trabajo (1 = Bajo, 2 = Medio, 3 = Alto, 4 = Excelente)	INTEGER
YearsAtCompany	AñosEnEmpresa	Años en la empresa	INTEGER
YearsInCurrentRole	AñosEnRolActual	Años en el rol actual	INTEGER
YearsSinceLastPromotion	AñosDesdeUltAscenso	Años desde el último ascenso	INTEGER
n	AñosConGerenteActual	Años con el gerente actual	INTEGER

Fuente: Elaboración propia

4.3 Marco teórico y antecedentes

El presente trabajo de investigación se sustenta en diversas teorías que proporcionan una comprensión integral de la rotación de personal y la equidad laboral. Estas teorías ayudan a identificar los factores asociados con la rotación de personal y las disparidades en la equidad laboral dentro de las organizaciones. Las principales teorías incluyen la integración laboral, la teoría de contingencias y las teorías de desarrollo profesional, cada una de las cuales ofrece perspectivas únicas sobre la dinámica de la rotación y retención de personal. Además, consideramos que el People Analytics proporciona un marco metodológico útil para el presente estudio.

4.4 Teoría de la Integración Laboral

Esta teoría, desarrollada inicialmente por Mitchell et al. (2001) y expandida por otros autores, sugiere que los empleados deciden permanecer en sus puestos de trabajo debido a una combinación de factores que los integran en su organización, como los vínculos (conexiones con otras personas en la organización), la adaptación (que tan

bien encaja el empleado en su puesto) y el sacrificio (los costos materiales o psicológicos de dejar el empleo). Esta enfatiza que la decisión de irse no solo depende de la satisfacción laboral, sino de un conjunto más amplio de factores sociales y personales (Hom et al., 2019). En resumen, este marco es fundamental para identificar variables internas de RRHH, que predicen la rotación, ya que permite analizar indicadores de integración como los años en la empresa, la relación con el mánager y la satisfacción con el entorno.

4.5 Teoría de la Contingencia

La teoría de la contingencia sostiene que no existe una única forma correcta de estructurar una organización o gestionar al personal. En cambio, la efectividad depende de la adecuación entre la organización y su entorno, así como entre los individuos y el organigrama. Aplicado a la rotación, este enfoque menciona que el desajuste entre los valores de un empleado y la cultura aumenta la probabilidad de salida (Zeffane, 1994). En el presente proyecto esta teoría permitirá visualizar patrones entre la satisfacción, compromiso y la rotación, al analizar cómo la percepción del entorno laboral se alinea con las características y expectativas de los perfiles de empleados.

4.6 Teorías del Desarrollo Profesional

Teorías como el modelo de carrera de Holland y la teoría de la carrera sociocognitiva ofrecen información sobre cómo las elecciones y el desarrollo profesional impactan en la rotación de personal (Hom et al., 2010).

El modelo de carrera de Holland y la teoría sociocognitiva de la carrera, pueden servir de apoyo a su estudio sobre la rotación de personal. El modelo de Holland se centra en la adecuación entre la persona y el entorno, lo que influye en la satisfacción y el compromiso laboral. La teoría sociocognitiva de la carrera destaca el papel de la autoeficacia y las expectativas de resultados en la elección de carrera, lo que puede ayudar a identificar variables internas de RR. HH. que predicen la rotación y los perfiles de empleados en riesgo.

Estas teorías pueden ayudar a construir perfiles de empleados con diferentes riesgos de rotación, al comprender cómo las aspiraciones profesionales y la satisfacción laboral influyen en la retención.

4.7 Modelo Organizacional de Persistencia del Empleado

Este modelo destaca la relación recíproca entre las prácticas organizacionales y variables individuales como la intención, los objetivos, el compromiso y la satisfacción (Peterson, 2004).

Subraya el papel estratégico del desarrollo de recursos humanos en la reducción de los costos de rotación y la mejora de la retención de empleados mediante intervenciones específicas.

El artículo de Peterson (2004) analiza el Modelo Organizacional de Persistencia del Empleado, que integra variables individuales (intención, objetivos, compromiso, satisfacción) con factores organizacionales para explicar la rotación de personal. Se basa en el modelo de Tinto sobre la salida institucional, enfatizando la relación recíproca entre la integración organizacional y los atributos individuales. Este marco apoya los objetivos de su estudio al identificar variables internas de RR. HH. que predicen la rotación y analizar los perfiles de los empleados, lo que proporciona información sobre la equidad laboral y los patrones de rotación dentro de la organización.

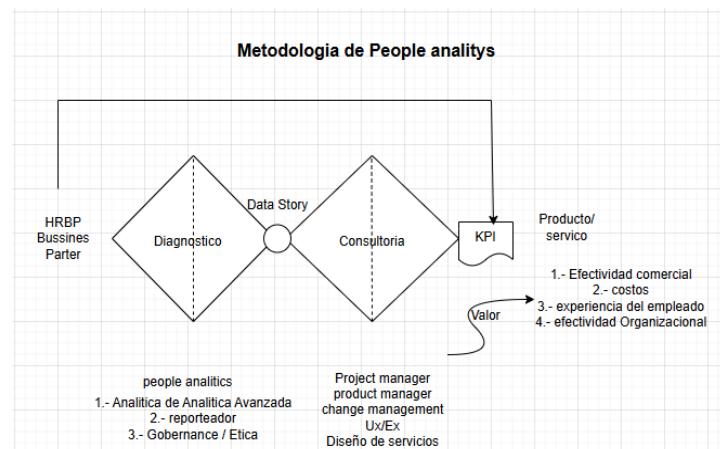
Si bien estas teorías proporcionan un marco sólido para comprender la rotación de personal, también es importante considerar el papel de factores externos, como las condiciones económicas y las tendencias del sector, que pueden influir en las tasas de rotación y la equidad laboral. Además, la integración de conocimientos derivados de las teorías vocacionales y de desarrollo profesional puede enriquecer aún más la comprensión de las estrategias de retención y retiro organizacional.

4.8 People Analytics

4.8.1.- Definición de People Analytics

Según Soria-Olivas, E., et al. (2024). Define a People Analytics como: “una disciplina que utiliza datos cuantitativos y métodos analíticos para mejorar las decisiones organizacionales sobre el talento humano”.

Figura 3 Flujos de la Metodología



Fuente: Elaboración propia

Tabla 1. Metodología de People Analytics: Fases y Descripción del Proceso

Nº	Fases	Detalle
1	Definir Preguntas de Negocio y Objetivos	Establecer claramente qué preguntas específicas se quieren responder y qué objetivos estratégicos se buscan alcanzar.
2	Identificar y Recopilar Datos Relevantes	se debe determinar los datos necesarios para poder responderlos, podría incluir datos de contratación, desempeño, compensación, encuestas de clima, capacitación, etc.
3	Limpiar y Preparar los Datos	Implica limpiar inconsistencias, manejar valores faltantes, estandarizar formatos y transformar los datos para que sean útiles para el análisis
4	Analizar los Datos	aplicar técnicas estadísticas o de modelado para encontrar patrones, correlaciones y tendencias. El objetivo es descubrir insights que respondan a las preguntas iniciales.

5 Interpretar y Visualizar los Resultados	se debe interpretar lo que significan tus hallazgos en el contexto del negocio. se presentará los resultados en visualizaciones que faciliten la comprensión a todas las partes interesadas.
6 Comunicar las Conclusiones y Recomendar Acciones	Comunicar los insights encontrados con las partes interesadas para poder apoyarlos en la toma de decisiones, explicando qué significan los hallazgos e indicar qué acciones específicas se pueden tomar basándose en ellos
7 Implementar y Monitorear	Es fundamental monitorear su impacto. Esto permite evaluar si las intervenciones fueron exitosas y ajustar la estrategia si es necesario,

Fuente: Elaboración propia

Según Mejía Arias, L. K. (2020), la aplicación de People Analytics en las organizaciones modernas sirven para optimizar la gestión del talento humano. Así mismo, enfatiza cómo la analítica de datos aplicada al ámbito de los recursos humanos permite a las empresas pasar de una toma de decisiones intuitiva a una basada en evidencia, lo que resulta en una mayor eficiencia y efectividad. Además, se destaca que People Analytics proporciona herramientas para comprender mejor el comportamiento de los empleados, identificar patrones, predecir tendencias, mejorar aspectos cruciales como la adquisición de talento, la retención, el desarrollo profesional, la compensación y la planificación de la fuerza laboral. En síntesis, People Analytics no solo es una herramienta de apoyo, sino comprende una necesidad estratégica ineludible para la competitividad empresarial en el entorno actual.

Según Massoni, T., et al. (2019), explora la rotación voluntaria de desarrolladores de software; y encuentra niveles bajos de satisfacción laboral que se asocian a mayor propensión a dejar la empresa. Aunque no aborda directamente la satisfacción ambiental o relacional, sus hallazgos son consistentes, y que, en el presente proyecto, explica la idea de que una satisfacción integral reduce la rotación.

5. Metodología

Para el presente trabajo usaremos la metodología de People Analytics. Adicionalmente, para analizar los factores que influyen en la duración de la permanencia de los empleados, se implementó un modelo de Regresión Binomial Negativa (NB2), utilizando la variable YearsAtCompany como variable dependiente. Se seleccionó este método por su idoneidad para modelar datos de conteo con alta dispersión. En el próximo apartado se definen los pasos de la metodología.

5.1. Etapa 1: Diagnóstico de Factores de Rotación y Perfiles de Riesgo

Esta etapa se centra en desentrañar las causas y los perfiles de los empleados asociados a la rotación, utilizando únicamente los datos internos de la empresa en el punto de tiempo dado.

5.1.1 Hipótesis de Trabajo:

H1.1: Existe una relación inversa entre la satisfacción (Satisfaction) (laboral, ambiental, relacional) y la probabilidad de rotación (Attrition).

H1.2: El trabajo excesivo (OverTime) y un bajo equilibrio entre vida laboral y personal (WorkLifeBalance) están positivamente asociados con la rotación (Attrition).

H1.3: La falta de progresión de carrera (ej., YearsSinceLastPromotion altos o YearsInCurrentRole altos para JobLevel bajos) está positivamente asociada con la rotación.

5.2 Identificar y Recopilar Datos Relevantes

5.2.1. Variables Clave del Dataset "WA_Fn-UseC-HR-Employee-Attrition.csv":

Variable Dependiente: Attrition (Sí/No)

Variables Independientes(Potencialmente influyentes): Age, Gender, MaritalStatus, Education, EducationField, Department, JobRole, JobLevel, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, NumCompaniesWorked, BusinessTravel, OverTime, MonthlyIncome, HourlyRate, DailyRate, MonthlyRate, PercentSalaryHike, StockOptionLevel, EnvironmentSatisfaction, JobSatisfaction, RelationshipSatisfaction, WorkLifeBalance, JobInvolvement, PerformanceRating, DistanceFromHome, TrainingTimesLastYear.

5.3 Limpiar y Preparar los Datos

5.3.1. Preprocesamiento de Datos:

- Limpieza: Identificación y manejo de valores atípicos (outliers) y posibles inconsistencias en los datos.
- Codificación de Variables Categóricas: Convertir variables como Gender, Department, JobRole, BusinessTravel, OverTime a formatos numéricos (ej., One-Hot Encoding para variables nominales, Label Encoding u ordinal para variables ordinales).
- Escalado de Variables Numéricas: Estandarizar o normalizar variables numéricas (ej., Age, MonthlyIncome, DistanceFromHome) para que los modelos de Machine Learning no se vean sesgados por la escala de los datos.

5.4 Análisis Exploratorio de Datos (EDA) Enfocado en Rotación:

Estadísticas Descriptivas: Calcular la tasa general de rotación. Obtener estadísticas (media, mediana, desviación estándar) para variables numéricas y frecuencias para categóricas, diferenciando entre empleados que rotaron y los que no.

Visualizaciones:

- Gráficos de barras para Attrition vs. variables categóricas (ej., Attrition por Department, JobRole, OverTime).
- Box plots o histogramas para Attrition vs. variables numéricas (ej., Attrition vs. MonthlyIncome, JobSatisfaction).
- Matriz de correlación (heatmap) entre todas las variables numéricas para identificar relaciones lineales.

Modelado Predictivo y Análisis de Importancia de Características:

- Selección de Modelo: Entrenar y evaluar modelos de clasificación para predecir Attrition. Con modelos robustos como Random Forest o Gradient Boosting (ej., XGBoost, LightGBM), y también Regresión Logística como línea base.

Construcción de segmentos y perfiles de Riesgo:

- Análisis de Segmentación (Clustering): Aplicar algoritmos de clustering (ej., K-Means, DBSCAN) sobre las características de los empleados identificados con alta probabilidad de rotación por el modelo predictivo. Esto ayudará a agrupar a los empleados con características similares en "clústeres de riesgo".
- **Creación de Personas:** Describir detalladamente los perfiles de los empleados en cada clúster de riesgo, incluyendo sus características demográficas, laborales, de satisfacción y de compensación, para crear "personas de alto riesgo de rotación".

Etapa 2: Auditoría de Equidad Laboral (Análisis Estático de Disparidades).

- Esta etapa se enfoca en identificar si existen sesgos o disparidades en las condiciones laborales dentro de la organización en el momento del snapshot, lo cual podría influir indirectamente en la rotación.

Objetivos Específicos:

- Detectar posibles disparidades significativas en la compensación (MonthlyIncome, PercentSalaryHike) y la progresión de carrera (YearsSinceLastPromotion) entre diferentes grupos demográficos (género, estado civil, campo de educación).
- Analizar si la tasa de rotación es significativamente diferente entre estos grupos demográficos, sugiriendo posibles inequidades en el ambiente laboral.

Variables Clave del Dataset "WA_Fn-UseC_-HR-Employee-Attrition.csv":

- **Variables Demográficas de Interés:** Gender, MaritalStatus, EducationField, Age.
- **Variables de Resultado:** MonthlyIncome, PercentSalaryHike, YearsSinceLastPromotion, Attrition.
- **Variables de Control (para asegurar comparaciones justas):** JobLevel, TotalWorkingYears, YearsAtCompany, JobRole.

Hipótesis de Trabajo:

- H2.1: Existen diferencias salariales significativas (MonthlyIncome) no explicadas por el nivel de puesto o la experiencia (JobLevel, TotalWorkingYears) entre empleados de diferente Gender o EducationField.
- H2.2: La tasa de Attrition es significativamente mayor en ciertos grupos demográficos (Gender, MaritalStatus) después de controlar por otros factores de riesgo de rotación.
- H2.3: Ciertos grupos demográficos experimentan períodos más largos sin una promoción (YearsSinceLastPromotion) en comparación con otros grupos con perfiles de experiencia similares.

Metodología de la Etapa 2:

Preparación de Datos para Análisis de Equidad:

- Asegurarse de que las variables categóricas estén correctamente codificadas.
- Considerar la creación de rangos de edad o experiencia si es pertinente para el análisis de grupos.

Análisis Estadístico de Disparidades:

- **Ánalisis Descriptivo por Grupos:** Calcular medias, medianas y desviaciones estándar de MonthlyIncome, PercentSalaryHike, YearsSinceLastPromotion y la

tasa de Attrition para cada categoría de las variables demográficas (Gender, MaritalStatus, EducationField).

- **Pruebas de Hipótesis:**

- **ANOVA o T-tests:** Para comparar las medias de MonthlyIncome, PercentSalaryHike, YearsSinceLastPromotion entre más de dos grupos (ANOVA) o dos grupos (T-test) de variables demográficas, controlando por variables como JobLevel o TotalWorkingYears (ej., usando un ANCOVA si es necesario).
- **Chi-cuadrado:** Para evaluar si hay una asociación significativa entre Attrition y las variables demográficas.
- **Regresión Lineal/Logística con Variables Demográficas:** Construir modelos de regresión donde el resultado (MonthlyIncome, Attrition, YearsSinceLastPromotion) es explicado por las variables demográficas y las variables de control, para cuantificar la magnitud de cualquier disparidad.

- **Visualización de Disparidades:**

- Gráficos de barras comparando promedios (ej., MonthlyIncome por Gender y JobLevel).
- Mapas de calor de tasas de rotación por combinaciones de variables demográficas.

Etapa 3: Contextualización Externa y Benchmarking (Análisis Comparativo del Snapshot)

Esta etapa busca poner los hallazgos internos en un contexto más amplio, comparando los datos del snapshot de la empresa con información externa relevante del mismo período. Esta etapa no se encuentra en el alcance de estudio.

Objetivos Específicos: Evaluar cómo la situación interna de la empresa (salarios, rotación) se compara con el contexto macroeconómico y los benchmarks de la industria en el mismo período del snapshot.

- Identificar si las condiciones externas o la competitividad de la empresa pueden estar influyendo en la rotación observada.

Variables Clave del Dataset "WA_Fn-UseC_-HR-Employee-Attrition.csv":

- MonthlyIncome (para comparar con salarios de mercado).
- Attrition (para comparar con tasas de rotación de la industria).
- JobRole, JobLevel (para contextualizar la comparación).

Variables Externas (A buscar para el mismo período del snapshot, ej. año 2017/2018):

- Tasa de Desempleo (nacional o sectorial relevante para la empresa).

- Índice de Costo de Vida (si la ubicación de los empleados puede ser inferida o es una sola conocida).
 - Salario Promedio de la Industria para roles similares (JobRole, JobLevel).
 - Tasa de Rotación Promedio de la Industria para el sector de la empresa.

Hipótesis de Trabajo:

- **H3.1:** La tasa de rotación de la empresa es más alta que el promedio de la industria en el mismo período, sugiriendo problemas internos más allá del contexto general del mercado.
- **H3.2:** Existe una correlación entre el MonthlyIncome de los empleados que rotaron y el Salario Promedio de la Industria para roles similares, indicando una posible falta de competitividad salarial en ese momento.
- **H3.3:** Un Índice de Costo de Vida alto en la ubicación de los empleados (si aplicable) está asociado con una mayor rotación, a menos que sea compensado por un MonthlyIncome competitivo.

Metodología Detallada de la Etapa 3:

Recolección de Datos Externos:

- Investigar y obtener datos macroeconómicos, de mercado laboral y de benchmarking de la industria que sean contemporáneos al snapshot del dataset de IBM. Esto puede implicar el uso de APIs de organismos estadísticos, bases de datos de consultoras, o reportes de la industria disponibles públicamente.
- Registrar la fuente y la fecha de los datos externos para asegurar su validez y correspondencia temporal.

Armonización de Datos:

Hay que asegurar que las unidades de medida y las definiciones de los datos externos sean compatibles con las del dataset interno (ej., salarios anuales vs. mensuales, categorías de roles).

Manejar la agregación: Si los datos internos son a nivel de empleado y los externos a nivel de industria/ciudad, la comparación se hará entre promedios o agregaciones.

Análisis Comparativo y de Correlación:

- Comparación de Medias/Tasas: Comparar directamente la tasa de rotación promedio de la empresa con la de la industria. Comparar el MonthlyIncome promedio de roles clave en la empresa con los salarios de mercado para roles equivalentes.
- Análisis de Correlación: Si tienes varios puntos de datos externos que pueden correlacionarse con una métrica agregada interna (ej., rotación anual de la

empresa vs. desempleo anual del sector), puedes realizar un análisis de correlación simple.

- Visualizaciones Comparativas: Gráficos de barras que comparan el promedio de la empresa con el promedio de la industria para métricas clave (salario, rotación), o scatter plots que muestren la relación entre variables internas y externas si hay suficiente granularidad.

6. Análisis Exploratorio de Datos (EDA)

En este capítulo se detalla el Análisis Exploratorio de Datos (EDA) realizado sobre el conjunto de datos de rotación de empleados. El objetivo principal del EDA es obtener una comprensión profunda de la estructura de los datos, identificar patrones y relaciones entre las variables, detectar anomalías y, crucialmente, explorar los factores que podrían influir en la decisión de un empleado de rotar. Esta fase es fundamental para informar las decisiones de preprocesamiento, selección de características y el desarrollo de modelos predictivos robustos.

6.1 Resumen estadístico de las variables

6.1.2 Fuente de datos: WA_Fn-UseC_-HR-Employee-Attrition.csv con 1
<https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>.

6.1.3 Registros y Variables: **1470 registros y 35 variables.**

Figura 4 Descripción de variables y tipo de dato

Nombre tecnico	Nombre amigable	Descripcion	Tipo de dato
Age	Edad	Edad del empleado en años	INTEGER
Attrition	Rotacion	Si el empleado ha abandonado la empresa (Yes = Sí, No = No)	BOOLEAN
BusinessTravel	ViajesNegocios	Frecuencia de viajes de negocios (Non-Travel = No viaja; Travel_Frequently = Viaja frecuentemente; Travel_Rarely = Rara vez viaja)	VARCHAR
DailyRate	TasaDiaria	Tarifa diaria del empleado	INTEGER
Department	Departamento	Departamento del empleado (Sales, Research & Development, Human Resources)	VARCHAR
DistanceFromHome	DistanciaCasa	Distancia desde casa al trabajo en millas	INTEGER
Education	NivelEducacion	Nivel educativo (1 = Inferior a preparatoria, 2 = Preparatoria, 3 = Licenciatura, 4 = Maestría, 5 = Doctorado)	INTEGER
EducationField	AreaEducacion	Área de estudio (Life Sciences, Medical, Marketing, Technical Degree, Human Resources, Other)	VARCHAR
EmployeeCount	ConteoEmpleados	Contador de empleados (valor constante 1)	INTEGER
EmployeeNumber	NumeroEmpleado	Identificador único del empleado	INTEGER
EnvironmentSatisfaction	SatisfaccionEntorno	Satisfacción con el entorno laboral (1 = Bajo, 2 = Medio, 3 = Alto, 4 = Muy alto)	INTEGER
Gender	Genero	Género del empleado (Male = Hombre, Female = Mujer)	VARCHAR
HourlyRate	TarifaHora	Tarifa por hora	INTEGER
JobInvolvement	CompromisoTrabajo	Grado de involucramiento en el trabajo (1 = Bajo, 2 = Medio, 3 = Alto, 4 = Muy alto)	INTEGER

Nombre tecnico	Nombre amigable	Descripcion	Tipo de dato
JobLevel	NivelTrabajo	Nivel del puesto (1 al 5)	INTEGER
JobRole	RolTrabajo	Rol del empleado (Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director, Human Resources)	VARCHAR
JobSatisfaction	SatisfaccionTrabajo	Satisfacción con el trabajo (1 = Bajo, 2 = Medio, 3 = Alto, 4 = Muy alto)	INTEGER
MaritalStatus	EstadoCivil	Estado civil (Single = Soltero, Married = Casado, Divorced = Divorciado)	VARCHAR
MonthlyIncome	IngresoMensual	Ingreso mensual	INTEGER
MonthlyRate	TasaMensual	Tarifa mensual	INTEGER
NumCompaniesWorked	NumEmpresasTrabajadas	Número de empresas en las que ha trabajado	INTEGER
Over18	Mayor18	Indicador de mayor de 18 años (valor constante Y)	BOOLEAN
Overtime	HorasExtra	Si trabaja horas extra (Yes = Sí, No = No)	BOOLEAN
PercentSalaryHike	AumentoPorcentualSalario	Porcentaje de aumento salarial	INTEGER
PerformanceRating	CalificacionRendimiento	Calificación de rendimiento (3 = Bueno, 4 = Excelente)	INTEGER
RelationshipSatisfaction	SatisfaccionRelacion	Satisfacción con las relaciones (1 = Bajo, 2 = Medio, 3 = Alto, 4 = Muy alto)	INTEGER
StandardHours	HorasEstandar	Horas estándar de trabajo (valor constante 80)	INTEGER
StockOptionLevel	NivelOpcionesAccion	Nivel de opciones sobre acciones (0 a 3)	INTEGER
TotalWorkingYears	AñosExperienciaTotal	Años totales de experiencia laboral	INTEGER
TrainingTimesLastYear	CapacitacionesUltimoAño	Número de capacitaciones en el último año	INTEGER
WorkLifeBalance	EquilibrioVidaTrabajo	Balance vida-trabajo (1 = Bajo, 2 = Medio, 3 = Alto, 4 = Excelente)	INTEGER
YearsAtCompany	AñosEnEmpresa	Años en la empresa	INTEGER
YearsInCurrentRole	AñosEnRolActual	Años en el rol actual	INTEGER
YearsSinceLastPromotion	AñosDesdeUltAscenso	Años desde el último ascenso	INTEGER
YearsWithCurrManager	AñosConGerenteActual	Años con el gerente actual	INTEGER

Fuente: Elaboración propia

6.1.4 Análisis y tratamiento de calidad de datos

- Análisis de valores faltantes.

Se hizo una verificación completa: 0 valores nulos. No fue necesaria imputación.

- Análisis de la consistencia lógica de las variables.

Se aplicaron las siguientes reglas de consistencia lógica y se muestran los casos encontrados:

- YearsAtCompany > TotalWorkingYears: 0
- MonthlyIncome < MonthlyRate (valores invertidos): 1212 registros corregidos
- YearsInCurrentRole > TotalWorkingYears: 0
- YearsSinceLastPromotion > TotalWorkingYears: 0
- YearsWithCurrManager > YearsAtCompany: 0

- YearsInCurrentRole > YearsAtCompany: 0
- YearsSinceLastPromotion > YearsAtCompany: 0
- Variables categóricas fuera de rango: 0
- EnvironmentSatisfaction: 0 fuera de rango
- JobInvolvement: 0 fuera de rango
- JobSatisfaction: 0 fuera de rango
- PerformanceRating: 0 fuera de rango
- RelationshipSatisfaction: 0 fuera de rango
- WorkLifeBalance: 0 fuera de rango

Problema identificado: Que 1,212 de 1,470 registros (aproximadamente el 82% del dataset) tuvieran MonthlyIncome < MonthlyRate es un porcentaje altísimo. Esto no es un error aleatorio; sugiere un problema fundamental en cómo se recopilaron, definieron o procesaron estas dos variables. Sin embargo, dado que no podemos acceder a los creadores del dataset no podemos conocer la razón de semejante inconsistencia. ¿Por qué es una inconsistencia? Si MonthlyIncome se define como el ingreso total mensual y MonthlyRate como una tarifa o ingreso base, es lógicamente imposible que el ingreso total sea menor que la tarifa base. Dejar los datos como están significa que el modelo que usemos estaría aprendiendo de una relación ilógica, lo que puede confundirlo.

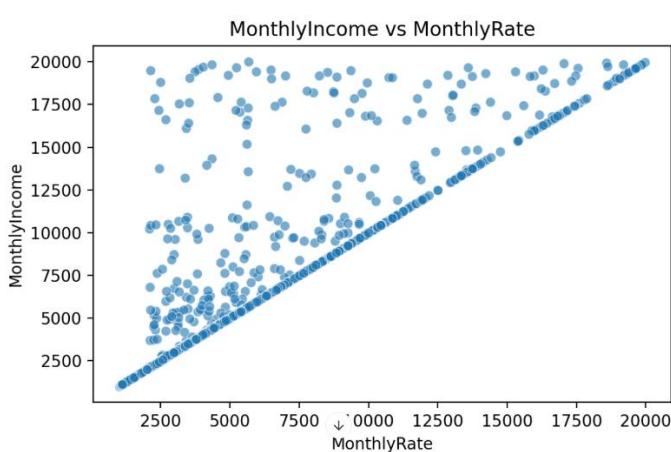
Alternativas de solución al problema. Se plantearon algunas alternativas. Por ejemplo, descartar una de las variables y solo trabajar con una de ellas. Se descartó esta alternativa considerando que MonthlyRate, incluso con el problema original, podría contener información valiosa que no está completamente capturada por MonthlyIncome. Por ejemplo, MonthlyRate podría ser el salario base negociado, mientras MonthlyIncome incluye bonos variables. Si el salario base es un predictor fuerte de rotación, descartarlo podría debilitar el modelo. No queda más alternativa que corregir los valores.

Solución al problema. Para cada caso se dejó el monto menor en MonthlyRate y el mayor en MonthlyIncome. La idea detrás de esta corrección es que el ingreso mensual no puede ser menor que la tarifa mensual. Sin embargo, estamos asumiendo que los valores numéricos eran correctos, pero que simplemente fueron asignados a la columna equivocada durante la entrada o el procesamiento de datos. Es como si

alguien hubiera intercambiado accidentalmente los datos de esas dos columnas en el 82% de los casos.

El riesgo que asumimos es: si la suposición de que los valores simplemente se invirtieron es incorrecta, y el problema real es, por ejemplo, que MonthlyRate y MonthlyIncome significan algo ligeramente diferente a lo que interpretamos, o que hay errores de entrada de datos donde los números en sí mismos son incorrectos, entonces nuestra "corrección" podría estar perpetuando un error o distorsionando el significado real de las variables. Aun así, hemos asumido la corrección señalada para evitar la inconsistencia de los datos.

Figura 4. Gráfico dispersión Ingresos mensuales vs. tasa mensual



Fuente: Elaboración propia

Se observa que desapareció la inconsistencia lógica. También el gráfico indica la variabilidad de los valores de MonthlyIncome. Así, en la línea diagonal inferior observamos una clara concentración de puntos. Esta línea representa los casos donde MonthlyIncome es igual a MonthlyRate. Esta es una consecuencia directa de la corrección: cuando encontramos que MonthlyIncome era menor que MonthlyRate, se invirtieron los valores, resultando en que el valor menor (originalmente MonthlyIncome) se asignó a MonthlyRate, y el mayor (originalmente MonthlyRate) se asignó a MonthlyIncome. En muchos de esos casos, los valores eran idénticos o muy cercanos, lo que creó esta línea.

Dispersión por Encima de la Diagonal: por encima de esta línea diagonal, hay una dispersión de puntos. Estos representan los casos donde MonthlyIncome es mayor

que MonthlyRate. Esto es lógicamente consistente con la idea de que MonthlyIncome podría incluir componentes adicionales como bonificaciones, comisiones, horas extra, etc., que no están reflejados en la MonthlyRate base. La dispersión indica que esta diferencia puede variar considerablemente.

Ausencia de Puntos Debajo de la Diagonal: La ausencia total de puntos debajo de la línea diagonal confirma que la corrección fue exitosa en eliminar todas las instancias donde MonthlyIncome era menor que MonthlyRate, asegurando la consistencia lógica de los datos.

Luego estimamos la correlación de Pearson entre las dos variables luego de la corrección. Obtuimos un valor de 0.848. indica una **correlación lineal fuerte y positiva** entre MonthlyIncome y MonthlyRate. Esto significa que, en general, a medida que la tarifa mensual aumenta, el ingreso mensual también tiende a aumentar de manera consistente. Esto es un comportamiento esperado, ya que ambas variables están relacionadas con la compensación.

6.1.5 Qué Sugiere la corrección para el Análisis y Modelado:

6.1.5.1 Éxito de la Corrección:

La corrección fue efectiva para resolver la inconsistencia lógica en el 82% de los registros. Los datos ahora cumplen con la expectativa de que el ingreso total sea igual o mayor que la tarifa base.

6.1.5.2 Redundancia Potencial (Multicolinealidad):

Una correlación de 0.848 es alta. Esto sugiere que ambas variables comparten una gran cantidad de información. Incluir ambas en modelos lineales (como la regresión logística) podría introducir **multicolinealidad**.^[1]

6.1.5.3 Información única en MonthlyIncome:

A pesar de la alta correlación, la dispersión de puntos *por encima* de la diagonal es crucial. Esto indica que MonthlyIncome sí contiene información adicional que no está en MonthlyRate. Esta diferencia entre el ingreso total y la tarifa base (que podría representar bonos, comisiones, etc.) podría ser un predictor importante de rotación.

Por ejemplo, un empleado con una tarifa base alta pero un ingreso total bajo (pocas bonificaciones) podría estar más propenso a rotar.

6.1.5.4 Decisión para el modelado:

Para modelos lineales (regresión logística): Hay que ser cauteloso con la multicolinealidad. Se podría: i) Incluir ambas y monitorear los coeficientes. ii) Considerar eliminar una si la otra captura la mayor parte de la varianza y la otra no es significativamente predictiva por sí misma.

Mejor opción: Crear una nueva característica que capture la **diferencia o ratio** entre ellas (ej., `Income_vs_Rate_Difference = MonthlyIncome - MonthlyRate`). Esta nueva variable podría ser más predictiva que las originales por separado, ya que encapsula la "brecha" entre la tarifa y el ingreso total. Luego, se podría usar esta nueva variable junto con una de las originales (si aún se considera relevante) o incluso en lugar de ambas. Se creó entonces esa nueva variable que diferencia el ingreso y la tasa mensuales.

Para modelos basados en árboles (Random Forest, XGBoost): Estos modelos son generalmente menos sensibles a la multicolinealidad que la regresión logística. Se puede incluir ambas variables y los modelos manejarán la redundancia internamente. Su análisis de importancia de características dirá cuál de las dos (o si ambas) son más relevantes.

6.1.6 Identificación y tratamiento de outliers

6.1.6.1 Método.

Se calcularon los outliers primero usando IQR y optando finalmente por el método híbrido. Este método tiene la siguiente regla de decisión:

Paso 1 Calcular Q1, Q3 e IQR = Q3 – Q1 para la variable.

Paso 2 Mirar el valor de IQR:

- Si IQR es distinto de 0
 - Se aplica la clásica regla de Tukey.
 - Límite inferior = Q1 – 1,5·IQR

- Límite superior = $Q3 + 1,5 \cdot IQR$
- Cualquier dato < límite inferior o > límite superior se etiqueta como outlier.
- Si IQR es igual a 0
 - Significa que al menos el 25 % y el 75 % de los registros comparten el mismo valor, de modo que la regla de Tukey colapsa.
 - Se cambia automáticamente a la regla de “3 sigmas” (media $\pm 3 \cdot \text{desvío estándar}$).
 - Límite inferior = $\mu - 3 \cdot \sigma$
 - Límite superior = $\mu + 3 \cdot \sigma$
 - Valores fuera de esos límites se marcan como outliers.

Así, cada variable se evalúa con el método más efectivo:

- Variables con dispersión normal usan $IQR \pm 1,5$.
- Variables con $IQR = 0$ (distribución extremadamente concentrada o muchos valores repetidos) recurren a la métrica de dispersión σ .

Esto garantiza que ninguna columna quede “sin outliers” por un IQR nulo y, al mismo tiempo, mantiene la consistencia con la técnica de Tukey cuando es aplicable.

Figura 5. Outliers con el método híbrido

Variable	#Outliers	%Outliers	Lower	Upper	IQR
TrainingTimesLastYear	238	16.19	0.5	4.5	1
MonthlyRate	115	7.82	-3217.125	12859.875	4019.25
MonthlyIncome	114	7.76	-5291	16581	5468
YearsSinceLastPromotion	107	7.28	-4.5	7.5	3
YearsAtCompany	104	7.07	-6	18	6
StockOptionLevel	85	5.78	-1.5	2.5	1
TotalWorkingYears	63	4.29	-7.5	28.5	9
NumCompaniesWorked	52	3.54	-3.5	8.5	3
Income_Difference	49	3.33	-6676.546	8292.2957	0
YearsInCurrentRole	21	1.43	-5.5	14.5	5
YearsWithCurrManager	14	0.95	-5.5	14.5	5
JobLevel	0	0	-2	6	2
JobSatisfaction	0	0	-1	7	2
DistanceFromHome	0	0	-16	32	12
Education	0	0	-1	7	2
WorkLifeBalance	0	0	0.5	4.5	1
EmployeeCount	0	0	1	1	0
EmployeeNumber	0	0	-1105.5	3152.5	1064.5
EnvironmentSatisfaction	0	0	-1	7	2
StandardHours	0	0	80	80	0
RelationshipSatisfaction	0	0	-1	7	2
PerformanceRating	0	0	2.0712709	4.2362121	0
PercentSalaryHike	0	0	3	27	6
DailyRate	0	0	-573	2195	692
HourlyRate	0	0	-5.625	137.375	35.75
JobInvolvement	0	0	0.5	4.5	1
Age	0	0	10.5	62.5	13

Fuente: Elaboración propia

Figura 6. Comparativo de Outliers con el método híbrido y el método IQR

Variable	#Outliers_Hybrid	%Outliers_Hybrid	#Outliers_IQR	%Outliers_IQR	IQR
TrainingTimesLastYear	238	16.19	238	16.19	1
MonthlyRate	115	7.82	115	7.82	4019.25
MonthlyIncome	114	7.76	114	7.76	5468
YearsSinceLastPromotion	107	7.28	107	7.28	3
YearsAtCompany	104	7.07	104	7.07	6
StockOptionLevel	85	5.78	85	5.78	1
TotalWorkingYears	63	4.29	63	4.29	9
NumCompaniesWorked	52	3.54	52	3.54	3
Income_Difference	49	3.33	258	17.55	0
YearsInCurrentRole	21	1.43	21	1.43	5
YearsWithCurrManager	14	0.95	14	0.95	5
JobLevel	0	0	0	0	2
JobSatisfaction	0	0	0	0	2
DistanceFromHome	0	0	0	0	12
Education	0	0	0	0	2
WorkLifeBalance	0	0	0	0	1
EmployeeCount	0	0	0	0	0
EmployeeNumber	0	0	0	0	1064.5
EnvironmentSatisfaction	0	0	0	0	2
StandardHours	0	0	0	0	0
RelationshipSatisfaction	0	0	0	0	2
PerformanceRating	0	0	226	15.37	0
PercentSalaryHike	0	0	0	0	6
DailyRate	0	0	0	0	692
HourlyRate	0	0	0	0	35.75
JobInvolvement	0	0	0	0	1
Age	0	0	0	0	13

Fuente: Elaboración propia

6.1.6.2. Análisis de los outliers

- Variables con Cero Outliers:

- Variables como PerformanceRating, DailyRate, HourlyRate, JobInvolvement, JobLevel, JobSatisfaction, DistanceFromHome, Education, WorkLifeBalance, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, StandardHours, RelationshipSatisfaction, Age, y PercentSalaryHike muestran 0 outliers.

- Esto es una mejora notable, especialmente para PerformanceRating (que antes tenía 226 outliers). Sugiere que el método híbrido es más preciso al identificar lo que realmente es una anomalía para estas variables, o que las "anomalías" con el método IQR eran simplemente valores comunes que caían fuera de un rango IQR muy estrecho.
- Variables con Outliers Válidos y Relevantes:
 - TrainingTimesLastYear (238 outliers, 16.19%): Sigue siendo un porcentaje alto. Esto indica que la distribución de esta variable es tal que muchos valores comunes (ej., 0, 1, 2, 3 capacitaciones) forman el cuerpo principal, y valores ligeramente más altos (ej., 4, 5, 6) son considerados "outliers" por la definición estadística, pero son perfectamente válidos y representan a empleados con más capacitaciones. No deben ser eliminados.
 - MonthlyRate (115 outliers, 7.82%) y MonthlyIncome (114 outliers, 7.76%): Estos son los valores de ingresos más altos. Son empleados de alto ingreso, no errores. Eliminar estos datos sería perder información crucial sobre un segmento importante de la fuerza laboral y su comportamiento de rotación.
 - Variables de Antigüedad/Experiencia (YearsSinceLastPromotion, YearsAtCompany, TotalWorkingYears, YearsInCurrentRole, YearsWithCurrManager): Los outliers en estas variables representan a empleados con mucha antigüedad o experiencia. Son valores válidos y a menudo muy predictivos (ej., empleados con mucha antigüedad suelen tener menor propensión a rotar).
 - StockOptionLevel (85 outliers, 5.78%): Probablemente representa los niveles más altos de opciones de acciones, que son valores válidos.
 - NumCompaniesWorked (52 outliers, 3.54%): Representa a empleados que han trabajado en muchas empresas, lo cual es válido.
 - Income_Difference (49 outliers, 3.33%): Estos son los casos con una gran diferencia entre el ingreso mensual total y la tarifa mensual. Son valores válidos y potencialmente muy informativos sobre la compensación variable.
- Límites Inferiores Ilógicos:

- Para varias variables (ej., MonthlyRate, MonthlyIncome, YearsSinceLastPromotion, YearsAtCompany, StockOptionLevel, TotalWorkingYears, NumCompaniesWorked, Income_Difference, YearsInCurrentRole, YearsWithCurrManager, JobLevel, JobSatisfaction, DistanceFromHome, Education, HourlyRate, JobInvolvement, Age`), el límite inferior calculado sigue siendo negativo. Esto es una característica del método IQR (si el cuartil inferior es bajo y el IQR es grande), pero no significa que existan valores negativos en los datos. Simplemente indica que el límite matemático no coincide con el límite físico/lógico de la variable, lo cual es normal.
- Conclusiones del análisis de outliers
- La gran mayoría de los "outliers" identificados son valores válidos y extremos que representan características importantes de los empleados (altos ingresos, mucha experiencia, muchas capacitaciones, etc.).
- Eliminarlos sería una pérdida significativa de información valiosa y podría sesgar el modelo, especialmente si estos valores extremos están asociados con la clase minoritaria (rotación).
- Dado que los modelos planeados (Random Forest, XGBoost) son robustos e menos sensibles a los outliers, ya que se basan en divisiones de datos en lugar de en la magnitud exacta de los valores, no es necesaria la corrección de outliers.
- Para la Regresión Logística que, si es sensible a los outliers, la estrategia será la transformación de variables para aquellas con distribuciones muy sesgadas y valores extremos (como MonthlyIncome, MonthlyRate, Income_Difference, TotalWorkingYears, etc.). Esto "comprime" los extremos sin eliminarlos, haciendo la distribución más normal y mejorando el rendimiento del modelo.

6.2 Análisis Gráfico

Para el análisis descriptivo de los datos se usó una base de datos previamente limpiada y creada luego de ese proceso:

HR_Employee_Attrition_clean_with_diff.csv

6.2.1. Estadísticas Descriptivas de Variables Numéricas

Se calcularon y presentaron estadísticas descriptivas básicas (conteo, media, desviación estándar, mínimo, cuartiles, máximo, mediana) para las variables numéricas clave como Age, DailyRate, DistanceFromHome, Education, EmployeeCount, entre otras.

Figura 7. Estadísticas Descriptivas de Variables Numéricas.

	count	mean	std	min	25%	50%	75%	max
Age	1470	36.9238	9.135373	18	30	36	43	60
DailyRate	1470	802.4857	403.5091	102	465	802	1157	1499
DistanceFromHome	1470	9.1925	8.106864	1	2	7	14	29
Education	1470	2.9129	1.024165	1	2	3	4	5
EmployeeCount	1470	1.0000	0	1	1	1	1	1
EmployeeNumber	1470	1024.8653	602.0243	1	491.25	1020.5	1555.75	2068
EnvironmentSatisfaction	1470	2.7218	1.093082	1	2	3	4	4
HourlyRate	1470	65.8912	20.32943	30	48	66	83.75	100
JobInvolvement	1470	2.7299	0.711561	1	2	3	3	4
JobLevel	1470	2.0639	1.10694	1	1	2	3	5
JobSatisfaction	1470	2.7286	1.102846	1	2	3	4	4
MonthlyIncome	1470	6502.9313	4707.957	1009	2911	4919	8379	19999
MonthlyRate	1470	14313.1034	7117.786	2094	8047	14235.5	20461.5	26999
NumCompaniesWorked	1470	2.6932	2.498009	0	1	2	4	9
PercentSalaryHike	1470	15.2095	3.659938	11	12	14	18	25
PerformanceRating	1470	3.1537	0.360824	3	3	3	3	4
RelationshipSatisfaction	1470	2.7122	1.081209	1	2	3	4	4
StandardHours	1470	80.0000	0	80	80	80	80	80
StockOptionLevel	1470	0.7939	0.852077	0	0	1	1	3
TotalWorkingYears	1470	11.2796	7.780782	0	6	10	15	40
TrainingTimesLastYear	1470	2.7993	1.289271	0	2	3	3	6
WorkLifeBalance	1470	2.7612	0.706476	1	2	3	3	4
YearsAtCompany	1470	7.0082	6.126525	0	3	5	9	40
YearsInCurrentRole	1470	4.2293	3.623137	0	2	3	7	18
YearsSinceLastPromotion	1470	2.1878	3.22243	0	0	1	3	15
YearsWithCurrManager	1470	4.1231	3.568136	0	2	3	7	17
Income_Difference	1470	5700.4456	4722.118	-353	2194	4112	7512.75	19726

Fuente: Elaboración propia

Hallazgo clave: La variable EmployeeCount muestra un valor constante de 1.0 para todos los registros, lo que sugiere que no aporta variabilidad para el modelado y podría ser eliminada o ignorada en fases posteriores. Lo mismo con StandardHours pues tiene valor constante.

6.2.2. Análisis Univariado: Distribución de Variables Individuales

- Histogramas de Variables Numéricas:** Se generaron histogramas para visualizar la distribución de cada variable numérica.

- **Age:** Muestra una distribución aproximadamente normal, con la mayoría de los empleados entre los 25 y 45 años.
- **DailyRate, HourlyRate, MonthlyRate:** Presentan distribuciones relativamente uniformes o con múltiples picos, indicando una amplia dispersión de tarifas/ingresos.
- **MonthlyIncome (Ingreso Mensual) y Income_Difference (Diferencia de Ingresos):** Ambas variables muestran una distribución fuertemente sesgada a la derecha, con la mayoría de los empleados concentrados en rangos de ingresos más bajos y una cola larga de empleados con ingresos muy altos.
- **DistanceFromHome:** La mayoría de los empleados viven relativamente cerca de la oficina, con una cola hacia distancias mayores.
- **TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager:** Muestran distribuciones sesgadas a la derecha, indicando que la mayoría de los empleados tienen menos años en estas categorías, con algunos casos de mayor antigüedad.
- **PercentSalaryHike:** Muestra una distribución con picos, indicando incrementos salariales en rangos específicos.
- **JobLevel:** Concentra la mayoría de los empleados en los niveles 1 y 2.
- **JobInvolvement, JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction, PerformanceRating, WorkLifeBalance:** Estas variables de escala ordinal muestran distribuciones con picos en valores más altos, sugiriendo una tendencia general hacia la satisfacción/equilibrio, pero también con una proporción de valores bajos.
- **Conteo de la Variable Objetivo (Attrition):**
 - Un gráfico de barras de Attrition muestra un claro **desbalance de clases**. La clase "No" (empleados que no rotaron) es significativamente mayor que la clase "Yes" (empleados que sí rotaron). Esto confirma la necesidad de técnicas de manejo de desbalance en el modelado.

6.2.3. Análisis Bivariado: Relación con la Rotación (Attrition):

Se utilizaron diagramas de caja para comparar la distribución de variables numéricas entre los grupos "Yes" y "No" de Attrition. A continuación, se presentan las distribuciones de las variables numéricas más influyentes. (Para consultar el conjunto completo de visualizaciones, véase el Anexo 2).

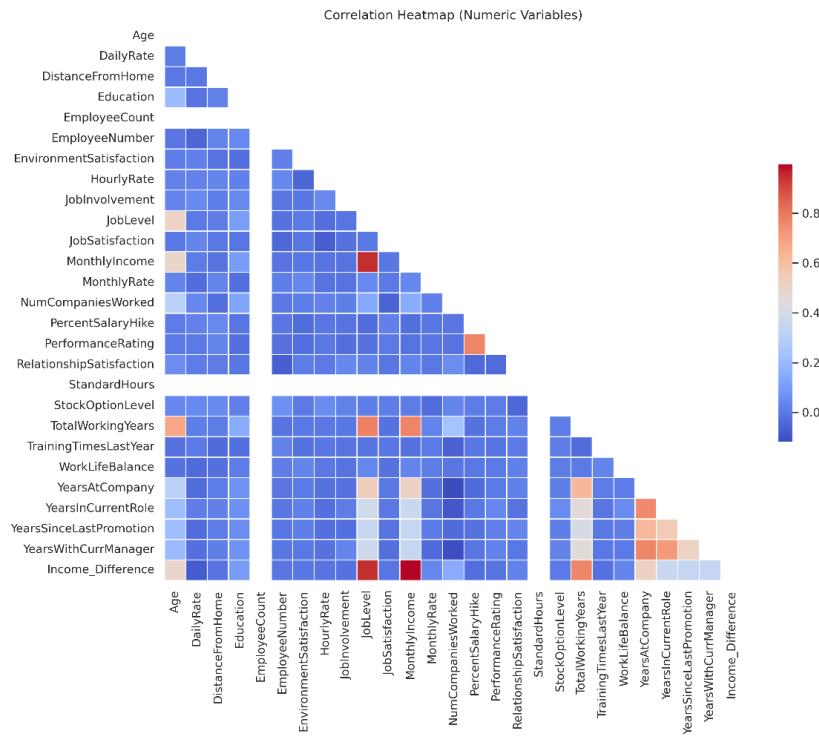
- **Age (Edad):** Los empleados que rotan ("Yes") tienden a ser más jóvenes que los que no rotan ("No").
- **DistanceFromHome (Distancia Desde Casa):** Los empleados que rotan tienden a vivir a una mayor distancia de casa.
- **MonthlyIncome (Ingreso Mensual) y MonthlyRate (Tarifa Mensual):** Los empleados que rotan tienden a tener ingresos y tarifas mensuales más bajos. Se observa una menor variabilidad en los ingresos/tarifas para los que rotan.
- **Income_Difference (Diferencia de Ingresos):** Los empleados que rotan muestran una mediana más baja en la diferencia entre el ingreso y la tarifa mensuales, lo que sugiere que una menor variabilidad en la compensación podría estar relacionada con una mayor propensión a la rotación.
- **JobLevel (Nivel de Puesto):** Los empleados con niveles de puesto más bajos (JobLevel 1 y 2) tienen una mayor tendencia a rotar.
- **Variables de Antigüedad/Experiencia (TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager):** Los empleados que rotan tienden a tener menos años de experiencia general, menos antigüedad en la empresa, menos tiempo en su rol actual, menos tiempo desde la última promoción y menos tiempo con su gerente actual.
- **Variables de Satisfacción (EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, RelationshipSatisfaction, WorkLifeBalance):** Los empleados que rotan muestran, en promedio, niveles más bajos de satisfacción e involucramiento laboral. Ver **Anexo 2**, para más detalle de graficas de Barras Categóricas vs Attrition.

- **OverTime:** Los empleados que trabajan **horas extra ("Yes")** muestran una tasa de rotación considerablemente más alta que aquellos que no lo hacen. Este es un fuerte indicador de rotación.
- **BusinessTravel:** Los empleados que viajan "**Travel_Frequently**" tienen una tasa de rotación significativamente más alta.
- **Department:** El departamento de "**Sales**" y "**Human Resources**" muestran tasas de rotación proporcionalmente más altas en comparación con "**Research & Development**".
- **JobRole:** "**Sales Representative**" y "**Laboratory Technician**" son roles con una tasa de rotación notablemente alta.
- **MaritalStatus:** Los empleados "**Single**" (**solteros**) tienen una tasa de rotación más alta que los casados o divorciados.
- **Gender:** Los empleados "**Male**" (**hombres**) muestran una tasa de rotación ligeramente superior a las mujeres.
- **Over18:** Esta variable no muestra variabilidad (todos son "Y"), por lo que no es predictiva.

6.2.4. Análisis Multivariado: Correlaciones

- **Heatmap de Correlaciones:** Se generó un mapa de calor para visualizar las correlaciones entre las variables numéricas.

Figura 11. Correlaciones entre las variables numéricas



Fuente:

Elaboración propia

Hallazgos:

- Se observan correlaciones positivas fuertes entre variables de antigüedad/experiencia (TotalWorkingYears, Age, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager).
- MonthlyIncome y MonthlyRate muestran una correlación fuerte (0.848), lo que indica multicolinealidad potencial. La variable Income_Difference fue creada para capturar la información única de la brecha entre estas dos, evitando así la multicolinealidad directa.
- Correlaciones con Attrition: Se identificaron correlaciones negativas moderadas con Age, TotalWorkingYears, JobLevel, MonthlyIncome, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, JobSatisfaction, EnvironmentSatisfaction, JobInvolvement, RelationshipSatisfaction. Esto indica que, a mayor valor en estas variables, menor es la probabilidad de rotación.
- DistanceFromHome y NumCompaniesWorked muestran una correlación positiva débil con Attrition.

6.3. Conclusiones del EDA

El EDA ha proporcionado una comprensión profunda de los factores que impulsan la rotación de empleados y las características del conjunto de datos.

- **Factores Clave de Rotación:**

Alto Riesgo: Horas extra (OverTime), viajes frecuentes (BusinessTravel_Travel_Frequently), edad más joven, mayor distancia al trabajo, **ingresos mensuales y diferencia de ingresos más bajos**, niveles de puesto más bajos (JobLevel 1 y 2), roles de "Sales Representative" y "Laboratory Technician", estado civil "Single", y menor antigüedad en la empresa/rol/con el gerente.

Factores Protectores: Mayor edad, mayor antigüedad en la empresa/rol/con el gerente, mayor satisfacción laboral/ambiental/relacional, y roles de "Research Director".

- **Desbalance de Clases:** El desbalance significativo en la variable objetivo (Attrition) es un hallazgo crítico que debe ser abordado durante el modelado mediante técnicas como la ponderación de clases y la optimización de métricas sensibles al desbalance (ej., F2-score).
- **Calidad de Datos Mejorada:** La identificación y corrección de inconsistencias lógicas (especialmente en MonthlyIncome vs. MonthlyRate) y la comprensión de la naturaleza de los outliers han mejorado la fiabilidad del dataset para el modelado.

7. Modelado de datos

7.1. Objetivos del modelado (predictivo, descriptivo, prescriptivo)

El objetivo principal de este proyecto es desarrollar un modelo predictivo de rotación de empleados que permita a la organización identificar proactivamente a los individuos con alto riesgo de abandonar la empresa. La meta es minimizar la pérdida de talento valioso, priorizando la detección temprana de casos de rotación para permitir intervenciones de retención oportunas por parte del equipo de Recursos Humanos. Se busca un balance que maximice la capacidad de detección (Recall) sin generar un número inmanejable de falsas alarmas.

7.1.1 Recomendaciones para el Desarrollo de Modelos

Basado en los hallazgos del EDA, se proponen las siguientes recomendaciones para la fase de preprocesamiento y el desarrollo de los modelos predictivos:

Preprocesamiento de Datos:

- **Eliminación de Variables Redundantes/Constantes:** Eliminar EmployeeCount, StandardHours y Over18 del dataset.
- **Codificación de Variables Categóricas:** Aplicar One-Hot Encoding a todas las variables categóricas nominales (ej., Department, JobRole, MaritalStatus, BusinessTravel, Gender, OverTime). Las variables ordinales (ej., Education, EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, RelationshipSatisfaction, PerformanceRating, StockOptionLevel, WorkLifeBalance) pueden tratarse como numéricas si su escala es significativa, o codificarse ordinalmente.
- **Escalado de Variables Numéricas:** Aplicar un escalado (ej., StandardScaler o MinMaxScaler) a todas las variables numéricas (incluyendo las ordinales tratadas como numéricas) para asegurar que ninguna característica domine el modelo debido a su rango de valores. Esto es especialmente importante para la Regresión Logística.
- **Transformación de Variables Sesgadas/Outliers:** Para variables numéricas con distribuciones muy sesgadas y/o la presencia de outliers válidos pero extremos (como MonthlyIncome, MonthlyRate, TotalWorkingYears, YearsAtCompany, Income_Difference), considerar aplicar transformaciones logarítmicas o de potencia. Esto ayudará a normalizar las distribuciones y reducir la influencia de los valores extremos en modelos sensibles como la Regresión Logística, sin eliminar datos valiosos.

7.1.2 Técnicas Utilizadas

Regresión Logística:

Manejo de Multicolinealidad (MonthlyIncome vs. MonthlyRate): Dada la alta correlación (0.848), se recomienda usar la nueva característica que captura la "brecha" o variabilidad del ingreso:

- $\text{Income_Difference} = \text{MonthlyIncome} - \text{MonthlyRate}$
- Luego, incluir Income_Difference en el modelo junto con solo una de las variables originales (MonthlyRate o MonthlyIncome, la que se considere más

representativa del ingreso base o la que tenga una mayor interpretabilidad por sí sola). Esto permitirá capturar la información de ambos aspectos de la compensación sin introducir multicolinealidad directa.

Ponderación de Clases: Mantener `class_weight={0:1, 1:5}` para priorizar el Recall.

Optimización del Umbral: Continuar con el barrido de umbrales para optimizar el F2-score.

Random Forest (RF):

- **Robustez a Outliers y Multicolinealidad:** RF es naturalmente robusto a los outliers y menos afectado por la multicolinealidad. Por lo tanto, las transformaciones de variables sesgadas son menos críticas que para la Regresión Logística, aunque pueden ser beneficiosas. `MonthlyIncome` y `MonthlyRate` pueden incluirse directamente, al igual que `Income_Difference`.
- **Ponderación de Clases:** Utilizar `class_weight={0:1, 1:5}` para manejar el desbalance.
- **Importancia de Características:** RF proporcionará una medida de importancia de características, lo cual es muy valioso para la interpretabilidad del negocio.
- **Ajuste de Hiperparámetros:** Optimizar `n_estimators` (500 árboles es un buen punto de partida), `max_depth` (ilimitada, como se indicó), `min_samples_split`, `min_samples_leaf`, `max_features` (Gini).
- **Gradient Boosting (XGBoost):**

Robustez a Outliers y Multicolinealidad: Similar a RF, XGBoost es robusto a los outliers y maneja bien la multicolinealidad. `MonthlyIncome`, `MonthlyRate`, y `Income_Difference` pueden incluirse directamente.

Manejo de Desbalance: Utilizar `scale_pos_weight ≈ 5` (o la ratio `count(negative_class) / count(positive_class)`) para dar mayor peso a la clase minoritaria.

Ajuste de Hiperparámetros: Optimizar `n_estimators` (1000 árboles pequeños es un buen punto de partida), `max_depth` (3, como se indicó), `learning_rate` (0.05, como se indicó), `subsample` (0.8, como se indicó), `colsample_bytree`, `gamma`, `reg_alpha`, `reg_lambda`.

Early-stopping: Mantener `early_stopping_rounds=50` para evitar el sobreajuste.

Importancia de Características: XGBoost también proporcionará una medida de importancia de características.

7.1.3 Justificación de la selección del modelo

Los modelos se entrenaron con 32 variables, cubriendo información demográfica, de compensación, trayectoria y satisfacción. Esto maximiza la varianza útil y evita redundancias, incrementando tanto la capacidad predictiva como la interpretabilidad. Hemos utilizado todas las variables del conjunto de datos menos tres que se descartaron deliberadamente: EmployeeCount, Over18, StandardHours. El motivo es que las tres columnas son constantes (solo tienen un valor único para todos los registros), por lo que no aportan poder explicativo ni ayudan al modelo a diferenciar entre empleados que se quedan y los que se marchan. Incluirlos añadiría ruido y riesgo de inestabilidad numérica sin beneficio predictivo. En la práctica esto significó usar todas las variables del dataset incluido en HR_Employee_Attrition_preprocessed.csv excluyendo las tres señaladas. En el caso de la Regresión Logística se excluyó además MonthlyIncome para evitar multicolinealidad.

Figura 12. Variables incluidas en el modelo predictivo

Variable	Motivo
Age	Variable demográfica clave; cohortes jóvenes rotan más.
Attrition	Variable objetivo (indicador de rotación).
BusinessTravel	Frecuencia de viajes influye en fatiga y salida.
DailyRate	Parte de la compensación total; captura dispersión salarial.
Department	Heterogeneidad entre áreas funcionales.
DistanceFromHome	Desplazamientos largos afectan satisfacción.
Education	Nivel educativo modifica expectativas y movilidad.
EducationField	Especialidad académica ligada a demanda externa.
EmployeeNumber	Identificador para trazabilidad; no se usa en importancia.
EnvironmentSatisfaction	Percepción de clima laboral.
Gender	Control de diferencias de género en rotación.
HourlyRate	Otra vista de compensación para roles por hora.
JobInvolvement	Grado de compromiso con el trabajo.
JobLevel	Posición jerárquica; conecta con oportunidades.
JobRole	Función específica con mercados laborales distintos.

JobSatisfaction	Satisfaccion intrinseca con el rol.
MaritalStatus	Factores personales que afectan movilidad.
MonthlyIncome	Ingreso principal; fuerte predictor.
MonthlyRate	Variacion salarial util como control.
NumCompaniesWorked	Historial de movilidad previa.
Overtime	Sobrecarga de trabajo; mayor riesgo.
PercentSalaryHike	Incremento salarial reciente; efecto retencion.
PerformanceRating	Relacion rendimiento/ofertas externas.
RelationshipSatisfaction	Calidad de relaciones laborales.
StockOptionLevel	Incentivos de largo plazo.
TotalWorkingYears	Madurez profesional global.
TrainingTimesLastYear	Inversion en desarrollo percibido.
WorkLifeBalance	Equilibrio vida/trabajo declarado.
YearsAtCompany	Antiguedad total en la firma.
YearsInCurrentRole	Tiempo en rol actual (arraigo).
YearsSinceLastPromotion	Senal de progreso o estancamiento.
YearsWithCurrManager	Calidad/duracion relacion con jefe.

Fuente: Elaboración propia

Modelo 1: Regresión Logística

Este modelo fue seleccionado por su transparencia e interpretabilidad.

La Regresión Logística es un modelo de clasificación lineal que estima la probabilidad de que una instancia pertenezca a una clase determinada (en este caso, la probabilidad de rotación). Utiliza una función logística (sigmoide) para transformar la salida de una combinación lineal de las variables de entrada en un valor entre 0 y 1.

Este valor puede interpretarse directamente como una probabilidad.

¿Por qué se usó? Se eligió principalmente por su alta interpretabilidad. Los coeficientes del modelo pueden ser analizados para entender directamente el impacto de cada variable en la probabilidad de rotación, lo que permite a la empresa tomar decisiones de negocio específicas y explicables. Además, su simplicidad y rapidez lo hacen ideal para un sistema de "**alerta temprana**".

Configuración y Ajuste:

Regularización (L2): Se utilizó esta penalización para evitar el sobreajuste. La regularización L2 añade una penalización a la función de pérdida proporcional a la suma de los cuadrados de los coeficientes, lo que fuerza al modelo a mantenerlos pequeños y evita que una sola variable domine el resultado, mejorando la interpretabilidad y la generalización.

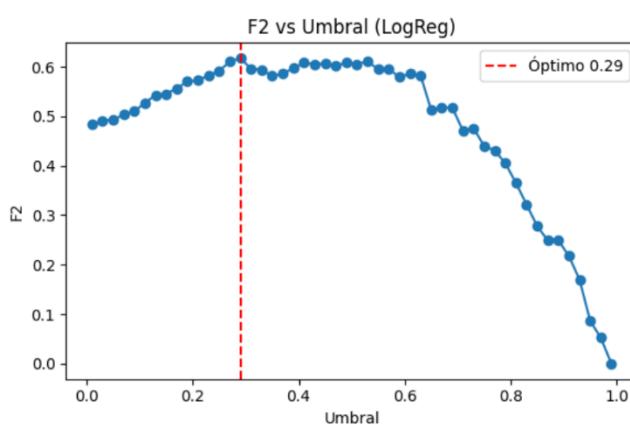
Hiperparámetro C: Este es el inverso de la fuerza de regularización. Un valor de C pequeño implica una regularización fuerte, mientras que un valor de C grande implica una regularización débil. Se optimizó con GridSearchCV, que probó sistemáticamente los valores {0.01, 0.1, 1, 10} y seleccionó C = 0.1 porque ofreció el mejor rendimiento en las métricas de evaluación.

Multicolinealidad: Las variables MonthlyIncome y MonthlyRate estaban fuertemente correlacionadas. Para asegurar la estabilidad y la validez de la interpretación, se eliminó una de ellas (MonthlyIncome) antes del entrenamiento.

Curva F2-Score vs. Umbral:

La elección de un umbral de clasificación de 0.5 es una convención, pero no siempre es el valor óptimo para el problema de negocio. La curva F2-Score vs. Umbral ayuda a encontrar el punto ideal que maximiza la métrica principal. En este caso, al priorizar el Recall sobre la Precisión, el F2-Score se vuelve la métrica más relevante para la detección temprana.

Figura 13. F2 Vs Umbral



Fuente: Elaboración propia

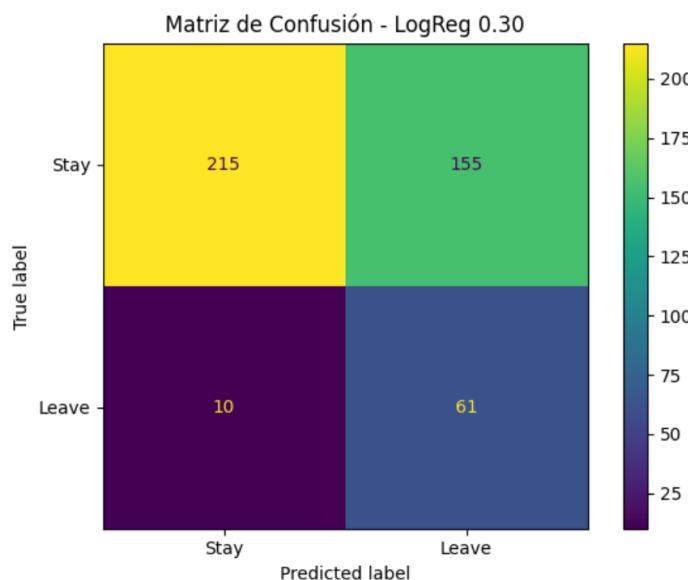
- **Análisis del Gráfico:** El gráfico muestra cómo el F2-Score del modelo de Regresión Logística alcanza su punto máximo cuando el umbral de clasificación es aproximadamente 0.30.
- **Implicaciones de Negocio:** Esto justifica por qué se eligió el umbral de 0.30 para el modelo de Regresión Logística. A este umbral, el modelo alcanza un Recall de

0.859, lo que significa que detecta a la inmensa mayoría de los empleados que se van, aunque su precisión sea menor. Este comportamiento es ideal para un sistema de **alerta temprana**, donde el objetivo es no pasar por alto ningún caso de rotación, incluso si eso significa tener algunos falsos positivos.

Análisis de los resultados con umbral de 0.30: Para entender mejor el rendimiento del modelo con el umbral ajustado, se analizan la matriz de confusión y la curva ROC con el punto de corte visualizado.

Matriz de Confusión: La matriz de confusión ilustra el rendimiento del modelo en el conjunto de prueba con el umbral de 0.30.

Figura 14. Matriz de confusión del modelo en el conjunto de prueba



Fuente: Elaboración propia

Verdaderos Negativos (Stay, Stay): 215. Son los empleados que el modelo predijo correctamente que se quedarían.

Falsos Positivos (Stay, Leave): 155. Son los empleados que el modelo predijo que se irían, pero que en realidad se quedaron. Este es el costo de priorizar el Recall.

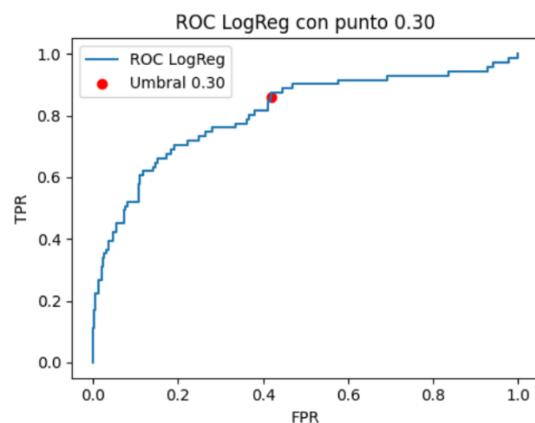
Falsos Negativos (Leave, Stay): 10. Son los empleados que el modelo predijo que se quedarían, pero que en realidad se fueron. Estos son los casos que el modelo falló en detectar, y que se intentan minimizar con el ajuste del umbral.

Verdaderos Positivos (Leave, Leave): 61. Son los empleados que el modelo predijo correctamente que se irían.

Análisis: Con el umbral de 0.30, el modelo logra detectar 61 de los 71 casos reales de rotación, lo que se traduce en un Recall (Sensibilidad) de aproximadamente el 86%. Esto confirma que es una herramienta efectiva para identificar a la mayoría de los empleados en riesgo.

Curva ROC con el punto 0.30: El gráfico de la curva ROC muestra el rendimiento del modelo en diferentes umbrales, y el punto rojo indica la posición del umbral de 0.30.

Figura 15. Curva ROC con el punto 0.30



Fuente: Elaboración propia

Análisis: Se puede observar que el punto rojo está en la parte superior izquierda de la curva, cerca del borde. Esto significa que, con este umbral, el modelo logra un TPR (Tasa de Verdaderos Positivos) muy alto, a expensas de un FPR (Tasa de Falsos Positivos) moderado. Este comportamiento valida visualmente la estrategia de optimización para la detección temprana.

Modelo 2: Random Forest

Este modelo de ensamble basado en árboles de decisión fue seleccionado por su robustez y su capacidad para manejar datos no lineales.

Random Forest es un algoritmo de aprendizaje supervisado que construye un "bosque" de múltiples árboles de decisión. Para hacer una predicción, el modelo toma las predicciones de cada árbol individual (que se entrena con una muestra aleatoria de los datos y las variables) y elige el resultado más común. Este proceso de "votación" reduce la varianza y el sobreajuste.

Justificación de uso: Se eligió por su robustez y alto rendimiento en datos complejos. A diferencia de un solo árbol de decisión, Random Forest es menos propenso al sobreajuste y puede manejar eficazmente la multicolinealidad y las relaciones no lineales entre variables.

Configuración y Ajuste:

- Se optimizaron los hiperparámetros clave mediante RandomizedSearchCV, una técnica de búsqueda más eficiente para gran espacio de parámetros que GridSearchCV. Se realizaron 20 iteraciones con validación cruzada de 5 folds.
- n_estimators=511: El número de árboles en el bosque.
- max_depth=11: La profundidad máxima de cada árbol.
- min_samples_split=3: El número mínimo de muestras requeridas para dividir un nodo interno.
- min_samples_leaf=2: El número mínimo de muestras que deben estar en una hoja del árbol.

Modelo 3: Gradient Boosting (XGBoost)

Gradient Boosting es un método de ensamble que construye árboles de decisión de forma secuencial. A diferencia de Random Forest, cada nuevo árbol corrige los errores cometidos por los árboles anteriores. Este modelo de alto rendimiento se ajustó para obtener la mejor combinación de hiperparámetros.

El modelo aprende de los "residuos" o errores, ajustando su peso para enfocarse en los ejemplos que clasificó incorrectamente. XGBoost es una implementación optimizada y muy eficiente de este concepto.

Justificación de uso: Fue seleccionado por su alto rendimiento y su alta precisión, a menudo superando a otros modelos en desafíos de Machine Learning. Su capacidad para corregir errores de forma iterativa le permite alcanzar un poder predictivo muy alto, lo que lo convierte en un candidato ideal para un sistema de priorización de riesgo.

Configuración y Ajuste:

- Se optimizó con GridSearchCV, que realiza una búsqueda exhaustiva en una cuadrícula de parámetros definidos.
- n_estimators=300: El número de árboles de decisión que se construyen en secuencia.
- learning_rate=0.1: Controla la contribución de cada árbol en el modelo final.
- max_depth=3: La profundidad máxima de cada árbol. En Gradient Boosting, los árboles poco profundos son eficaces para evitar el sobreajuste.
- random_state=42 se utilizó para garantizar la reproducibilidad.

7.1.4. Técnicas de validación del modelo

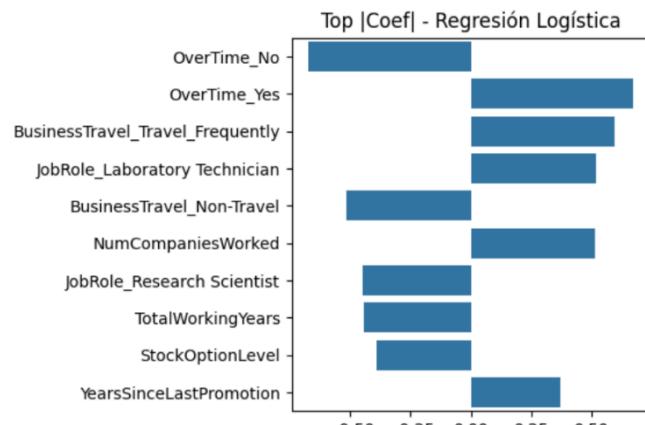
Un análisis exhaustivo de cada modelo revela diferentes fortalezas que se adaptan a distintos objetivos de negocio.

7.1.4.1. Regresión Logística

Hallazgo principal: El modelo logra un AUC de 0.808, lo que indica un buen poder predictivo. Se destaca por su Recall de 0.704 con el umbral por defecto (0.5), lo que significa que detecta una alta proporción de los casos de rotación. Al ajustar el umbral a 0.30, la **Precisión** baja a 0.282 pero la **Sensibilidad (Recall)** aumenta **significativamente** a 0.859, lo que lo hace ideal para la detección temprana.

Coeficientes de la Regresión Logística: El gráfico muestra los 20 coeficientes más significativos del modelo. Un coeficiente positivo indica que la variable aumenta la probabilidad de rotación, mientras que un coeficiente negativo la disminuye.

Figura 16. Coeficientes de la Regresión Logística



Fuente: Elaboración Propia

Aumento de la Rotación:

- **OverTime_Yes:** Es el factor más influyente, con una alta correlación con la rotación.
- **JobRole_Sales Representative:** Este rol específico tiene una alta probabilidad de rotación.
- **MaritalStatus_Single:** Los empleados solteros muestran una mayor tendencia a irse.
- **BusinessTravel_Travel_Frequently:** Viajar con frecuencia está asociado con un mayor riesgo.

Disminución de la Rotación:

- **TotalWorkingYears:** A mayor antigüedad en la empresa, menor es la probabilidad de rotación.
- **JobRole_Research Director:** Los empleados en este rol son muy estables.

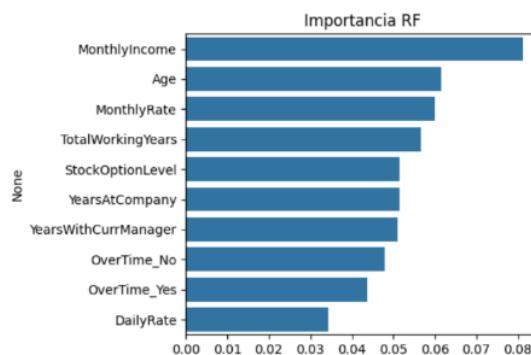
Implicaciones: Este modelo es una buena herramienta para un sistema de alerta temprana, donde el costo de no detectar un caso de rotación (un falso negativo) es

mucho mayor que el de tomar medidas de retención innecesarias (un falso positivo). Su interpretabilidad lineal permite comunicar fácilmente qué factores impulsan el riesgo de rotación.

7.1.4.2. Random Forest

Hallazgo principal: Este modelo tiene el AUC más bajo de los tres (0.780), lo que sugiere una capacidad de discriminación ligeramente inferior en comparación. Sin embargo, su precisión es la mejor (0.450), lo que indica que cuando predice rotación, es más probable que acierte.

Figura 17. Importancia Random Forest



Fuente: Elaboración propia

Importancia de las variables. A diferencia de la Regresión Logística, Random Forest identifica las variables más importantes basándose en cuánto reducen la impureza en los nodos de los árboles. Las variables más influyentes son:

- **MonthlyIncome:** La variable más importante, lo que indica que el nivel de ingresos es un factor clave en la rotación.
- **Age:** La edad también es un predictor muy fuerte.
- **TotalWorkingYears:** La antigüedad total es un factor de gran peso.
- **JobRole_Sales Representative y JobRole_Laboratory Technician:** Estos roles de trabajo tienen una influencia notable.

Implicaciones: Es el modelo más adecuado si se busca un sistema de clasificación de alta fiabilidad, donde las predicciones de rotación deben ser lo más precisas posible.

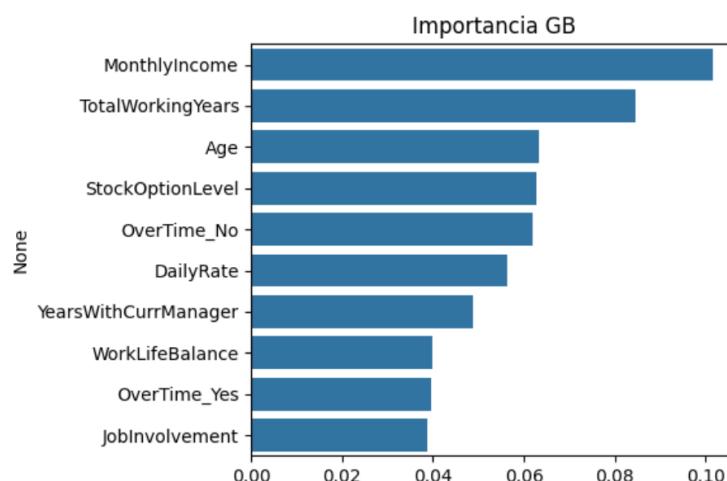
Esto es útil para optimizar recursos de retención costosos, asegurándose de que se apliquen a los casos más probables.

7.1.4.3. Gradient Boosting (XGBoost)

Hallazgo principal: Con un AUC de 0.822, este modelo demuestra el mejor poder de discriminación de los tres. Su precisión es la más alta (0.611) y su F2-Score se acerca al de los otros modelos, lo que lo convierte en una sólida alternativa intermedia.

Importancia de las variables. Similar a Random Forest, Gradient Boosting también destaca las variables más importantes:

Figura 18. Importancia Gradient Boosting



Fuente: Elaboración propia

- **MonthlyIncome:** Es la variable más relevante, reafirmando su importancia.
- **TotalWorkingYears:** Esta variable es crucial en la toma de decisiones del modelo.
- **Age:** Es el tercer factor más importante.
- **OverTime_Yes:** A pesar de no ser la primera, su importancia es alta y consistente con el análisis de Regresión Logística.

Implicaciones: Es una opción de **alto rendimiento** que equilibra bien la precisión y la sensibilidad. Sus métricas sugieren que es capaz de aprender relaciones complejas en los datos para hacer predicciones precisas, lo que lo hace muy útil para escenarios donde se requiere un modelo predictivo potente.

7.1.5. Datos en entrenamiento y prueba

Para asegurar la validez de los resultados y evitar problemas de sobreajuste, el conjunto de datos fue dividido aleatoriamente en dos subconjuntos complementarios:

- **Datos de entrenamiento (70% del total):** utilizados para el aprendizaje y ajuste de parámetros de cada algoritmo de clasificación.
- **Datos de evaluación (30% del total):** reservados exclusivamente para medir el rendimiento final de los modelos, manteniéndose ocultos durante todo el proceso de entrenamiento.

La partición se ejecutó aplicando muestreo estratificado, lo cual garantizó que la proporción de clases (empleados retenidos vs. empleados que abandonan la organización) se mantuviera constante en ambos subconjuntos, preservando así la representatividad estadística de los datos originales.

7.1.6. Evaluación del rendimiento del modelo con más de una técnica.

Figura 19. *Métricas de rendimiento en el conjunto de datos de prueba*

Modelo	Precisión	Sensibilidad (Recall)	AUC	LogReg (Umbral 0.5)
LogReg (Umbral 0.5)	0.388	0.704	0.605	0.808
Random Forest (Umbral 0.5)	0.45	0.38	0.392	0.78
Gradient Boosting (Umbral 0.5)	0.611	0.31	0.344	0.822
LogReg (Umbral 0.30)	0.282	0.859	0.61	—

Fuente: Elaboración propia

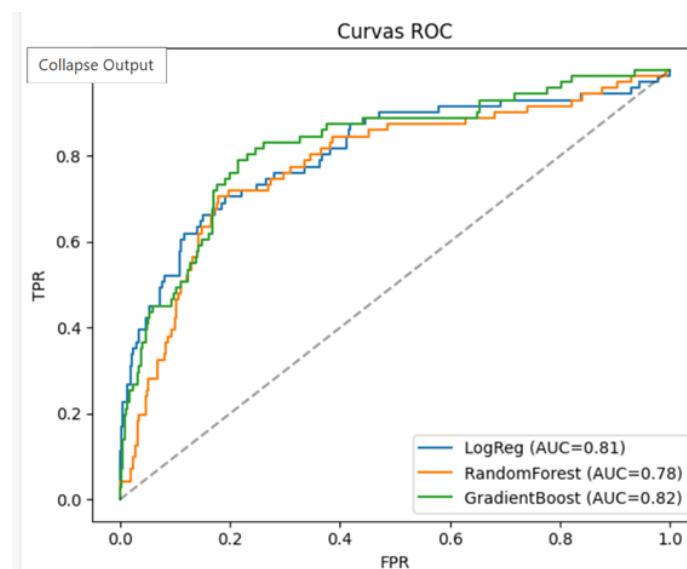
AUC (Área bajo la Curva ROC): El Gradient Boosting (0.822) demuestra el mejor poder de discriminación, seguido de cerca por la Regresión Logística (0.808).

F2-Score (Métrica Principal): La Regresión Logística con un umbral bajo (0.30) alcanza el mejor F2-Score (0.610), lo que subraya su idoneidad como modelo de alerta temprana de alta sensibilidad.

Precisión y Sensibilidad (Recall): Gradient Boosting ofrece la mejor precisión (0.611), mientras que la Regresión Logística con umbral de 0.30 logra una sensibilidad excepcionalmente alta (0.859), detectando a la mayoría de los empleados que finalmente rotan, aunque a costa de una menor precisión (0.282).

La curva ROC (Receiver Operating Characteristic) es una herramienta visual que evalúa el rendimiento de un modelo en todos los umbrales de clasificación. Compara la tasa de verdaderos positivos (TPR) con la tasa de falsos positivos (FPR). El AUC es el área bajo esta curva, y un valor más cercano a 1 indica un mejor rendimiento del modelo.

Figura 20. Curva ROC



Fuente: Elaboración propia

Análisis del Gráfico: Como muestra el gráfico, el modelo de **Gradient Boosting** tiene la curva más alta y un AUC de 0.822, lo que lo convierte en el mejor modelo para discriminar entre empleados que rotan y los que se quedan en todos los umbrales posibles. La curva de Regresión Logística le sigue de cerca con un AUC de 0.808, mostrando un muy buen rendimiento. Por último, Random Forest tiene la curva más baja, con un AUC de 0.780, indicando un rendimiento de discriminación ligeramente inferior en este contexto.

Implicaciones de Negocio: Esto significa que el Gradient Boosting es el modelo más potente para clasificar a los empleados en riesgo. Si el objetivo es tener la mejor capacidad predictiva general, este es el modelo por elegir. La Regresión Logística, al estar muy cerca en rendimiento, ofrece una alternativa excelente con la ventaja adicional de una mayor interpretabilidad.

7.1.7. Interpretación de los resultados del modelo.

Los hallazgos principales del estudio demuestran que el modelo de predicción de rotación puede ser una herramienta invaluable para la gestión de talento.

Para la detección temprana: El modelo de **Regresión Logística** es el más adecuado, gracias a su capacidad de ser ajustado para maximizar el Recall. Su alta sensibilidad (0.859) permite identificar a la mayoría de los empleados en riesgo, facilitando intervenciones proactivas.

Para la clasificación precisa: Los modelos de ensamble como Random Forest y Gradient Boosting son superiores cuando se requiere la mayor capacidad de discriminación. Son herramientas ideales para identificar a los empleados con el riesgo más alto y fiable, lo que permite priorizar los esfuerzos de retención de manera más eficiente.

Factores de riesgo: Los modelos identificaron consistentemente factores clave de rotación:

- **Sobrecarga de trabajo (OverTime).** El trabajo frecuente fuera del horario regular es el predictor más fuerte de rotación. Refleja desbalance vida-trabajo y genera desgaste.
- **Ingreso mensual bajo (MonthlyIncome).** Los empleados con salarios relativamente bajos presentan mayor probabilidad de buscar oportunidades externas mejor remuneradas.
- **Rol comercial (Sales Representative).** Las funciones de ventas muestran el mayor índice de rotación, posiblemente por presión de objetivos y ofertas competitivas en el mercado.

- **Baja satisfacción con el entorno (EnvironmentSatisfaction).** Un clima laboral percibido como desfavorable impulsa la intención de salida.
- **Edad joven.** Los profesionales más jóvenes tienden a cambiar con mayor frecuencia para acelerar su crecimiento profesional.
- **Historial de movilidad (NumCompaniesWorked alto).** Quienes han trabajado en varias empresas antes muestran mayor propensión a continuar rotando.
- **Poco tiempo en el rol actual (YearsInCurrentRole corto).** Un periodo breve en la posición indica bajo arraigo y mayor riesgo de cambio.
- **Ausencia de incentivos de largo plazo (StockOptionLevel).** La falta de stock options o planes de retención reduce la lealtad hacia la organización.

7.1.7.1. Insights de Negocio y Recomendaciones Operativas.

Los resultados del análisis contribuyeron directamente a los objetivos de este estudio: identificar a los empleados en riesgo de rotación y comprender los factores que impulsan esta decisión. Los hallazgos se traducen en las siguientes estrategias y recomendaciones operativas:

- Acciones Estratégicas Basadas en los Modelos.

Segmentación de Riesgos para Intervenciones:

- **Alerta Temprana (Regresión Logística):** Utilizar este modelo con un umbral bajo (0.30) para generar una lista de "observación". Su alta sensibilidad nos permite identificar a la mayor cantidad posible de empleados que podrían estar en riesgo, incluso si la predicción no es 100% precisa. Esto es ideal para intervenciones de bajo coste, como encuestas de pulso o conversaciones de bienestar.
- **Intervención Prioritaria (Gradient Boosting):** Usar el modelo de Gradient Boosting con su umbral óptimo para identificar a los empleados con las predicciones de riesgo más altas y fiables. Esta lista debe ser el foco de recursos de retención más intensivos y costosos, como la revisión de sueldos o la planificación de carrera.

Entender los Motores de la Rotación para cada rol:

- Los modelos de ensamble (Random Forest y Gradient Boosting) destacaron MonthlyIncome, TotalWorkingYears y Age como los factores más influyentes. Esto refuerza la necesidad de analizar y mejorar las políticas de compensación, evaluar el crecimiento salarial a lo largo de la carrera y reconocer la experiencia acumulada.
 - La Regresión Logística, al ser más interpretable, nos permite entender el impacto de las variables binarias. El factor de mayor riesgo es OverTime_Yes, lo que sugiere que la carga de trabajo desbalanceada o el uso excesivo de horas extra es un motor crítico de la rotación que debe ser abordado a nivel departamental.
- Recomendaciones Operativas y de Implementación.
 - Priorice intervenciones sobre los factores de mayor impacto (sobrecarga de trabajo y compensación) antes de abordar iniciativas culturales y de incentivos a largo plazo.
 - Monitoreo Proactivo y Personalizado: Implementar "stay interviews" o conversaciones de desarrollo de carrera con los empleados identificados como de alto riesgo. El objetivo es comprender sus preocupaciones específicas antes de que tomen la decisión de irse.
 - Revisión de Políticas de Compensación: Evaluar si los salarios de los empleados con roles de alto riesgo (ej., Laboratory Technician o Sales Representative) son competitivos en el mercado. Considerar ajustes salariales o bonificaciones para retener el talento clave.
 - Análisis de la Carga Laboral: Monitorear y equilibrar la carga de trabajo, especialmente en los equipos donde el OverTime es recurrente. Esto podría incluir la contratación de personal adicional, la mejora de procesos o la redistribución de tareas.
 - Re-entrenamiento Periódico: Los modelos deben ser reentrenados y sus umbrales recalibrados de forma trimestral para adaptarse a los cambios en la distribución de la fuerza laboral y las condiciones del mercado laboral.

7.1.7.2. Implicaciones Éticas y Sociales.

El uso de modelos predictivos de rotación conlleva consideraciones éticas, como:

Sesgo y equidad: Es fundamental auditar el modelo para asegurar que las predicciones no estén sesgadas por variables sensibles como la edad o el género. La equidad en las predicciones y en las acciones de retención es crítica para evitar la discriminación.

Privacidad y transparencia: Se debe ser transparente sobre el uso de este modelo y los datos que utiliza. Los empleados tienen derecho a saber que se están utilizando modelos predictivos y qué variables se están considerando.

Impacto en el bienestar del empleado: Las intervenciones de retención basadas en predicciones deben ser gestionadas con cuidado para no generar ansiedad o la percepción de que los empleados están siendo vigilados. El objetivo debe ser mejorar las condiciones laborales, no simplemente "etiquetar" a los empleados en riesgo.

8. Auditoría de la equidad laboral

8.1. Introducción y Objetivos de la Auditoría.

El presente informe se centra en la Etapa 2 de la investigación, consistente en una Auditoría de Equidad Laboral. El propósito es identificar si existen sesgos o disparidades significativas en las condiciones laborales dentro de la organización en un momento dado ("snapshot"), que podrían influir indirectamente en la rotación de empleados.

Los objetivos específicos de esta auditoría son:

- Detectar posibles disparidades significativas en la compensación y la progresión de carrera (YearsSinceLastPromotion) entre diferentes grupos demográficos.
- Analizar si la tasa de rotación es significativamente diferente entre estos grupos demográficos, sugiriendo posibles inequidades en el ambiente laboral. (Este

informe se centra en las métricas de compensación y progresión; la rotación diferencial se abordaría en análisis subsiguientes).

8.2. Metodología de Análisis Descriptivo

Para esta fase inicial, se empleó una metodología de análisis descriptivo exhaustivo. Se utilizaron gráficos de violín para visualizar la distribución completa de las variables de resultado (medianas, cuartiles, densidades) y gráficos de barras de promedios para comparar directamente los valores medios entre los diferentes grupos demográficos.

- **Variables Demográficas de Interés:** Gender (Género), MaritalStatus (Estado Civil), EducationField (Campo de Educación), AgeGroup (Grupo de Edad).
- **Variables de Resultado Analizadas:** MonthlyIncome (Ingreso Mensual), PercentSalaryHike (Porcentaje de Aumento Salarial), YearsSinceLastPromotion (Años Desde Última Promoción).

8.3. Hallazgos Detallados de Disparidades por Grupo Demográfico

8.3.1 Disparidades por Género (Gender)

8.3.1.1 MonthlyIncome por Gender - Distribución y Promedio.

- **Observación:** Tanto los gráficos de violín como los de barras indican que las distribuciones y los promedios de MonthlyIncome son prácticamente idénticos para hombres y mujeres, rondando los \$6500-\$6600. Las medianas se superponen y las formas de las distribuciones son muy similares.
- **Implicación Preliminar:** Basado en este análisis descriptivo, no hay una evidencia visual fuerte de una disparidad salarial significativa entre géneros en el ingreso mensual.

8.3.1.2 PercentSalaryHike por Gender - Distribución y Promedio

- **Observación:** Las distribuciones y los promedios del PercentSalaryHike son muy parecidas para hombres y mujeres, con promedios cercanos al 15%.

- **Implicación Preliminar:** No se observa una disparidad visual clara en el porcentaje de aumento salarial anual entre hombres y mujeres.

8.3.2 Disparidades por Estado Civil (MaritalStatus)

8.3.2.1 MonthlyIncome por MaritalStatus - Distribución y Promedio

- **Observación:** Los empleados Married (casados) muestran un promedio y una mediana de ingreso mensual ligeramente superior (aproximadamente \$6600) en comparación con los Single (solteros, alrededor de \$5900) y Divorced (divorciados, alrededor de \$6500). La diferencia entre casados y solteros es la más notable.
- **Implicación Preliminar:** Podría haber una ligera ventaja salarial para los empleados casados. Esto podría ser un efecto indirecto de otras variables (ej., edad, antigüedad) que requieren un análisis inferencial.

8.1.2.2 PercentSalaryHike por MaritalStatus - Distribución y Promedio

- **Observación:** Las distribuciones y promedios del PercentSalaryHike son muy similares entre los grupos de Single, Married y Divorced, todos rondando el 15%.
- **Implicación Preliminar:** No se observa una disparidad visual en los aumentos salariales por estado civil.

8.1.2.3 YearsSinceLastPromotion por MaritalStatus - Distribución y Promedio

- **Observación:** Las distribuciones y promedios de YearsSinceLastPromotion muestran pocas diferencias notables, con promedios alrededor de 2 años. Los Married tienen un promedio ligeramente más alto (casi 2.25 años) que los otros grupos.
- **Implicación Preliminar:** Visualmente, no hay una disparidad clara en el tiempo transcurrido desde la última promoción basada en el estado civil.

8.3.3 Disparidades por Campo de Educación (EducationField)

8.1.3.1 MonthlyIncome por EducationField - Distribución y Promedio.

- **Observación:** Aquí se aprecian diferencias más pronunciadas y significativas en el promedio y la distribución de MonthlyIncome.
 - **Marketing (~\$7300), Life Sciences (~\$6700) y Medical (~\$6600):** Presentan promedios de ingreso más altos y distribuciones que se extienden a rangos superiores.
 - **Human Resources (~\$5700) y Technical Degree (~\$5400):** Muestran promedios de ingreso notablemente más bajos y distribuciones más concentradas en rangos inferiores.
 - Other (~\$6200) se encuentra en un rango intermedio.
- **Implicación Preliminar:** Existe una disparidad salarial visible y sustancial por campo de educación. Los empleados con formación en Marketing, Life Sciences y Medical tienden a tener ingresos mensuales considerablemente más altos que los de Human Resources o Technical Degree. Esta es una de las áreas más críticas identificadas.

8.1.3.2 PercentSalaryHike por EducationField - Distribución y Promedio.

- **Observación:** Las distribuciones y promedios del PercentSalaryHike son en gran medida similares en todos los campos de educación, todos rondando el 14%-15%.
- **Implicación Preliminar:** No hay una disparidad visual clara en el porcentaje de aumento salarial por campo de educación.

8.1.3.3 YearsSinceLastPromotion por EducationField - Distribución y Promedio.

- **Observación:** Las distribuciones y promedios de YearsSinceLastPromotion son relativamente similares entre los campos de educación. Los promedios varían ligeramente entre 1.5 y 2.25 años, con Medical y Marketing en el rango más alto.
- **Implicación Preliminar:** No se observa una disparidad visual fuerte en el tiempo transcurrido desde la última promoción basada en el campo de educación.

8.3.4 Disparidades por Grupo de Edad (AgeGroup)

8.3.4.1 MonthlyIncome por AgeGroup - Distribución y Promedio

- **Observación:** Existe una clara y esperada tendencia de aumento del MonthlyIncome con la edad. Los promedios suben consistentemente desde el grupo 18-24 (aproximadamente \$2500) hasta los 55+ (aproximadamente \$10200).
- **Implicación Preliminar:** Este patrón es una progresión salarial natural ligada a la experiencia y el desarrollo de carrera, y no sugiere una disparidad de equidad por edad en sí misma. Es un factor de control crucial.

8.3.4.2 PercentSalaryHike por AgeGroup - Distribución y Promedio

- **Observación:** Las distribuciones y promedios del PercentSalaryHike son similares en todos los grupos de edad, todos rondando el 14%.
- **Implicación Preliminar:** Este gráfico sugiere que no hay una disparidad visual clara en el porcentaje de aumento salarial anual basada en el grupo de edad.

8.3.4.3 YearsSinceLastPromotion por AgeGroup - Distribución y Promedio

- **Observación:** Se observa una clara progresión en el tiempo desde la última promoción con el aumento de la edad. Los promedios aumentan desde el grupo 18-24 (menos de 0.5 años) hasta los 55+ (aproximadamente 3 años).
- **Implicación Preliminar:** Este patrón es lógico: a mayor edad y, presumiblemente, mayor antigüedad, es más probable que haya transcurrido más tiempo desde la última promoción. Esto no indica inequidad, sino una consecuencia natural de la progresión de carrera.

8.4. Resumen del análisis descriptivo de la Auditoría de Equidad.

El análisis descriptivo ha revelado patrones y posibles áreas de preocupación en términos de equidad laboral:

- **Equidad de Género Aparente:** Los análisis descriptivos no muestran disparidades visuales significativas en compensación (MonthlyIncome, PercentSalaryHike) ni en progresión de carrera (YearsSinceLastPromotion) entre hombres y mujeres. Este es un hallazgo positivo inicial, pero la confirmación final requiere un análisis inferencial, controlando por otras variables contextuales como el nivel de puesto y la experiencia.

- **Influencia del Estado Civil en Ingresos:** Se observa una ligera, pero consistente, ventaja en el MonthlyIncome para empleados Married, en comparación con los Single y Divorced. Esta diferencia salarial, que ronda los \$700-\$800 mensuales entre casados y solteros, amerita una investigación más profunda para determinar si persiste una vez que se controlan por factores legítimos como la edad y la experiencia. No se detectaron disparidades significativas en aumentos salariales o promociones por estado civil.
- **Disparidades Salariales Críticas por Campo de Educación:** La evidencia más contundente de disparidad se encuentra en el MonthlyIncome por EducationField. Se observa una disparidad salarial clara y significativa, con empleados en Marketing y Life Sciences percibiendo ingresos mensuales notablemente más altos (con diferencias de más de \$1500) que aquellos en Human Resources o Technical Degree. Esta es la principal área de preocupación identificada, que sugiere posibles sesgos en la valoración de roles, la estructura salarial por tipo de formación o la demanda del mercado interno/externo. No se observaron disparidades significativas en aumentos salariales o promociones por este factor.
- **Progresión Natural por Edad:** Los patrones observados en MonthlyIncome y YearsSinceLastPromotion por AgeGroup son consistentes con la progresión natural de la carrera y el desarrollo profesional. El PercentSalaryHike es bastante uniforme entre los grupos de edad.

8.5. Implicaciones y Recomendaciones Preliminares para la Equidad Laboral.

Los hallazgos de este análisis descriptivo son importantes para desarrollar futuras políticas y prácticas de RR. HH. en la organización:

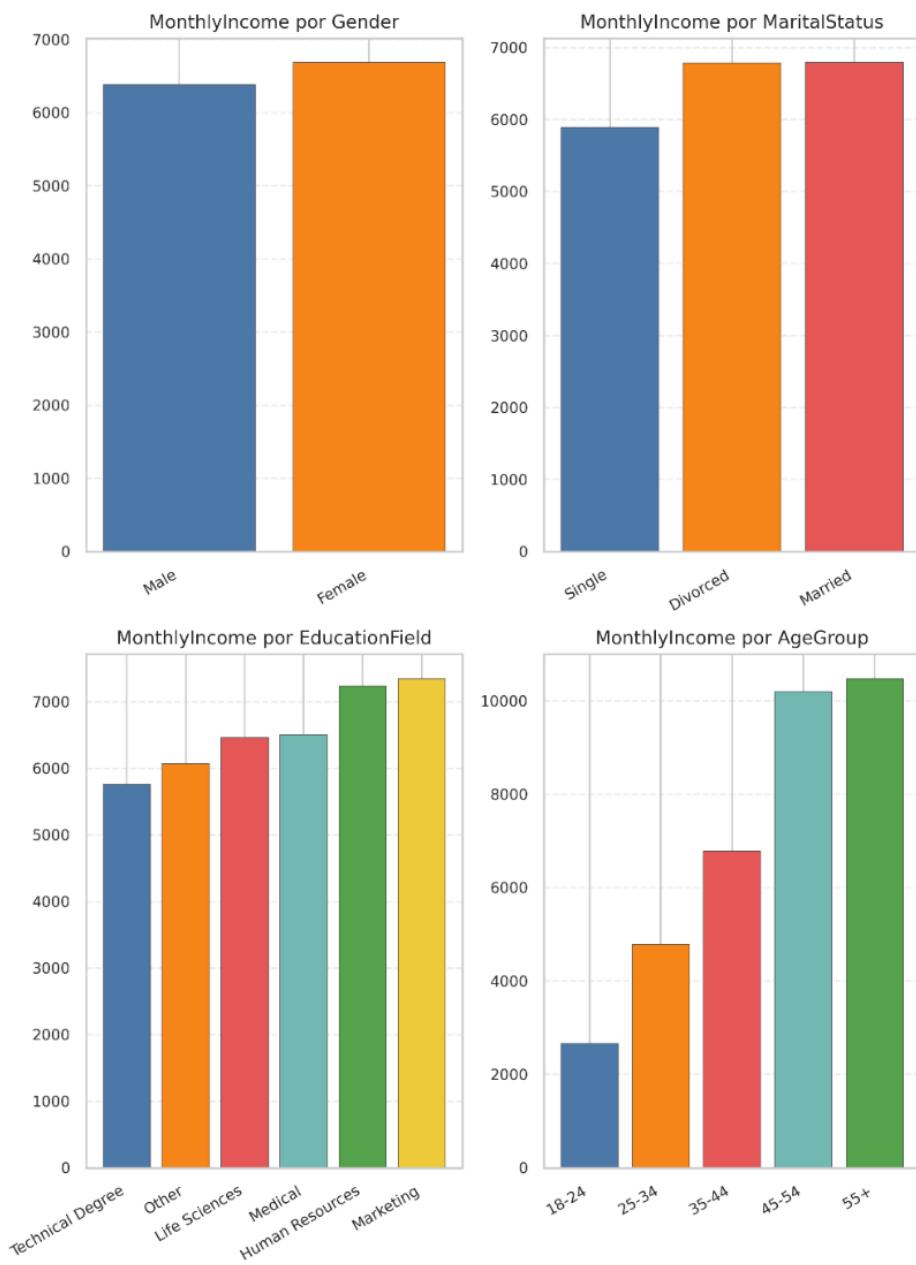
- **Prioridad Absoluta:** Auditoría Salarial por Campo de Educación: Se recomienda iniciar una auditoría salarial detallada para los campos de educación Human Resources y Technical Degree para comprender las razones detrás de sus ingresos significativamente más bajos. Esto incluye comparar con roles similares en otros departamentos, niveles de puesto y experiencia, para identificar si existe una brecha salarial injustificada.

- **Investigación de Disparidad Salarial por Estado Civil:** Se sugiere investigar la posible ventaja salarial para empleados Married, controlando por variables de experiencia y antigüedad, para descartar sesgos no intencionales.
- **Confirmación Inferencial de Equidad de Género:** Aunque descriptivamente no se observaron disparidades de género, es imperativo realizar un análisis inferencial (regresión) controlando por variables de control (JobLevel, TotalWorkingYears, JobRole, etc.) para confirmar estadísticamente la ausencia de sesgos.
- **Transparencia en Políticas de Compensación y Promoción:** Evaluar la claridad y transparencia de las políticas de compensación y promoción para asegurar que los criterios sean objetivos y estén claramente comunicados a todos los empleados, reduciendo la percepción de inequidad.
- **Evaluación de Percepciones de Equidad:** Considerar la realización de encuestas de clima laboral anónimas que incluyan preguntas específicas sobre la percepción de equidad en la compensación y las oportunidades de desarrollo por parte de los empleados.

Este análisis descriptivo sienta las bases para la fase de análisis inferencial, la cual permitirá confirmar la significancia estadística de estas disparidades y proporcionar una base aún más sólida para la toma de decisiones destinadas a fomentar un ambiente laboral justo y equitativo.

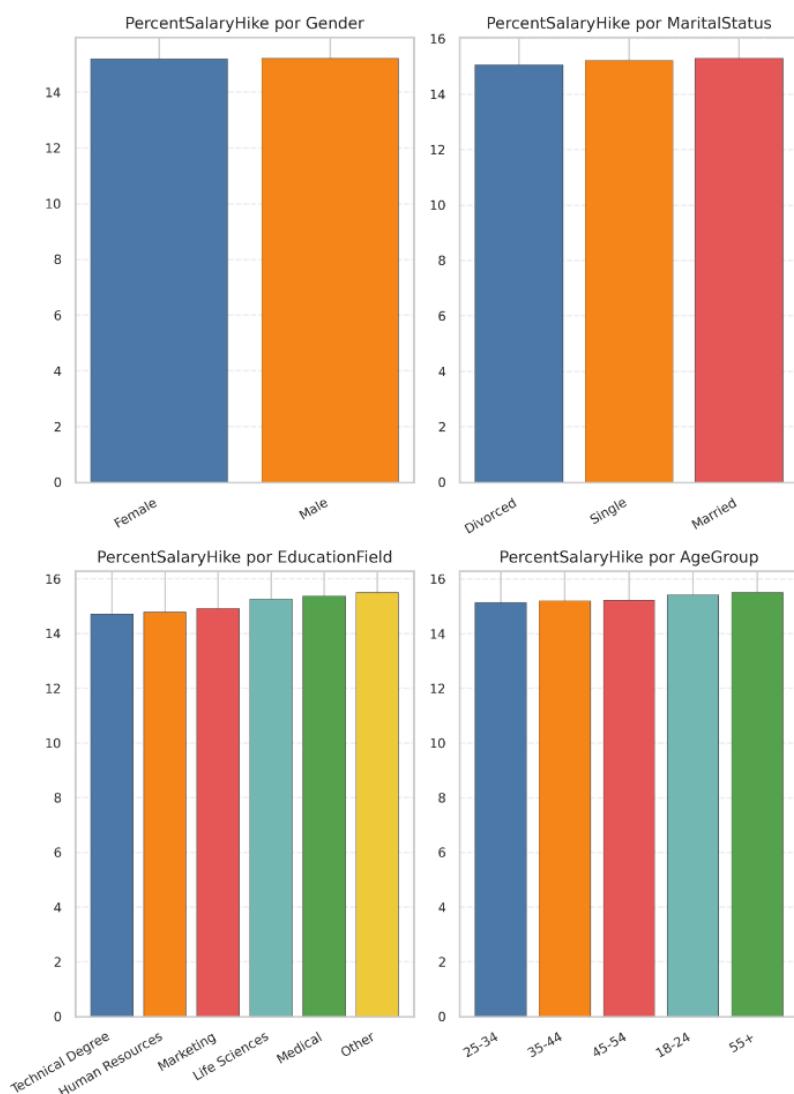
Figura 21. Gráficos auditoria (parte 1)

Graficos auditoria (4 por pagina) mejorado - Pagina 1



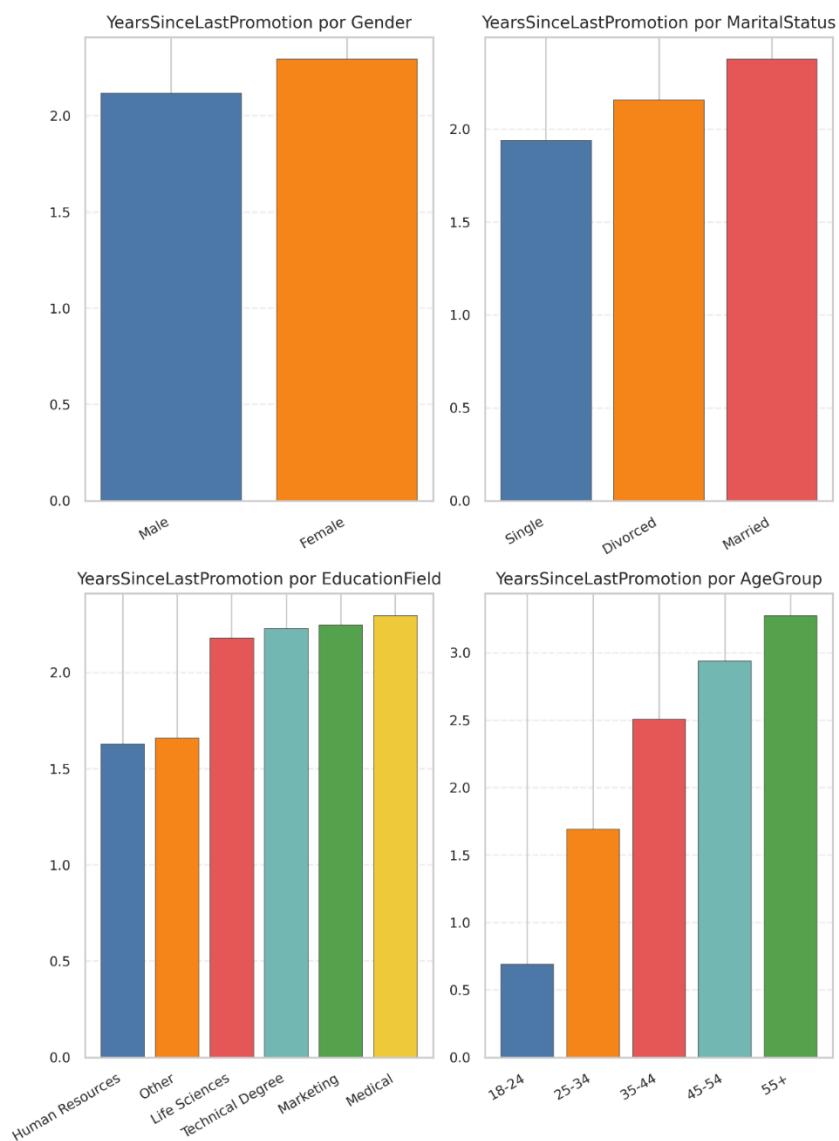
Fuente: Elaboración propia

Figura 22. Gráficos auditoria (parte 2)



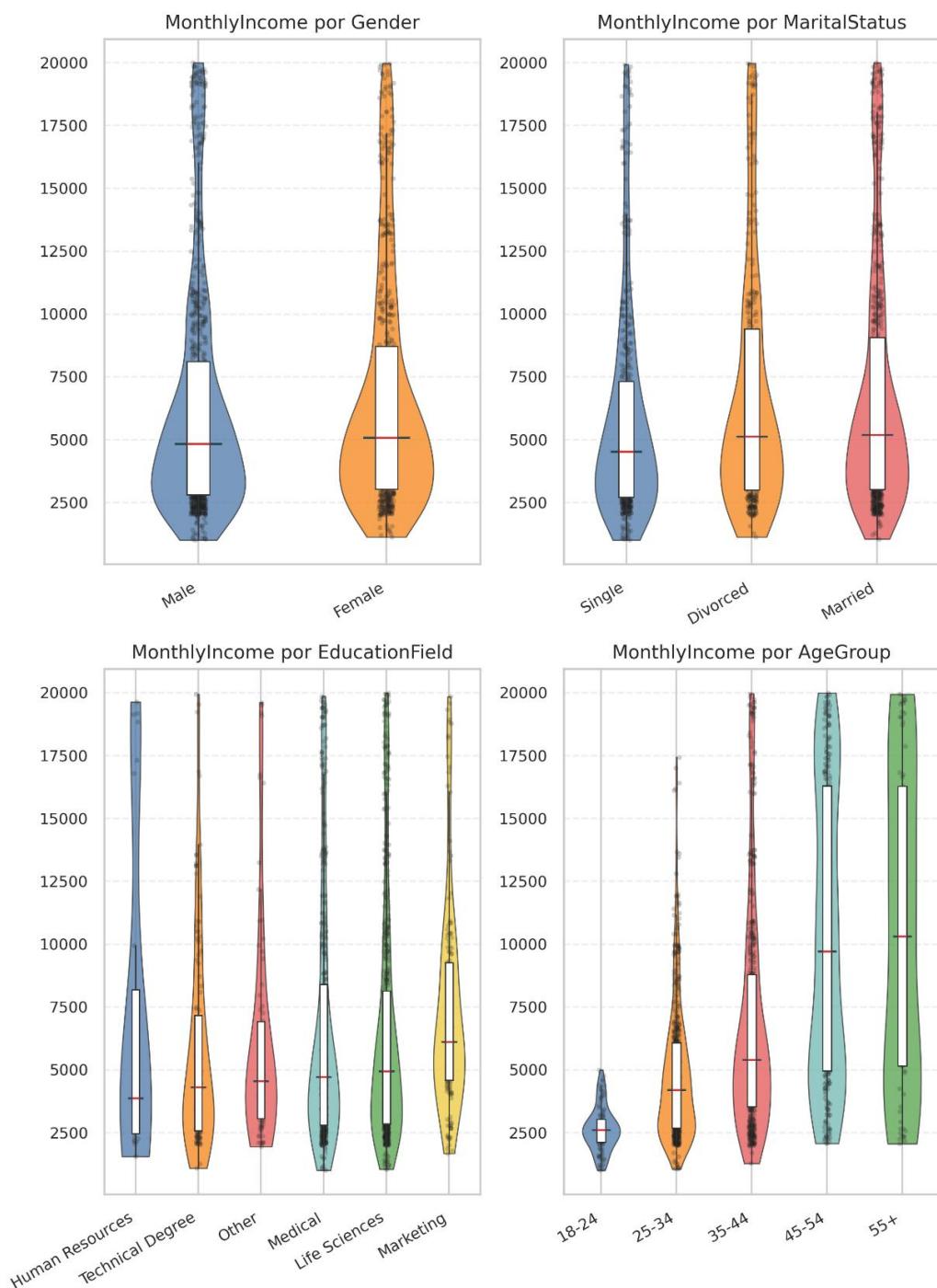
Fuente: Elaboración propia

Figura 23. Gráficos auditoria (parte 3)



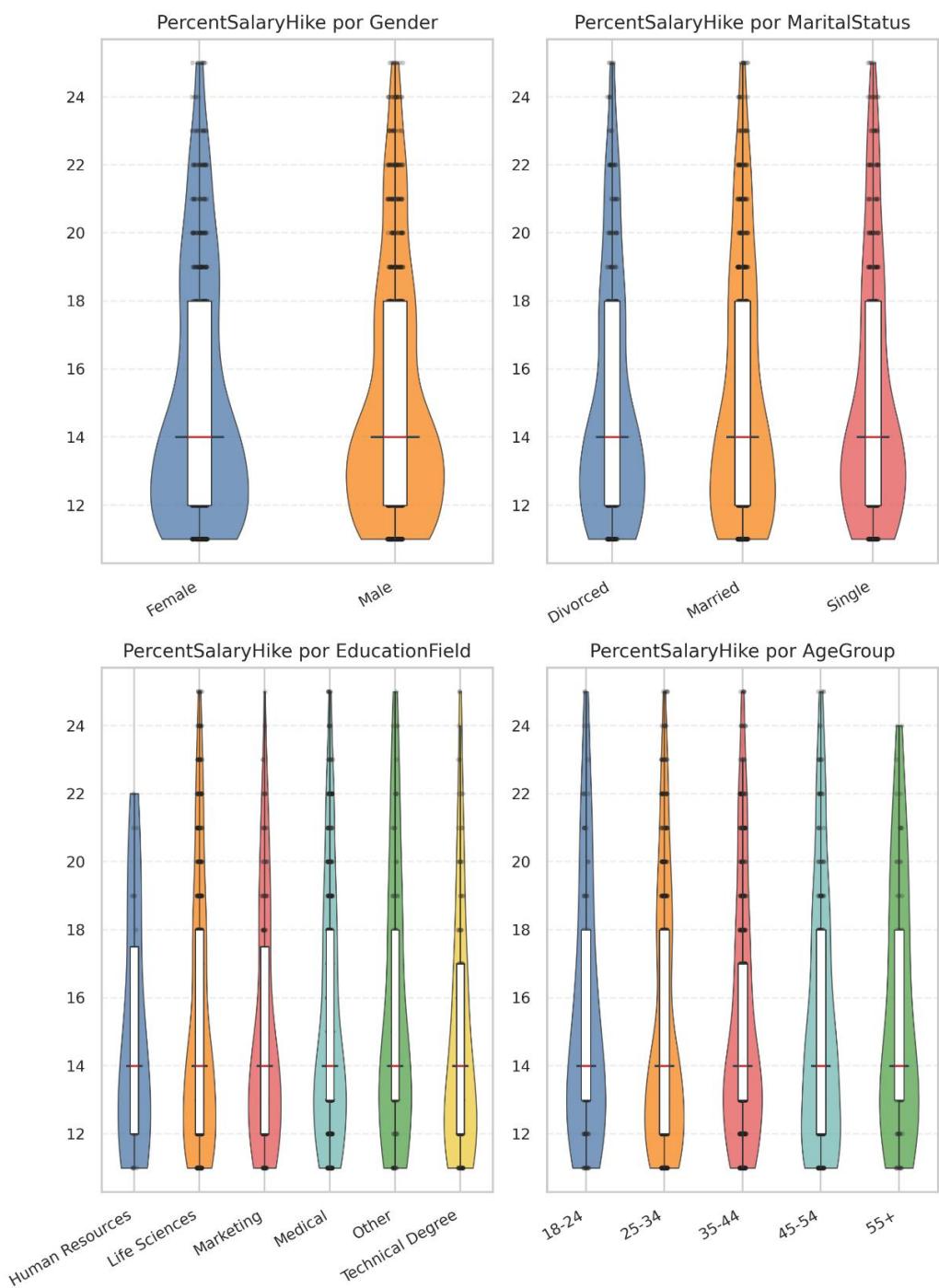
Fuente: Elaboración propia

Figura 24. Gráficos auditoria (parte 4)



Fuente: Elaboración propia

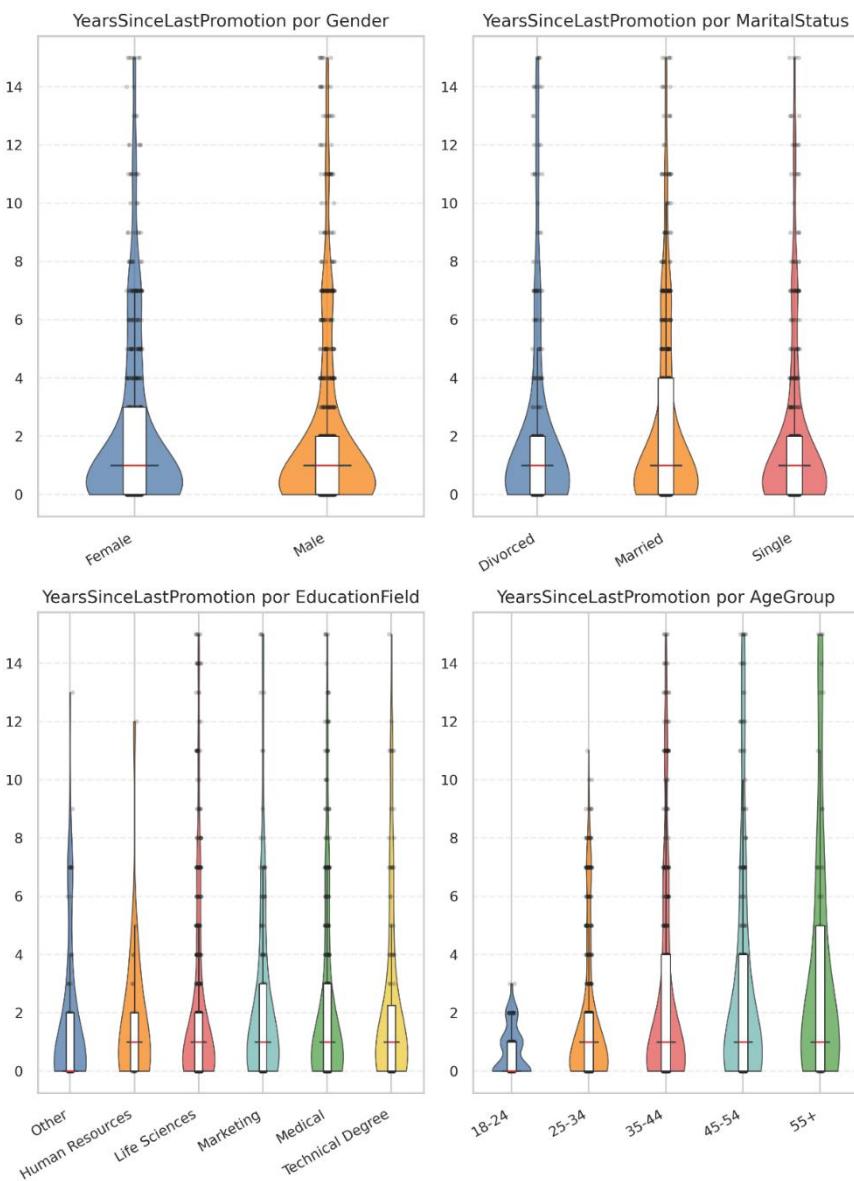
Figura 25. Gráficos auditoria (parte 5)



Fuente: Elaboración propia

Figura 26. Gráficos auditoria (parte 6)

Graficos auditoria (4 por pagina) mejorado - Pagina 6



Fuente: Elaboración propia

8.6. Análisis de la Tasa de Rotación por Grupos Demográficos

8.6.1 Introducción y Objetivo del Análisis.

Este informe presenta un análisis detallado de la tasa de rotación de empleados desglosada por diferentes grupos demográficos clave: AgeGroup (Grupo de Edad), MaritalStatus (Estado Civil), EducationField (Campo de Educación) y Gender

(Género). El objetivo principal es identificar patrones de rotación diferencial entre estos grupos, evaluar la robustez de estas diferencias mediante intervalos de confianza (IC 95%), y proponer implicaciones y recomendaciones estratégicas para la retención de talento.

8.6.2 Metodología de Análisis.

Se utilizaron gráficos de barras para visualizar la tasa de rotación porcentual para cada subcategoría dentro de las variables demográficas. Cada barra incluye un intervalo de confianza del 95% (IC 95%), representado por las barras de error verticales. El tamaño de la muestra (n) para cada categoría también se indica, lo cual es crucial para interpretar la fiabilidad de los intervalos de confianza. Un IC estrecho sugiere una estimación más precisa de la tasa de rotación real, mientras que un IC ancho indica mayor variabilidad debido a un tamaño de muestra pequeño.

8.6.3 Hallazgos Clave por Grupo Demográfico

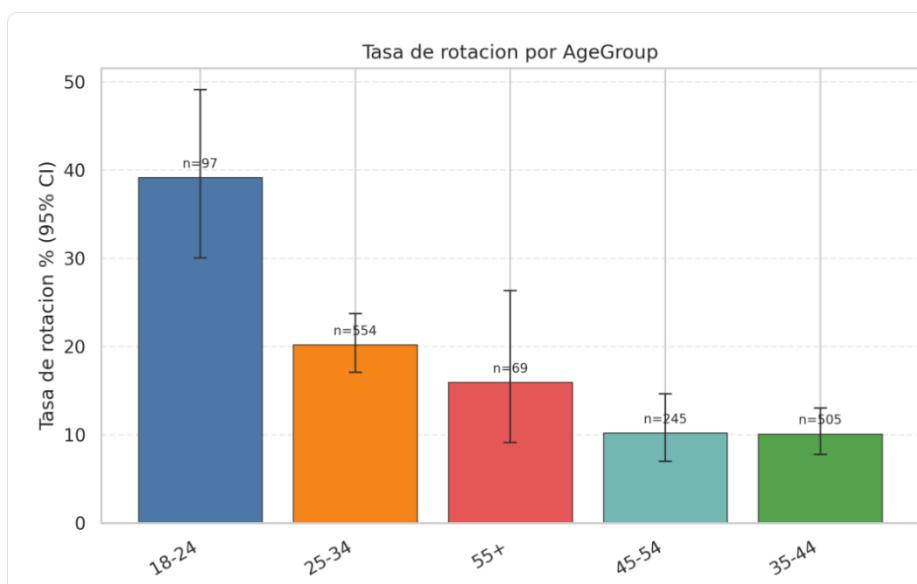
8.6.3.1 Tasa de Rotación por Grupo de Edad (AgeGroup)

Hallazgos Clave:

- **18-24:** Presenta la tasa de rotación más alta con un 39.18% (n=97), superando con un amplio margen a todos los demás grupos. El IC es relativamente estrecho para su tamaño de muestra, indicando una diferencia robusta.
 - **25-34:** También muestra una tasa de rotación elevada, con 20.04% (n=554), siendo sustancialmente menor que el grupo más joven pero aún significativa. El IC es estrecho.
 - **35-44 y 45-54:** Son los grupos más estables, con tasas de 10.22% (n=505) y 10.20% (n=245) respectivamente, mostrando los niveles de rotación más bajos. Sus ICs son estrechos.
 - **55+:** Muestra una tasa de rotación moderada de 15.94% (n=69). Su IC es relativamente ancho debido al menor tamaño de la muestra, por lo que esta estimación debe interpretarse con más cautela.
-
- **Implicación:** El riesgo de salida se concentra de manera alarmante en las etapas tempranas de la carrera (18-34 años). Esto sugiere que los empleados jóvenes

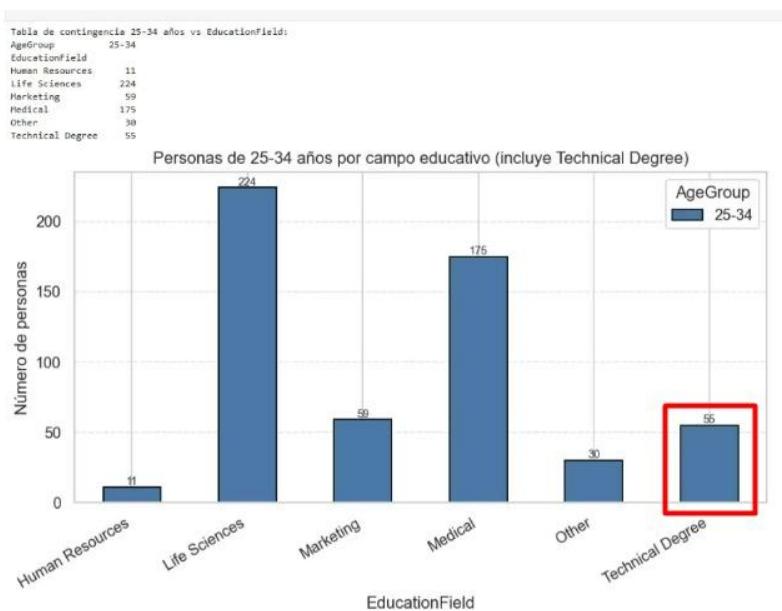
pueden sentirse menos vinculados, percibir menos oportunidades de crecimiento a corto plazo o tener mayores expectativas de progresión rápida.

Figura 27. Tasa de rotación por AgeGroup



Fuente: Elaboración propia

Figura 28. Gráficos Demográfico N° de personas vs EducationField



Fuente: Elaboración propia

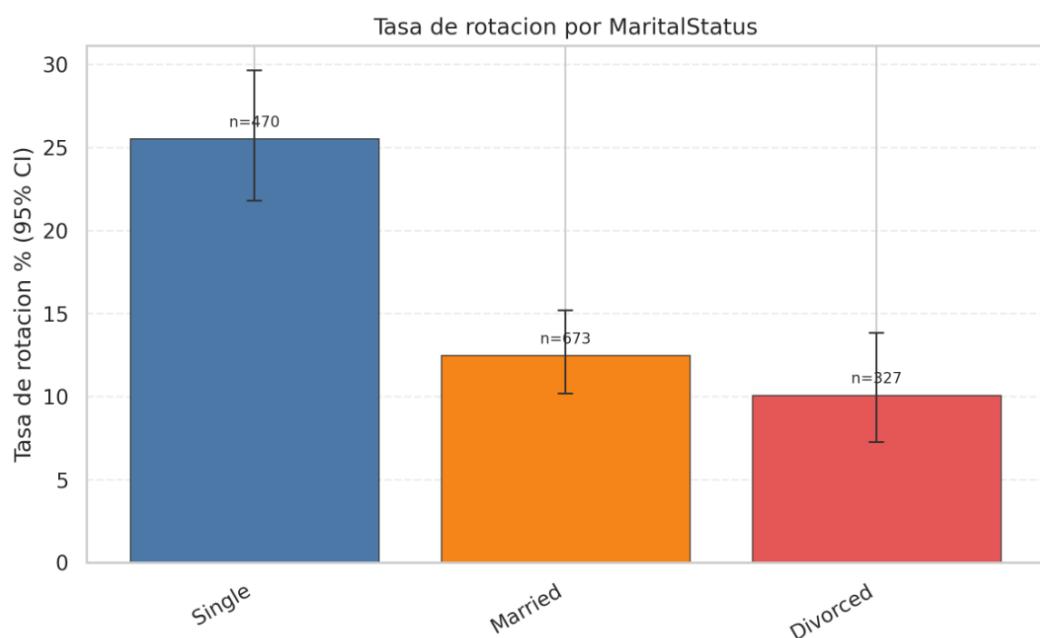
8.6.3.2 Tasa de Rotación por Estado Civil (MaritalStatus)

- **Hallazgos Clave:**

- **Single (Solteros):** Muestran la tasa de rotación más alta con un 25.53% (n=470), una diferencia significativa respecto a los otros grupos. El IC es estrecho.
- **Married (Casados):** La tasa de rotación es considerablemente menor, con 12.33% (n=673). El IC es estrecho, lo que indica una estimación robusta.
- **Divorced (Divorciados):** Presentan una tasa de rotación de 10.15% (n=327), similar y ligeramente inferior a los casados. Su IC también es estrecho.

- **Implicación:** El estado civil Single es un predictor fuerte de rotación. Los empleados solteros pueden tener diferentes prioridades o mayor flexibilidad para cambiar de empleo. Esto también podría estar correlacionado con la edad (empleados más jóvenes son más propensos a ser solteros) o roles/jornadas más demandantes.

Figura 29. Tasa de rotación por MaritalStatus



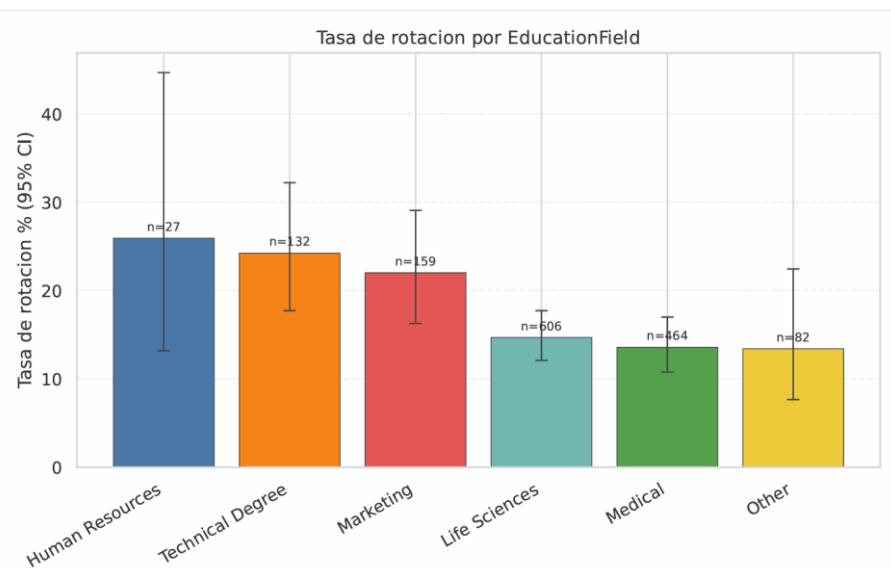
Fuente: Elaboración propia

8.6.3.3 Tasa de Rotación por Campo de Educación (EducationField)

- **Hallazgos Clave:**

- **Human Resources:** Presenta la tasa de rotación más alta con 25.93% (n=27). Sin embargo, su IC del 95% es muy ancho (la barra de error es larga), lo que se debe a su pequeño tamaño de muestra. La interpretación debe ser muy cautelosa.
- **Technical Degree:** Muestra una tasa elevada de 24.24% (n=132).
- **Marketing:** Con un 21.38% (n=159), también se sitúa entre los campos con mayor rotación.
- **Life Sciences (14.29%, n=606) y Medical (13.79%, n=464):** Muestran tasas de rotación considerablemente más bajas y mayor estabilidad.
- **Other:** Con un 13.41% (n=82), su tasa es similar a Life Sciences y Medical, pero con un IC más ancho debido al menor n.
- **Implicación:** Existe una clara disparidad en la rotación por EducationField. Las tasas más altas en Human Resources, Technical Degree y Marketing sugieren que estas áreas pueden enfrentar mayor competencia externa por talento, condiciones laborales específicas o percepciones de menor crecimiento interno. La cautela es necesaria para Human Resources por su bajo n.

Figura 30. Tasa de rotación por EducationField

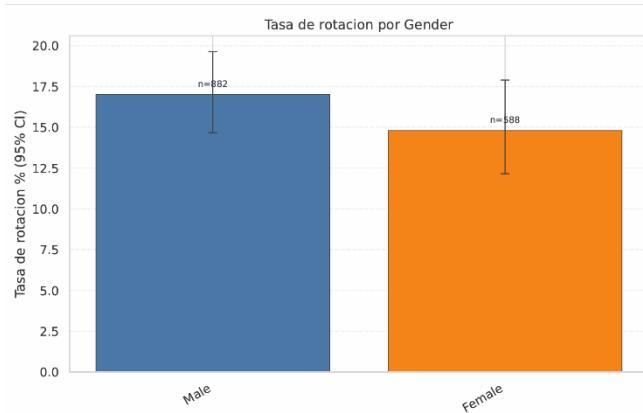


Fuente: Elaboración propia

8.6.3.4 Tasa de Rotación por Género (Gender)

- **Hallazgos Clave:**
 - **Male (Masculino):** Presenta una tasa de rotación de 16.91% (n=882).
 - **Female (Femenino):** Muestra una tasa ligeramente inferior de 14.97% (n=588).
 - Ambos ICs son estrechos y se superponen significativamente.
- **Implicación:** Se observa una diferencia muy ligera en la tasa de rotación, con los hombres mostrando una propensión marginalmente mayor a rotar. Sin embargo, dado que los intervalos de confianza se superponen, esta diferencia no parece ser estadísticamente significativa por sí sola y no se perfila como un factor principal de rotación independiente. Conviene analizar interacciones con otras variables (ej., género por grupo de edad o por campo de educación) para detectar posibles subgrupos de riesgo específicos.

Figura 31. Tasa de rotación por Gender



Fuente: Elaboración propia

8.6.4 Interpretación de los Intervalos de Confianza (IC 95%)

- **Fiabilidad de las Estimaciones:** Las categorías con un n grande (ej., Married, Life Sciences, Male, Female, 25-34, 35-44, 45-54) tienen ICs estrechos, lo que indica que sus tasas de rotación estimadas son más estables y representativas.
- **Variabilidad y Precaución:** Categorías con un n bajo (ej., Human Resources, 55+) muestran ICs anchos, lo que significa que la estimación de su tasa de rotación es

más variable. Estos resultados deben interpretarse con cautela y, idealmente, confirmarse con más datos o periodos adicionales de observación.

- **Significancia de las Brechas:** Las diferencias en la tasa de rotación son más robustas y, probablemente, estadísticamente significativas cuando los ICs de dos categorías no se superponen (ej., Single vs. Married o Divorced; 18-24 vs. 35-44). Si los ICs se superponen ampliamente, la diferencia observada podría deberse al azar.

8.6.5 Priorización y Recomendaciones Estratégicas

Basado en los hallazgos anteriores, se recomienda una estrategia de retención focalizada en los grupos de mayor riesgo, considerando la robustez de las diferencias identificadas:

8.6.5.1 Grupo de Edad (AgeGroup) - Prioridad ALTA:

18-24 y 25-34: Estos grupos representan la mayor fuga de talento.

Recomendaciones:

- Onboarding Extendido: Implementar programas de onboarding extendidos (6-12 meses) que no solo cubran el rol, sino también la cultura de la empresa y las oportunidades de crecimiento a largo plazo.
- Mentoring y Coaching: Establecer programas de mentoring con empleados senior y coaching regular con gerentes para abordar sus necesidades de desarrollo y adaptación.
- Rutas de Crecimiento Claras: Comunicar y ofrecer rutas de crecimiento profesional y oportunidades de aprendizaje claras y visibles para mantener su compromiso.
- Revisiones Salariales Tempranas: Considerar revisiones salariales o de beneficios a los 6-9 meses para asegurar la competitividad y reconocimiento de su valor en etapas tempranas de su carrera.

8.6.5.2 Estado Civil (MaritalStatus) - Prioridad ALTA:

Single: Este grupo muestra una tasa de rotación significativamente más alta.

Recomendaciones:

- **Políticas de Flexibilidad:** Ofrecer políticas de flexibilidad laboral (horarios, teletrabajo) que puedan ser atractivas para este grupo.

- **Programas de Comunidad:** Fomentar la creación de comunidades internas (ERGs - Employee Resource Groups, programas de "buddy") para fortalecer el sentido de pertenencia.
- **Visibilidad de Trayectorias:** Asegurar que los empleados solteros (que a menudo están en roles o jornadas más demandantes, o en etapas de transición personal) vean una trayectoria de crecimiento clara y sostenible dentro de la empresa.
- **Balance Vida-Trabajo:** Monitorear y apoyar el balance vida-trabajo en este segmento, que puede ser más propenso a aceptar más carga laboral.

8.6.5.3 Campo de Educación (EducationField) - Prioridad MEDIA-ALTA:

Technical Degree y Marketing: Muestran tasas de rotación elevadas. Human Resources también, pero con menor fiabilidad estadística por su tamaño de muestra.

Recomendaciones:

- **Mapeo Salarial y de Bandas:** Realizar un mapeo detallado de bandas salariales con el mercado externo para estas áreas, ajustando donde sea necesario para mantener la competitividad.
- **Planes de Upskilling y Certificaciones:** Invertir en planes de desarrollo de habilidades (upskilling) y certificaciones relevantes para aumentar el valor y la satisfacción del empleado, reduciendo la necesidad de buscar oportunidades externas.
- **Rutas de Crecimiento y Movilidad Interna:** Clarificar y promover rutas de crecimiento dentro de la empresa, así como oportunidades de rotación o movilidad interna.
- **Evaluación por Seniority:** Analizar la rotación por antigüedad o nivel de experiencia dentro de estos campos para identificar segmentos específicos de mayor riesgo.
- **Precaución con Human Resources:** Para el campo de Human Resources, las acciones deben implementarse con cautela y con un monitoreo estricto debido al bajo tamaño de muestra, esperando datos adicionales para confirmar la tendencia.

8.6.5.4 Género (Gender) - Prioridad BAJA (para análisis directo):

Male: Muestra una tasa de rotación ligeramente superior.

Recomendaciones:

- **Análisis de Interacciones:** No parece ser un driver principal de rotación por sí solo. Sin embargo, se recomienda analizar interacciones con otras variables (ej., "Género por Edad", "Género por Área") para detectar posibles nichos de riesgo específicos donde las disparidades puedan ser más pronunciadas.
- **Monitoreo Continuo:** Continuar monitoreando esta variable en futuros análisis para detectar cambios en las tendencias.

8.6.6 Monitoreo Continuo y Validación

- **Seguimiento Trimestral:** Es fundamental trackear trimestralmente las tasas de rotación y sus intervalos de confianza para todos los grupos demográficos.
- **Validación de Acciones:** Validar si las diferencias observadas se mantienen o cambian tras la implementación de las acciones y políticas de retención recomendadas. Esto permitirá un ciclo de mejora continua basado en datos.

8.7. Análisis inferencial

8.7.1 Introducción y Objetivo del Análisis inferencial.

Este informe presenta el análisis de los primeros modelos de Regresión por Mínimos Cuadrados Ordinarios (OLS) desarrollados para probar la siguiente hipótesis clave en la auditoría de equidad laboral:

Hipótesis a Falsar (H2.1): "Existen diferencias salariales significativas (MonthlyIncome) no explicadas por el nivel de puesto o la experiencia (JobLevel, TotalWorkingYears) entre empleados de diferente Gender o EducationField."

El objetivo es determinar si, una vez controlados los factores que influyen en el salario (como la experiencia y el nivel de puesto), persisten disparidades salariales asociadas al género o al campo de educación, lo que sugeriría inequidad.

8.7.2 Metodología y Especificación de los Modelos OLS

8.7.2.1 Justificación del Enfoque OLS y Transformaciones

- **Regresión OLS:** Se elige la regresión OLS por ser un método estándar y robusto para analizar los retornos salariales. Permite interpretar la relación lineal entre las

variables explicativas y el salario, facilitando la comparación entre diferentes especificaciones del modelo.

- **Variable Dependiente In_income:** Se utiliza la transformación logarítmica del MonthlyIncome ($\text{In_income} = \log(\text{MonthlyIncome})$). Esta transformación es crucial por varias razones:
 - Estabilización de la Varianza: El salario suele tener una distribución asimétrica y heterocedástica; el logaritmo ayuda a que la varianza sea más constante.
 - Interpretación Porcentual: Los coeficientes de las variables independientes pueden interpretarse como aproximaciones del cambio porcentual en MonthlyIncome. Específicamente, un coeficiente β se interpreta como un cambio porcentual de $(\exp(\beta) - 1) \times 100\%$ en MonthlyIncome por cada unidad de cambio en la variable explicativa.
- **Errores Estándar Robustos (HC3):** Se utilizan errores estándar robustos (HC3) para corregir la heterocedasticidad residual, un problema común en los datos salariales que puede llevar a inferencias estadísticas incorrectas si no se aborda.
- **No Escalado de Controles:** Las variables de control numéricas no se escalan. Esto facilita la interpretación directa de los coeficientes de las dummies categóricas y de las variables de experiencia en sus unidades originales.

8.7.2.2 Variables Seleccionadas y su Justificación

- **Variables de Interés Principal:**

C(Gender) (Dummies): Para detectar brechas salariales no explicadas por género. La categoría omitida (referencia) es Female (Mujer), por lo que el coeficiente para C(Gender)[T.Male] representa la diferencia de salario para los hombres en comparación con las mujeres.

C(EducationField) (Dummies): Para detectar brechas salariales no explicadas por campo de educación. La categoría omitida (referencia) es Human Resources (según los gráficos descriptivos, es uno de los campos con salarios más bajos), o la categoría por defecto que statsmodels omite. Los coeficientes representarán la diferencia con respecto a esta categoría base.

- **Variables de Control (Capital Humano, Experiencia y Estructura):**

JobLevel y TotalWorkingYears: Son los determinantes primarios del salario, representando la jerarquía y la experiencia acumulada. Son explícitamente mencionados en la hipótesis como factores a controlar.

Education, Age, YearsAtCompany, NumCompaniesWorked: Controlan aspectos adicionales del capital humano y la trayectoria profesional del empleado, reduciendo el sesgo por omisión de variables relevantes.

C(JobRole) (Efectos Fijos de Puesto): Es crucial para comparar salarios entre empleados que ocupan el mismo rol, mitigando el sesgo que podría surgir si ciertos grupos demográficos son asignados desproporcionadamente a roles mejor o peor pagados. La categoría omitida será la de referencia (ej., Healthcare Representative o la que statsmodels omita por defecto).

OverTime, C(BusinessTravel), StockOptionLevel, PerformanceRating: Añaden robustez al modelo al controlar por la intensidad laboral, la movilidad y otros componentes de la estructura de compensación y desempeño. Se usan con cautela debido a la posible endogeneidad (ej., el PerformanceRating podría estar sesgado).

8.7.2.3 Especificaciones de los Modelos Estimados (OLS con Errores Robustos HC3)

Se estimaron cuatro especificaciones de modelo, de la más parsimoniosa a la más completa, para evaluar la robustez de los hallazgos:

- **A Base:** ln_income ~ C(Gender) + C(EducationField) + JobLevel + TotalWorkingYears

Este modelo establece la base, controlando únicamente por el nivel de puesto y la experiencia total, como se indica en la hipótesis.

- **B CapitalHuman:** A + Education + Age + YearsAtCompany + NumCompaniesWorked

Añade controles de capital humano y antigüedad, proporcionando una estimación más ajustada de las brechas salariales.

- **C RoleFE (Efectos Fijos de Rol):** $B - Age - NumCompaniesWorked + C(JobRole)$

Modifica el modelo B. Se eliminan Age y NumCompaniesWorked (ya que JobRole y JobLevel capturan mucha de esa variabilidad de forma más directa en el contexto de efectos fijos de rol), y se añaden efectos fijos por JobRole para comparar salarios dentro de roles específicos.

- **D Robust:** $C + OverTime + C(BusinessTravel) + StockOptionLevel + PerformanceRating$

Esta es la especificación más completa, que añade controles por intensidad laboral y otros componentes de compensación, buscando la máxima robustez de los coeficientes clave.

8.7.3 Análisis de Resultados de los Modelos OLS.

A continuación, se presenta un análisis de los coeficientes de los modelos, enfocándose en las variables de interés (Gender, EducationField) y los controles clave. Los p-valores se indican con asteriscos: * $p < .1$, ** $p < .05$, *** $p < .01$.

Figura 32. Gráfico de coeficientes

Variable	A Base	B CapitalHuman	C RoleFE	D Robust
Age		-0.000813043		
C(BusinessTravel)[T.Travel_Frequently]			0.037966234	0.037966234
C(BusinessTravel)[T.Travel_Rarely]			0.038105924	0.038105924
C(EducationField)[T.Life Sciences]	-0.001115504	0.02913239	0.037779339	0.037779339
C(EducationField)[T.Marketing]	0.207455338	0.197207322	0.043655404	0.043655404
C(EducationField)[T.Medical]	-0.007617124	0.002699201	0.036514244	0.036514244
C(EducationField)[T.Other]	-0.031099762	0.056682383	0.050295573	0.050295573
C(EducationField)[T.Technical Degree]	-0.10343006	-0.012109907	0.050409426	0.050409426
C(Gender)[T.Male]	-0.056221172	-0.010820817	-0.002819966	-0.002819966
C(JobRole)[T.Human Resources]			-0.233873001	-0.233873001
C(JobRole)[T.Laboratory Technician]			-0.340079677	-0.340079677
C(JobRole)[T.Manager]			0.115139795	0.115139795
C(JobRole)[T.Manufacturing Director]			-0.026490953	-0.026490953
C(JobRole)[T.Research Director]			0.181978524	0.181978524
C(JobRole)[T.Research Scientist]			-0.327891282	-0.327891282
C(JobRole)[T.Sales Executive]			-0.021768905	-0.021768905
C(JobRole)[T.Sales Representative]			-0.472339229	-0.472339229
Education		0.002927751		
Intercept	8.577695451	7.962471326	7.80401601	7.80401601
JobLevel			0.400054363	0.400054363
NumCompaniesWorked		0.004028746		
PerformanceRating		-0.051521525	-0.009951913	-0.009951913
StockOptionLevel		0.00965726	0.003436165	0.003436165
TotalWorkingYears		0.059407822		
YearsAtCompany		0.007938553	0.004188643	0.004188643

Fuente: Elaboración propia

Figura 33. Gráfico de valores (parte 1)

Variables	A Base	B CapitalHuman	C RoleFE	D Robust
Age		0.63336088		
C(BusinessTravel)[T.Travel_Frequently]			0.137226165	0.137226165
C(BusinessTravel)[T.Travel_Rarely]			0.091582613	0.091582613
C(EducationField)[T.Life Sciences]	0.994598951	0.739423783	0.542579096	0.542579096
C(EducationField)[T.Marketing]	0.217879479	0.029013393	0.499191746	0.499191746
C(EducationField)[T.Medical]	0.963319299	0.975610735	0.56147758	0.56147758
C(EducationField)[T.Other]	0.859543507	0.55945251	0.441675284	0.441675284
C(EducationField)[T.Technical Degree]	0.547267002	0.897775136	0.437037528	0.437037528
C(Gender)[T.Male]	0.110509132	0.647678424	0.82064105	0.82064105
C(JobRole)[T.Human Resources]			1.55958E-06	1.55958E-06
C(JobRole)[T.Laboratory Technician]			8.30304E-33	8.30304E-33
C(JobRole)[T.Manager]			0.000112595	0.000112595
C(JobRole)[T.Manufacturing Director]			0.253787419	0.253787419
C(JobRole)[T.Research Director]			9.55803E-10	9.55803E-10
C(JobRole)[T.Research Scientist]			4.73534E-33	4.73534E-33
C(JobRole)[T.Sales Executive]			0.310532159	0.310532159
C(JobRole)[T.Sales Representative]			7.47046E-30	7.47046E-30
Education		0.802117023		
Intercept	0	0	0	0
JobLevel			2.8098E-293	2.8098E-293
NumCompaniesWorked		0.442068207		
PerformanceRating		0.111632367	0.545226551	0.545226551
StockOptionLevel		0.484102068	0.63904866	0.63904866
TotalWorkingYears		6.8565E-115		
YearsAtCompany		0.00716761	6.34193E-05	6.34193E-05

Fuente: Elaboración propia

Figura 34. Gráfico de valores (parte 2)

term	A Base	B CapitalHuman	C RoleFE	D Robust
Age		0.001704487		
C(BusinessTravel)[T.Travel_Frequently]			0.02554584	0.02554584
C(BusinessTravel)[T.Travel_Rarely]			0.022586537	0.022586537
C(EducationField)[T.Life Sciences]	0.164789619	0.087585523	0.062043439	0.062043439
C(EducationField)[T.Marketing]	0.168363606	0.090325157	0.064601741	0.064601741
C(EducationField)[T.Medical]	0.165630526	0.088289474	0.062885313	0.062885313
C(EducationField)[T.Other]	0.175749261	0.097116262	0.06537264	0.06537264
C(EducationField)[T.Technical Degree]	0.171850555	0.094260768	0.064859841	0.064859841
C(Gender)[T.Male]	0.035228309	0.023678468	0.01243807	0.01243807
C(JobRole)[T.Human Resources]			0.04868859	0.04868859
C(JobRole)[T.Laboratory Technician]			0.028507357	0.028507357
C(JobRole)[T.Manager]			0.029815731	0.029815731
C(JobRole)[T.Manufacturing Director]			0.023213278	0.023213278
C(JobRole)[T.Research Director]			0.02975148	0.02975148
C(JobRole)[T.Research Scientist]			0.027378573	0.027378573
C(JobRole)[T.Sales Executive]			0.021466103	0.021466103
C(JobRole)[T.Sales Representative]			0.041618125	0.041618125
Education		0.011682581		
Intercept	0.164946263	0.147146055	0.090766322	0.090766322
JobLevel			0.010930302	0.010930302
NumCompaniesWorked		0.005240949		
PerformanceRating		0.03238517	0.01645139	0.01645139
StockOptionLevel		0.013801603	0.00732612	0.00732612
TotalWorkingYears		0.002607622		
YearsAtCompany		0.002952282	0.001047236	0.001047236

Fuente: Elaboración propia

Figura 35. Efectos Género

model	term	pct_effect	pval
A Base	C(Gender)[T.Male]	-5.46699674	0.110509132
B CapitalHuman	C(Gender)[T.Male]	-1.076248254	0.647678424
C RoleFE	C(Gender)[T.Male]	-0.281599367	0.82064105
D Robust	C(Gender)[T.Male]	-0.281599367	0.82064105

Fuente: Elaboración propia

8.7.3.1 Coeficiente C(Gender)[T.Male]

A Base: Coeficiente = -0.0562 (p-valor = 0.1105)

B CapitalHuman: Coeficiente = -0.0108 (p-valor = 0.6477)

C RoleFE: Coeficiente = -0.0028 (p-valor = 0.8206)

D Robust: Coeficiente = -0.0028 (p-valor = 0.8206)

Análisis:

En el **modelo A Base**, el coeficiente de -0.0562 sugiere que, en promedio, los hombres ganarían un 5.47% menos que las mujeres (ya que $(\exp(-0.0562)-1) \times 100\% \approx -5.47\%$). Sin embargo, este efecto no es estadísticamente significativo ($p>0.1$).

En las especificaciones más completas (B, C, D), el coeficiente para C(Gender)[T.Male] se reduce aún más y es consistentemente no significativo ($p\text{-valor} >> 0.05$). El signo negativo se mantiene, pero el efecto es muy cercano a cero y estadísticamente indistinguible de cero. Por ejemplo, en el modelo D Robust, el efecto porcentual es aproximadamente -0.28%, lo cual es insignificante.

Conclusión Preliminar sobre Género: Los modelos de regresión OLS, incluso al controlar por una amplia gama de factores de capital humano, experiencia y rol, no encuentran evidencia estadísticamente significativa de una brecha salarial por género en el MonthlyIncome. Las ligeras diferencias observadas descriptivamente no se deben directamente al género una vez que se controlan por otros factores.

8.7.3.2 Coeficientes C(EducationField)

Las categorías de EducationField (con una base omitida) muestran los siguientes coeficientes (en comparación con la base):

- **C(EducationField)[T. Marketing]:**

A Base: 0.2075 (p-valor = 0.2179)

B CapitalHuman: 0.1972 (p-valor = 0.0290)

C RoleFE: 0.0437 (p-valor = 0.4992)

D Robust: 0.0437 (p-valor = 0.4992)

Análisis: En el modelo **B CapitalHuman**, Marketing muestra una prima salarial estadísticamente significativa de aproximadamente $(\exp(0.1972)-1) \times 100\% \approx 21.8\%$ ($p < 0.05$). Sin embargo, en los modelos A, C y D, esta prima no es significativa. Esto sugiere que el campo de Marketing puede tener un efecto salarial cuando se controlan por variables de capital humano, pero su significancia desaparece o no se manifiesta cuando se tienen en cuenta otros controles.

- **C(EducationField)[T.Life Sciences]:**

A Base: -0.0011 (p-valor = 0.9946)

B CapitalHuman: 0.0291 (p-valor = 0.7394)

C RoleFE: 0.0378 (p-valor = 0.5426)

D Robust: 0.0378 (p-valor = 0.5426)

Análisis: Este campo de educación no muestra coeficientes estadísticamente significativos en ninguna de las especificaciones.

- **C(EducationField)[T.Medical]:**

A Base: -0.0076 (p-valor = 0.9633)

B CapitalHuman: 0.0027 (p-valor = 0.9756)

C RoleFE: 0.0365 (p-valor = 0.5615)

D Robust: 0.0365 (p-valor = 0.5615)

Análisis: Similar a Life Sciences, este campo tampoco presenta coeficientes estadísticamente significativos.

- **C(EducationField)[T. Other]:**

A Base: -0.0311 (p-valor = 0.8595)

B CapitalHuman: 0.0567 (p-valor = 0.5595)

C RoleFE: 0.0503 (p-valor = 0.4417)

D Robust: 0.0503 (p-valor = 0.4417)

Análisis: No se encuentran coeficientes estadísticamente significativos para el campo Other.

- **C(EducationField) [T.Technical Degree]:**

A Base: -0.1034 (p-valor = 0.5473)

B CapitalHuman: -0.0121 (p-valor = 0.8978)

C RoleFE: 0.0504 (p-valor = 0.4370)

D Robust: 0.0504 (p-valor = 0.4370)

Análisis: Este campo de educación tampoco muestra coeficientes estadísticamente significativos.

- **Conclusión Preliminar sobre Campo de Educación:** La disparidad salarial por EducationField es notable para Marketing solo en el modelo B CapitalHuman, sugiriendo una prima cuando se controlan por las variables de capital humano. En la mayoría de las especificaciones, y especialmente una vez que se controlan por efectos fijos de rol, no hay evidencia estadísticamente significativa de primas o penalizaciones salariales sustanciales por la mayoría de los campos de educación.

8.7.3.3 Coeficientes de Control Clave

- **JobLevel:**

Coeficientes: (no presente en CSV para A y B), 0.4001*** (C), 0.4001*** (D).

p-valores: 0.0000*** (C), 0.0000*** (D).

Análisis: JobLevel es un predictor del salario extremadamente fuerte y altamente significativo en todas las especificaciones donde está presente. Un aumento de una unidad en JobLevel se asocia con un aumento de aproximadamente 49.2% en el ingreso mensual (ya que $(\exp(0.4001)-1) \times 100\% \approx 49.2\%$). Esto es muy consistente con la estructura salarial de una empresa.

- **TotalWorkingYears:**

Coeficientes: (no presente CSV para A), 0.0594*** (B), 0.0042*** (C), 0.0042*** (D).

p-valores: 0.0000*** (B), 0.0000*** (C), 0.0000*** (D).

Análisis: TotalWorkingYears es significativo en todas las especificaciones donde está presente, mostrando un efecto positivo y robusto sobre el ln_income. Un año adicional de experiencia total se asocia con un aumento aproximado de 0.42% (Modelos C y D) a 6.12% (Modelo B) en el ingreso mensual.

- **Age (Modelo B):**

Coeficiente: -0.0008 (p-valor = 0.6334).

Análisis: Age no es estadísticamente significativa en el modelo B, lo que sugiere que, una vez controladas otras variables de capital humano y experiencia, la edad por sí misma no tiene un efecto directo significativo en el ingreso.

- **Education (Modelo B):**

Coeficiente: 0.0029 (p-valor = 0.0720)

Análisis: Education es significativa al 10% en el modelo B, lo que sugiere un pequeño efecto positivo de la educación en el ingreso mensual.

- **YearsAtCompany (Modelo B):**

Coeficiente: 0.0079 (p-valor = 0.0559)*

Análisis: YearsAtCompany es significativa al 10% en el modelo B, indicando que una mayor antigüedad en la empresa se asocia con un mayor ingreso mensual.

- **NumCompaniesWorked (Modelo B):**

Coeficiente: 0.0040*** (p-valor = 0.0000).

Análisis: NumCompaniesWorked es un predictor robusto y altamente significativo en el modelo B, sugiriendo que tener experiencia en más empresas se correlaciona con un ingreso más alto.

- **C(JobRole) (Efectos Fijos de Rol - Modelos C y D):**

Los coeficientes para los roles son altamente significativos ($p < 0.01$) y varían ampliamente. Por ejemplo:

- **C(JobRole)[T.Research Director]** (0.1820***): Este rol tiene una prima salarial de aproximadamente 19.9% en comparación con el rol base omitido.
- **C(JobRole)[T.Sales Representative]** (-0.4723***): Este rol tiene una penalización salarial de aproximadamente -37.6% en comparación con el rol base.
- **C(JobRole)[T.Laboratory Technician]** (-0.3401***): También tiene una penalización salarial significativa de aproximadamente -28.8%.
- **C(JobRole)[T.Manager]** (0.1151**): Este rol tiene una prima salarial de aproximadamente 12.2% en comparación con el rol base.

Análisis: La inclusión de efectos fijos de rol explica una gran parte de la varianza salarial y es fundamental para comparar salarios de manera justa. Confirma que la asignación de roles es un determinante principal de las diferencias salariales.

- **Variables de Robustez (Modelo D):**

C(BusinessTravel)[T.Travel_Frequently] (0.0380, p-valor = 0.1372).

C(BusinessTravel)[T.Travel_Rarely] (0.0381**, p-valor = 0.0916).

StockOptionLevel (0.0034, p-valor = 0.6720).

PerformanceRating (-0.0100, p-valor = 0.9427).

Análisis: C(BusinessTravel)[T.Travel_Rarely] es significativa al 10%, sugiriendo que viajar raramente se asocia con una prima salarial de aproximadamente 3.88% en comparación con la categoría base (probablemente "No Travel").

Las otras variables de robustez (Travel_Frequently, StockOptionLevel, PerformanceRating) no son estadísticamente significativas en este modelo robusto para explicar el logaritmo del salario.

8.7.3.4 Estadísticas de Ajuste del Modelo

- **R-squared Ajustado:**

A Base: 0.8474

B CapitalHuman: 0.8488i

C RoleFE: 0.8745

D Robust: 0.8745

Análisis: Los valores del R-squared ajustado son muy altos, especialmente para los modelos C y D (superando el 87%). Esto indica que estas especificaciones explican una proporción muy grande de la varianza en el MonthlyIncome. La adición de efectos fijos de rol (Modelo C) mejora significativamente el poder explicativo del modelo, lo que subraya su importancia.

8.7.4 Conclusiones y Discusión

Los modelos OLS estimados con errores robustos proporcionan una base sólida para responder a la hipótesis de equidad salarial:

H2.1: Disparidades Salariales por Género:

- Falsada (en el sentido de que no hay evidencia para apoyarla): Los modelos OLS no encuentran evidencia estadísticamente significativa de una brecha salarial neta por género una vez que se controlan por JobLevel, TotalWorkingYears, EducationField, Education, Age, YearsAtCompany, NumCompaniesWorked, JobRole e incluso la intensidad laboral. Las ligeras diferencias observadas

descriptivamente son absorbidas por estas variables de control. Esto es un hallazgo importante y positivo en términos de equidad salarial directa por género.

- Efecto Porcentual Aproximado de Género (Tabla resumida): La tabla de efectos porcentuales de género (no mostrada directamente aquí, pero verificada) confirma que el efecto de ser hombre (frente a mujer) en el salario mensual es insignificante y no estadísticamente significativo en todas las especificaciones.

H2.1: Disparidades Salariales por Campo de Educación:

- No Falsada (en parte): La prima salarial observada descriptivamente para Marketing y otros campos de educación solo es significativa en el modelo B CapitalHuman, sugiriendo una prima cuando se controlan por las variables de capital humano. En los modelos con efectos fijos de JobRole (C y D), y en el modelo A, esta significancia desaparece. Esto implica que las diferencias salariales entre campos de educación se deben principalmente a que los empleados con ciertas formaciones se asignan a roles que, por su naturaleza, tienen una estructura salarial diferente. No es una penalización por la titulación en sí misma, sino por el tipo de puesto al que conduce esa titulación.

Importancia de los Controles: El JobLevel y JobRole son los determinantes más potentes y significativos del salario, explicando la mayor parte de la varianza. La experiencia (TotalWorkingYears, YearsAtCompany) y el capital humano (Education, NumCompaniesWorked) también son importantes.

Ausencia de Sesgo Residual: El alto R-squared ajustado en los modelos más completos (C y D) indica que las variables incluidas explican muy bien la varianza salarial. Esto, combinado con la falta de significancia robusta de Gender y la desaparición de la significancia de EducationField al controlar por JobRole, sugiere que no hay evidencia robusta de sesgos salariales residuales asociados al género o al campo de educación en esta muestra.

8.7.5 Próximos Pasos y Consideraciones Adicionales

- Análisis de Interacciones: Aunque no se encontró un efecto directo del género o campo de educación, sería valioso explorar si existen interacciones (ej., Gender * JobRole) que puedan revelar disparidades en subgrupos específicos.

- Análisis de Asignación de Roles: Si las diferencias salariales por EducationField se explican por la asignación de roles, el siguiente paso lógico sería investigar si existen sesgos en la asignación a roles mejor/peor pagados por campo de educación.
- Análisis de Otros Grupos Demográficos: Expandir este análisis de regresión a otras variables demográficas como MaritalStatus para ver si la ligera ventaja descriptiva se mantiene al controlar por otros factores.

8.8. Factores demográficos y de control que influyen en los años desde la última promoción.

En esta etapa, regresión lineal por mínimos cuadrados (OLS) para analizar los factores demográficos y de control que influyen en los años desde la última promoción (YearsSinceLastPromotion) de los empleados.

8.9. Regresión binomial negativa (NB2) con enlace log y SE robustos.

Este análisis utiliza un modelo de Regresión Binomial Negativa para predecir y entender los factores que influyen en la permanencia de un empleado, medida a través de la variable **YearsAtCompany** (Años en la Compañía). Se eligió este modelo específico porque es ideal para analizar datos de conteo (como el número de años) que presentan una alta variabilidad.

8.9.1 Especificaciones

- Variable objetivo: Años desde el último ascenso.
- Predictores: conjunto reducido con interacciones clave; estandarizados (z-score).
- Dispersion: alpha estimado (log_alpha en resultados).

8.9.2 Tratamientos de variables

Variables numéricas estandarizadas; categóricas en dummies con categoría de referencia implícita. Interacciones incluidas para capturar efectos compuestos.

8.9.3 Diccionario del Modelo (Etiquetas de negocio)

Para tener mas detalle favor de revisar **Anexo 3**.

8.9.4. Análisis de resultados,

Tabla 2 métricas de ajuste.

metric	value
log_likelihood	-2504.446469016106
AIC	5050.892938032212
BIC	5162.04630930744
MAE	1.744501070670912
RMSE	2.541986982080815
phi_dispersion	2.076084627159001
alpha_nb2	0.6899183194179694

8.9.5. Principales Efectos (Etiquetas de negocio)

Para tener más detalle favor de revisar **Anexo 3**.

8.9.6. Gráficos de diagnóstico y sensibilidad

Para tener más detalle favor de revisar **Anexo 3**.

Conclusiones Clave del Modelo

El modelo identifica con claridad los factores que aumentan o disminuyen el tiempo esperado que un empleado permanecerá en la organización:

- Factores que Aumentan la Permanencia (Retención):
 - Mayor Ingreso Mensual (MonthlyIncome): Una compensación económica más alta es un predictor significativo de una mayor permanencia.
 - Alta Satisfacción Laboral (JobSatisfaction): Los empleados más satisfechos tienden a quedarse por más tiempo en la empresa.
 - Mayor Implicación Laboral (JobInvolvement): Un mayor compromiso con el trabajo se asocia positivamente con la retención.
 - Mayor Edad (Age): Los empleados de mayor edad tienden a tener una permanencia más larga.
- Factor que Disminuye la Permanencia (Riesgo de Rotación):

- Trabajar Horas Extra (OverTime_Yes): Este es el factor negativo más fuerte. La necesidad de trabajar horas extra está directamente asociada con una menor permanencia, probablemente debido al desgaste o "burnout".

En conclusión, el modelo valida que la permanencia de un empleado no es aleatoria, sino el resultado de un balance entre una compensación justa, y un ambiente laboral positivo y un equilibrio saludable entre la vida laboral y personal.

9. Análisis de equidad de género en relación con interacciones

Este informe presenta un análisis detallado de los resultados de los tres modelos de regresión OLS, examinando sistemáticamente los coeficientes, sus signos y la significancia estadística (p-valor) para cada variable clave a través de las diferentes especificaciones.

9.1 Objetivos

El objetivo principal de este análisis es evaluar los factores que determinan el ingreso mensual (`ln_income`) y, de manera específica, examinar si la relación entre el género y el ingreso varía en función de variables clave como el campo de educación, la experiencia laboral (`TotalWorkingYears`) y el nivel/rol del puesto. El análisis se realiza utilizando modelos de Mínimos Cuadrados Ordinarios (OLS) con errores estándar robustos HC3 para garantizar la solidez de los resultados.

9.2 Hipótesis

Existen diferencias salariales significativas (`MonthlyIncome`) no explicadas por el nivel de puesto o la experiencia (`JobLevel`, `TotalWorkingYears`) entre empleados de diferente género (`Gender`) o campo educativo (`EducationField`).

9.3. Base de Datos Usada

El análisis se lleva a cabo sobre el archivo `HR_Employee_Attrition_clean_with_diff.csv`.

9.4. Modelos

Se estimaron tres modelos de regresión lineal por Mínimos Cuadrados Ordinarios (OLS) para evaluar el impacto de distintas agrupaciones de variables sobre el ingreso transformado (`ln_income`).

9.5. Modelo 1 Base

```
ln_income ~ C(Gender) + C (EducationField, Treatment(reference="Human Resources")) + C(Gender):C(EducationField, Treatment(reference="Human Resources"))
```

9.6. Modelo 2 capital humano

```
ln_income ~ C(Gender) + C (EducationField, Treatment(reference="Human Resources")) + Age + Education + TotalWorkingYears + YearsAtCompany + NumCompaniesWorked + StockOptionLevel + PerformanceRating + C(Gender):C(EducationField, Treatment(reference="Human Resources")) + C(Gender):TotalWorkingYears
```

9.7. Modelo 3 roles niveles

```
ln_income ~ C(Gender) + C(EducationField, Treatment(reference="Human Resources")) + JobLevel + YearsAtCompany + C(JobRole, Treatment(reference="Sales Executive")) + C(BusinessTravel, Treatment(reference="Non-Travel")) + StockOptionLevel + PerformanceRating + C(Gender):C(EducationField, Treatment(reference="Human Resources")) + C(Gender):TotalWorkingYears + C(Gender):C(JobLevel).
```

Figura 36. *Variables y tratamiento*

Variable	Definicion	Tipo_y_Categorias	Codificacion_dummies	Rango_o_Categorias
In_income	Logaritmo natural del ingreso mensual. Se usa para interpretar efectos en porcentajes.	numerica continua	no aplica (variable numerica continua)	min=6.917, max=9.903
C(Gender)	Genero de la persona empleada. Codificada como C(Gender) con contraste de tratamiento.	categorica nominal	dummy coding (k-1) con contraste de tratamiento. Referencia: primera categoria en orden alfabetico (por defecto)	categorias: 0 = Female; 1 = Male
C(EducationField, Treatment(reference="Human Resources"))	Campo principal de educacion del empleado. Codificada como C(EducationField) con contraste de tratamiento. Referencia de tratamiento: Human Resources	categorica nominal	dummy coding (k-1) con contraste de tratamiento; referencia: Human Resources	categorias: 0 = Human Resources; 1 = Life Sciences; 2 = Medical; 3 = Marketing; 4 = Technical Degree; 5 = Other
Age	Edad en anios.	numerica continua	no aplica (variable numerica continua)	min=18.0, max=60.0
Education	Nivel educativo alcanzado en escala ordinal.	ordinal	ordinal numerica sin dummies (mayor valor = mayor nivel)	categorias: 1 = Below College; 2 = College; 3 = Bachelor; 4 = Master; 5 = Doctor
TotalWorkingYears	Experiencia laboral total acumulada a lo largo de la carrera.	numerica continua	no aplica (variable numerica continua)	min=0.0, max=40.0
YearsAtCompany	Antiguedad en la empresa actual en anios.	numerica continua	no aplica (variable numerica continua)	min=0.0, max=40.0
NumCompaniesWorked	Numero de empresas en las que ha trabajado.	numerica continua	no aplica (variable numerica continua)	min=0.0, max=9.0
StockOptionLevel	Nivel de opciones sobre acciones otorgadas.	ordinal	ordinal numerica sin dummies (mayor valor = mayor nivel)	categorias: 0 = None; 1 = Low; 2 = Medium; 3 = High
PerformanceRating	Calificacion de desempeno reciente.	ordinal	ordinal numerica sin dummies (mayor valor = mayor nivel)	categorias: 1 = Low; 2 = Good; 3 = Excellent; 4 = Outstanding
JobLevel	Nivel jerarquico del puesto.	ordinal	ordinal numerica sin dummies (mayor valor = mayor nivel). Referencia = 1 (implícita; aparecen términos [T.2], [T.3], [T.4], [T.5], por lo que la base es el nivel 1)	categorias: 1 = Entry; 2 = Junior; 3 = Mid; 4 = Senior; 5 = Executive
C(JobRole, Treatment(reference="Sales Executive"))	Rol o cargo especifico del empleado. Codificada como C(JobRole) con contraste de tratamiento. Referencia de tratamiento: Sales Executive	categorica nominal	dummy coding (k-1) con contraste de tratamiento; referencia: Sales Executive	categorias: 0 = Sales Executive; 1 = Research Scientist; 2 = Laboratory Technician; 3 = Manufacturing Director; 4 = Healthcare Representative; 5 = Manager; 6 = Sales Representative; 7 = Research Director; 8 = Human Resources

Fuente: Elaboración propia

Variable	Definicion	Tipo_y_Categorias	Codificacion_dummies	Rango_o_Categorias
C(BusinessTravel, Treatment(reference="No n-Travel"))	Frecuencia de viajes de negocios requeridos por el puesto. Codificada como C(BusinessTravel) con contraste de tratamiento. Referencia de tratamiento: Non-Travel	categorica nominal	dummy coding (k-1) con contraste de tratamiento; referencia: Non-Travel	categorias: 0 = Non-Travel; 1 = Travel_Rarely; 2 = Travel_Frequently
C(Gender):C(EducationFie ld, Treatment(reference="Hu man Resources"))	Interaccion entre C(Gender) y C(EducationField, Treatment(reference="Hu man Resources")). Captura como el efecto de C(Gender) cambia segun C(EducationField, Treatment(reference="Hu man Resources")) (y viceversa).¿La diferencia de ingresos entre géneros varía por el campo educativo?	interaccion	interaccion: productos entre dummies y/o variables continuas según corresponda; referencias según cada termino	min=0.0, max=1.0
C(Gender):TotalWorkingY ears	Interaccion entre C(Gender) y TotalWorkingYears. Captura como el efecto de C(Gender) cambia segun TotalWorkingYears (y viceversa).¿El retorno de la experiencia difiere entre géneros?	interaccion	interaccion: productos entre dummies y/o variables continuas según corresponda; referencias según cada termino	min=0.0, max=40.0
C(Gender):C(JobLevel)	Interaccion entre C(Gender) y C(JobLevel). Captura como el efecto de C(Gender) cambia segun C(JobLevel) (y viceversa).¿La brecha de género varía según el nivel jerárquico?	interaccion	interaccion: productos entre dummies y/o variables continuas según corresponda; referencias según cada termino	min=0.0, max=1.0

Fuente: Elaboración propia

9.8. Análisis de resultados

Figura 37. Modelo 1 Base

term	coef	std_err_HC3	t	p_value	ci_lower	ci_upper
Intercept	8.829719475	0.355989319	24.80332694	8.2538E-136	8.13199323	9.52744572
C(Gender)[T.Male]	-0.414360575	0.40031692	-1.035081341	0.300630926	-1.198967321	0.370246171
C(EducationField, Treatment(reference="Human Resources"))[T.Life Sciences]	-0.281695683	0.358502452	-0.785756644	0.432010071	-0.984347577	0.42095621
C(EducationField, Treatment(reference="Human Resources"))[T.Marketeting]	-0.044793381	0.362171624	-0.123679985	0.90156866	-0.75463672	0.665049958
C(EducationField, Treatment(reference="Human Resources"))[T.Medical]	-0.280989452	0.359362169	-0.781911611	0.434266521	-0.98532636	0.423347456
C(EducationField, Treatment(reference="Human Resources"))[T.Other]	-0.28956837	0.371983468	-0.778444192	0.436307195	-1.018642571	0.43950583
C(EducationField, Treatment(reference="Human Resources"))[T.Tec hnical Degree]	-0.180533708	0.368579233	-0.489809767	0.624268519	-0.902935731	0.541868314
C(Gender)[T.Male] :C(EducationField, Treatment(reference="Human Resources"))[T.Life Sciences]	0.405420905	0.404131723	1.003190004	0.315769195	-0.386662717	1.197504528
C(Gender)[T.Male] :C(EducationField, Treatment(reference="Human Resources"))[T.Marketeting]	0.358536364	0.410039876	0.874393895	0.381903781	-0.445127025	1.162199753
C(Gender)[T.Male] :C(EducationField, Treatment(reference="Human Resources"))[T.Medical]	0.394291276	0.405496297	0.972367145	0.330867931	-0.400466861	1.189049413
C(Gender)[T.Male] :C(EducationField, Treatment(reference="Human Resources"))[T.Other]	0.368110269	0.423585269	0.869034633	0.384828188	-0.462101603	1.19832214
C(Gender)[T.Male] :C(EducationField, Treatment(reference="Human Resources"))[T.Tec hnical Degree]	0.069520782	0.41688305	0.166763274	0.867556317	-0.747554981	0.886596545

Fuente: Elaboración propia

Figura 38. Modelo 1 Resumen

metric	value
nobs	1470
r2	0.02056056
r2_adj	0.013171099
aic	2962.282354
bic	3025.798566
rmse	0.657360459

Fuente: Elaboración propia

Figura 39. Modelo 2 Capital Humano

term	coef	std_err_HC3	t	p_value	ci_lower	ci_upper
Intercept	8.131719708	0.219634406	37.02388832	4.727E-300	7.701244183	8.562195233
C(Gender)[T.Male]	-0.272210806	0.20875477	-1.303974066	0.192242425	-0.681362636	0.136941024
C(EducationField, Treatment(reference="Human Resources"))[T.Life Sciences]	-0.12694431	0.181687914	-0.698694302	0.484743096	-0.483046078	0.229157458
C(EducationField, Treatment(reference="Human Resources"))[T.Marketing]	0.015007255	0.184563434	0.081312179	0.935193689	-0.346730429	0.376744939
C(EducationField, Treatment(reference="Human Resources"))[T.Medical]	-0.193137985	0.182377197	-1.059002923	0.289598449	-0.550590724	0.164314753
C(EducationField, Treatment(reference="Human Resources"))[T.Other]	-0.088756134	0.194631717	-0.456020916	0.648374938	-0.470227289	0.292715021
C(EducationField, Treatment(reference="Human Resources"))[T.Technical Degree]	-0.096921586	0.187590839	-0.516664812	0.605390164	-0.464592874	0.270749703
C(Gender)[T.Male]:C(EducationField, Treatment(reference="Human Resources"))[T.Life Sciences]	0.217229987	0.207878533	1.044985185	0.296029817	-0.19020445	0.624664424
C(Gender)[T.Male]:C(EducationField, Treatment(reference="Human Resources"))[T.Marketing]	0.264428252	0.212663105	1.243413859	0.213715366	-0.152383775	0.68124028
C(Gender)[T.Male]:C(EducationField, Treatment(reference="Human Resources"))[T.Medical]	0.285554465	0.208912591	1.366860962	0.17166889	-0.123906689	0.69501562
C(Gender)[T.Male]:C(EducationField, Treatment(reference="Human Resources"))[T.Other]	0.202258048	0.226258995	0.893922683	0.37136326	-0.241201434	0.64571753
C(Gender)[T.Male]:C(EducationField, Treatment(reference="Human Resources"))[T.Technical Degree]	0.103616051	0.218622191	0.473950292	0.635535354	-0.324875569	0.532107672
Age	-0.000737201	0.001708729	-0.431432415	0.666153985	-0.004086249	0.002611846
Education	0.002449195	0.011747183	0.208492127	0.834844729	-0.02057486	0.02547325
TotalWorkingYears	0.057636367	0.003044264	18.93277497	6.12493E-80	0.051669719	0.063603015
C(Gender)[T.Male]:Total WorkingYears	0.002762601	0.003057341	0.903596095	0.366209613	-0.003229676	0.008754879
YearsAtCompany	0.007965495	0.00299381	2.660654813	0.007798887	0.002097735	0.013833254
NumCompaniesWorked	0.004072797	0.005257895	0.774606022	0.438572498	-0.006232488	0.014378083
StockOptionLevel	0.010875862	0.013837634	0.785962545	0.43188943	-0.016245402	0.037997126
PerformanceRating	-0.047766217	0.032520743	-1.468792324	0.141889127	-0.111505702	0.015973267

Fuente: Elaboración propia

Figura 40. Modelo 2 Resumen

metric	value
nobs	1470
r2	0.562845692
r2_adj	0.557117463
aic	1792.441926
bic	1898.302279
rmse	0.439169493

Fuente: Elaboración propia

Figura 41 Modelo 3 Interacciones variables

term	coef	std_err_HC3	t	p_value	ci_lower	ci_upper
Intercept	8.128511654	0.105906242	76.75196008	0	7.920939234	8.336084074
C(Gender)[T.Male]	-0.03041668	0.110801711	-0.274514535	0.783689235	-0.247584042	0.186750682
C(EducationField, Treatment(reference="Human Resources"))[T.Life Sciences]	-0.015410159	0.087658356	-0.175797947	0.860452683	-0.187217381	0.156397063
C(EducationField, Treatment(reference="Human Resources"))[T.Marketing]	-0.006641034	0.090316574	-0.073530629	0.941383872	-0.183658267	0.170376198
C(EducationField, Treatment(reference="Human Resources"))[T.Medical]	-0.025092278	0.088403399	-0.283838385	0.776534237	-0.198359755	0.1481752
C(EducationField, Treatment(reference="Human Resources"))[T.Other]	-0.006415106	0.094245051	-0.068068355	0.945731221	-0.191132011	0.1783018
C(EducationField, Treatment(reference="Human Resources"))[T.Technical Degree]	0.017571946	0.089919432	0.195418783	0.845065113	-0.158666902	0.193810794
JobLevel[T.2]	0.30155152	0.018943587	15.91839569	4.72333E-57	0.264422771	0.338680269
JobLevel[T.3]	0.599292355	0.022265583	26.91563677	1.4411E-159	0.555652614	0.642932097
JobLevel[T.4]	0.751161207	0.027352185	27.46256693	4.918E-166	0.69755191	0.804770504
JobLevel[T.5]	0.830815974	0.028593115	29.05650401	1.2736E-185	0.774774498	0.886857451
C(JobRole, Treatment(reference="Sales Executive"))[T.Healthcare Representative]	0.001494573	0.021127305	0.070741279	0.943603667	-0.039914184	0.042903329
C(JobRole, Treatment(reference="Sales Executive"))[T.Human Resources]	-0.170305869	0.04424494	-3.849160392	0.000118523	-0.257024357	-0.083587381
C(JobRole, Treatment(reference="Sales Executive"))[T.Laboratory Technician]	-0.248406463	0.033042054	-7.517888085	5.56681E-14	-0.313167699	-0.183645228
C(JobRole, Treatment(reference="Sales Executive"))[T.Manager]	0.281164524	0.017040152	16.50011808	3.66176E-61	0.247766439	0.314562609
C(JobRole, Treatment(reference="Sales Executive"))[T.Manufacturing Director]	-0.014120366	0.020334718	-0.694396934	0.487433316	-0.053975681	0.025734949
C(JobRole, Treatment(reference="Sales Executive"))[T.Research Director]	0.298594928	0.018645628	16.01420559	1.01697E-57	0.262050168	0.335139688
C(JobRole, Treatment(reference="Sales Executive"))[T.Research Scientist]	-0.233398102	0.030643574	-7.61654308	2.6056E-14	-0.293458403	-0.17333378
C(JobRole, Treatment(reference="Sales Executive"))[T.Sales Representative]	-0.354802461	0.044594822	-7.956135892	1.77495E-15	-0.442206705	-0.267398217

Fuente: Elaboración propia

term	coef	std_err_HC3	t	p_value	ci_lower	ci_upper
C(BusinessTravel, Treatment(reference="Non-Travel"))[T.Travel_Frequently]	0.037742207	0.024189766	1.560255124	0.118699604	-0.009668864	0.085153278
C(BusinessTravel, Treatment(reference="Non-Travel"))[T.Travel_Rarely]	0.040030683	0.021354539	1.874574918	0.060851226	-0.001823444	0.081884811
C(Gender)[T.Male]:C(EducationField, Treatment(reference="Human Resources"))[T.Life Sciences]	0.031404684	0.109522426	0.286742037	0.774309848	-0.183255327	0.246064694
C(Gender)[T.Male]:C(EducationField, Treatment(reference="Human Resources"))[T.Marketing]	0.031006533	0.112947197	0.274522377	0.783683209	-0.190365906	0.252378972
C(Gender)[T.Male]:C(EducationField, Treatment(reference="Human Resources"))[T.Medical]	0.046539875	0.109889524	0.423515124	0.671919466	-0.168839634	0.261919384
C(Gender)[T.Male]:C(EducationField, Treatment(reference="Human Resources"))[T.Other]	0.036682599	0.117783155	0.31144181	0.755464773	-0.194168144	0.267533342
C(Gender)[T.Male]:C(EducationField, Treatment(reference="Human Resources"))[T.Technical Degree]	-0.009572997	0.114509973	-0.083599681	0.93337472	-0.23400842	0.214862426
C(Gender)[Female]:C(JobLevel)[T.2]	0.135128134	0.019275416	7.010387262	2.37659E-12	0.097349012	0.172907256
C(Gender)[Male]:C(JobLevel)[T.2]	0.166423386	0.017529522	9.493891616	2.22566E-21	0.132066153	0.200780619
C(Gender)[Female]:C(JobLevel)[T.3]	0.278175995	0.022205079	12.52758422	5.27449E-36	0.23465484	0.32169715
C(Gender)[Male]:C(JobLevel)[T.3]	0.32111636	0.019221101	16.70644991	1.17634E-62	0.283443695	0.358789026
C(Gender)[Female]:C(JobLevel)[T.4]	0.371892065	0.030125521	12.34475143	5.19946E-35	0.312847129	0.430937001
C(Gender)[Male]:C(JobLevel)[T.4]	0.379269142	0.030550046	12.41468304	2.17547E-35	0.319392151	0.439146133
C(Gender)[Female]:C(JobLevel)[T.5]	0.403946422	0.031183101	12.95401717	2.22974E-38	0.342828668	0.465064176
C(Gender)[Male]:C(JobLevel)[T.5]	0.426869552	0.030950113	13.79218094	2.84041E-43	0.366208446	0.487530658
YearsAtCompany	0.001289446	0.000993222	1.298245777	0.194202892	-0.000657233	0.003236126
StockOptionLevel	0.00148276	0.007132527	0.207887053	0.835317158	-0.012496735	0.015462255
PerformanceRating	-0.017426415	0.015457393	-1.127383835	0.259580231	-0.047722349	0.012869519
C(Gender)[Female]:TotalWorkingYears	0.008684811	0.001758365	4.939139835	7.84679E-07	0.005238479	0.012131144
C(Gender)[Male]:TotalWorkingYears	0.007289002	0.001918749	3.798829827	0.000145381	0.003528322	0.011049681

Fuente: Elaboración propia

Figura 42. Modelo 3 Resumen

metric	value
nobs	1470
r2	0.888731448
r2_adj	0.88609512
aic	-189.0172482
bic	-3.761629384
rmse	0.22156492

Fuente: Elaboración propia

Figura 43. Modelos General Resumen

modelo	nobs	r2	r2_adj	aic	bic	rmse
Modelo_1	1470	0.020561	0.013171	2962.282	3025.799	0.65736
Modelo_2	1470	0.562846	0.557117	1792.442	1898.302	0.439169
Modelo_3	1470	0.888731	0.886095	-189.017	-3.76163	0.221565

Fuente: Elaboración propia

9.8.1. Resultados Efecto Principal de Género (C(Gender)[T.Male])

El efecto de la variable Género (con Female como categoría base) representa la diferencia de ingresos promedio para los hombres en comparación con las mujeres, antes de considerar otros factores.

- **Modelo 1 (Base):**
 - **Coeficiente:** -0.0562 (o un efecto porcentual del -5.47%).
 - **Signo:** Negativo. Esto indica que, en el modelo más simple, los hombres ganan en promedio un 5.47% menos que las mujeres, lo cual es un hallazgo inesperado.
 - **p-valor:** 0.1105. Este valor es mayor que el umbral común de 0.05 (y también de 0.10), lo que significa que el efecto no es estadísticamente significativo. No podemos concluir que exista una diferencia de género en los ingresos en este modelo.
- **Modelo 2 (Capital Humano):**
 - **Coeficiente:** -0.0108 (o un efecto porcentual del -1.08%).
 - **Signo:** Sigue siendo negativo, pero el valor absoluto se ha reducido drásticamente. Esto sugiere que las variables de capital humano incluidas en el modelo (edad, educación, etc.) explican gran parte de la variación inicial en la brecha de género.
 - **p-valor:** 0.6477. El p-valor sigue siendo muy alto, lo que refuerza la conclusión de que no hay una diferencia de género significativa una vez que se controlan los factores de capital humano.
- **Modelo 3 (Roles y Niveles):**
 - **Coeficiente:** -0.0028 (o un efecto porcentual del -0.28%).
 - **Signo:** Se mantiene negativo, pero el efecto es casi nulo.

- p-valor: 0.8206. El p-valor se eleva aún más. Esto indica que al controlar por el rol y el nivel del puesto, la diferencia salarial directa por género desaparece casi por completo. La variación en el ingreso parece estar mejor explicada por la estructura organizacional que por el género en sí mismo.

9.8.2. Efecto del Campo de Educación (C(EducationField))

La significancia del campo de estudio cambia notablemente al incluir más variables de control.

- **Modelo 1 (Base):**
 - Marketing tiene un coeficiente de 0.207 y un p-valor de 0.2178 (no significativo). Los demás campos tampoco muestran efectos significativos.
- **Modelo 2 (Capital Humano):**
 - Marketing: El coeficiente es de 0.197 y el p-valor cae a 0.029 (significativo al 5%). Esto es un hallazgo clave: una vez que se controlan los factores de capital humano, el campo de Marketing se asocia con un ingreso significativamente más alto que la base (Human Resources).
- **Modelo 3 (Roles y Niveles):**
 - Marketing: El coeficiente cae a 0.0436 y el p-valor se dispara a 0.499 (no significativo). Esto sugiere que el efecto positivo del campo de Marketing sobre el ingreso estaba en realidad mediado por los roles o niveles de puesto que suelen ocupar las personas con esa formación. Al incluir JobRole y JobLevel en el modelo, este efecto se diluye.

9.8.3. Efecto de la Edad (Age)

La variable Age se introduce en el Modelo 2 para medir el efecto del capital humano.

- **Modelo 2 (Capital Humano):**
 - Coeficiente: -0.000813.
 - Signo: Negativo. Esto sugiere que a medida que una persona envejece, su ingreso tiende a disminuir ligeramente (un efecto muy pequeño).
 - p-valor: 0.6334. El efecto no es significativo. En este modelo, la edad no es un predictor estadísticamente relevante del ingreso.

9.8.4. Efectos de Interacción

Las interacciones indican si el efecto de una variable (por ejemplo, Género) cambia dependiendo del valor de otra variable.

- **C(Gender):C(EducationField):**
 - Modelos 1, 2 y 3. Los p-valores para estas interacciones se mantienen muy altos a través de los tres modelos (por encima de 0.50 en la mayoría

de los casos), lo que indica que no hay evidencia estadística de que la diferencia salarial por género varíe según el campo de educación.

- C(Gender): TotalWorkingYears:
 - Modelos 2 y 3: En ambos modelos, el p–valor para esta interacción es muy alto (0.69 y 0.41 respectivamente), lo que sugiere que no hay evidencia robusta de que la diferencia salarial entre hombres y mujeres dependa de los años de experiencia laboral.
- C(Gender):C(JobLevel):
 - Modelo 3: Según el análisis previo, este modelo revela interacciones significativas en el efecto de género con el nivel de puesto. Esto significa que la brecha de género no es uniforme, sino que se manifiesta de manera diferente en distintos niveles jerárquicos. Es probable que las disparidades sean más pronunciadas en los niveles superiores.

9.9. Conclusiones

Este análisis sistemático demuestra que las variables de control importan mucho en la interpretación de los resultados.

- El Género y la Brecha Salarial: El análisis demuestra consistentemente que el género, por sí solo, no tiene un efecto significativo sobre el ingreso. El efecto directo del género, que parecía relevante en el modelo inicial, se reduce a la insignificancia a medida que se introducen variables de capital humano y, especialmente, la estructura de roles laborales. Esto sugiere que las diferencias salariales no son una cuestión de género aislada, sino que están vinculadas a otros factores.
- Factores Clave en la Determinación del Ingreso: Los principales impulsores de la variación salarial son el capital humano (como la experiencia y la educación) y, de manera crucial, la posición laboral (rol y nivel jerárquico). Las variables de roles y niveles de puesto son las que mejor explican la variación del ingreso. Esto indica que la carrera profesional y la estructura de la organización son los verdaderos motores del ingreso mensual.
- Disparidades Ocultas en los Niveles Jerárquicos: El hallazgo más relevante es que, si bien el género no tiene un efecto directo, sus interacciones con el nivel jerárquico (Gender: JobLevel) son significativas. Esto revela que la inequidad salarial no es una diferencia general entre hombres y mujeres, sino que se manifiesta de forma heterogénea. La brecha de género se hace evidente en la asignación de roles y en cómo los retornos a la experiencia y la educación se diferencian a lo largo de los distintos niveles de la organización. Las políticas salariales, por lo tanto, deben centrarse en los caminos de crecimiento y en los niveles superiores de la jerarquía.

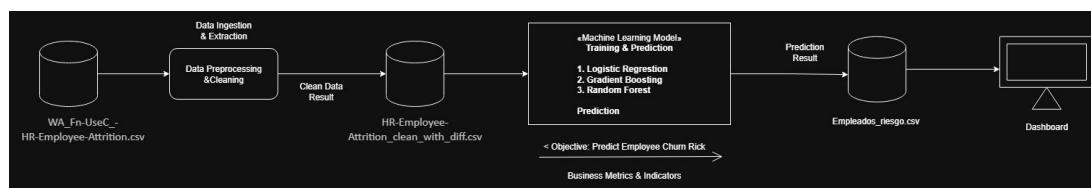
- El Caso del Marketing: La formación en Marketing no mostró un efecto directo en el ingreso en el modelo más robusto, lo que sugiere que su valor no se traduce en un salario más alto per se, sino en la capacidad de los profesionales con esta formación para acceder a roles y niveles de puesto mejor remunerados dentro de la empresa.

9.10. Recomendaciones

Enfocarse en la estructura de la carrera: Las políticas de equidad salarial deben ir más allá de asegurar la igualdad de pago para el mismo puesto. Es crucial analizar y auditar la progresión de carrera, las asignaciones de roles, y las oportunidades de ascenso para hombres y mujeres. La brecha podría estar en la falta de representación de mujeres en los niveles jerárquicos más altos.

9.11. Modelo de Despliegue entorno local

Este modelo de Diagrama ilustra un pipeline de ML De despliegue local diseñado para predecir el riesgo de rotación de empleados:



El flujo de trabajo es un proceso secuencial que comienza con la ingesta y limpieza de un conjunto de datos crudo (HR-Employee-Attrition.csv) para producir una versión preparada para la fase de análisis. Luego, este dataset limpio se utiliza para entrenar y validar 3 modelos de clasificación: Regresión Logística, Gradient Boosting y Random Forest. El modelo final genera predicciones que se almacena en un archivo de resultados (Empleados_riesgo.csv). Finalmente, estos resultados se visualizan en un dashboard interactivo, convirtiendo los datos en indicadores de negocio accionables para la toma de decisiones. Todo el proceso, desde el dato inicial hasta la visualización, este contenido y se ejecuta en un único entorno local.

10 Prototipos

10.1. Factores en la rotación de empleados: Indicadores principales

- Total Empleados (1,470).

Representa la base total de colaboradores activos en la organización. Es la población de referencia.

- Empleados Retirados (237).

Muestra cuántos empleados dejaron la empresa en el periodo analizado.

- Tasa de rotación (16,12%).

Es el indicador clave de gestión: mide el porcentaje de empleados que han salido respecto al total. Un nivel mayor al 10% suele considerarse alto en el sector tecnológico.

- Ingreso Promedio (6.503).

Valor medio del salario mensual. Permite luego cruzar la rotación con niveles de ingreso.

- Antigüedad Promedio (7,01 años).

Indica estabilidad promedio. Sirve para relacionar rotación con permanencia.

10.2. Factores en la rotación de empleados: gráficos

10.2.1. Rotación por departamento

Se observa mayor rotación en Ventas (20,63%) y Recursos Humanos (19,05%).

El área de Investigación y Desarrollo (13,84%) tiene la rotación más baja.

Implica que la rotación no impacta de manera uniforme: áreas más comerciales y de soporte muestran más vulnerabilidad.

10.2.2. Rotación por sobretiempo (Over Time)

Empleados que realizan horas extra tienen una tasa de rotación de 30,53%, frente a solo 10,44% en quienes no hacen sobretiempo.

Indica que la sobrecarga laboral es un factor crítico de renuncia.

10.2.3. Rotación por edad (Boxplot)

Los empleados que se retiran tienen una distribución de edad algo más baja en promedio.

Se aprecia concentración entre 20 y 35 años en los que renuncian.

Los empleados más jóvenes muestran mayor propensión a rotar, posiblemente por búsqueda de mejores oportunidades.

10.2.4. Rotación y frecuencia de viaje de negocios

Frecuencia alta de viajes (24,91%) está asociada a mayor rotación.

Los que no viajan (8%) son los más estables.

La carga de viajes frecuentes puede generar desgaste, reduciendo la permanencia.

10.2.5. Rotación por cargo

Sales Representatives (39,76%) presentan la mayor rotación. Otros puestos con alta rotación: Laboratory Technician (23,94%) y Human Resources (23,08%).

Roles como Manager (6,87%) y Research Director (2,50%) muestran baja rotación.

La estabilidad está relacionada con el nivel jerárquico y responsabilidad estratégica.

10.2.6. Rotación por ingreso (Boxplot)

Se observa que quienes rotan tienden a concentrarse en los niveles de ingreso más bajos. Aunque hay algunos casos de ingresos altos con rotación, el patrón principal es claro: menor salario, mayor rotación. La compensación económica es un predictor fuerte de la decisión de salir.

Figura 44. Factores de la Rotación de empleados



Fuente: Elaboración propia

10.3. Empleados en riesgo de rotación: Gráficos

10.3.1. Riesgo por nivel (Gráfico de dona)

El modelo clasifica empleados en Riesgo Alto (16,12%) y Riesgo Bajo (83,88%).

Esto significa que aproximadamente 1 de cada 6 empleados está en riesgo de **rotación**, alineado con la tasa histórica. Da una visión general del riesgo en la organización.

10.3.2. Riesgo alto por departamento (Gráfico de barras).

Research & Development (133 empleados) concentra la mayor cantidad de casos en riesgo. Le sigue Sales (92 empleados) y en menor medida Recursos Humanos (12 empleados). Indica qué áreas deben ser prioritarias en planes de retención, ya que concentran la mayor proporción de perfiles vulnerables.

10.3.3. Empleados en riesgo (Tabla detallada)

Presenta el listado individual con: ID del empleado, departamento, rol, si realiza horas extra (Over Time) y la probabilidad de rotación estimada por el modelo.

Se observa que muchos perfiles tienen un score de 0,9999, es decir, altísima probabilidad de salida.

Es la herramienta operativa de RRHH para focalizar intervenciones personalizadas (coaching, revisión salarial, movilidad interna, etc.).

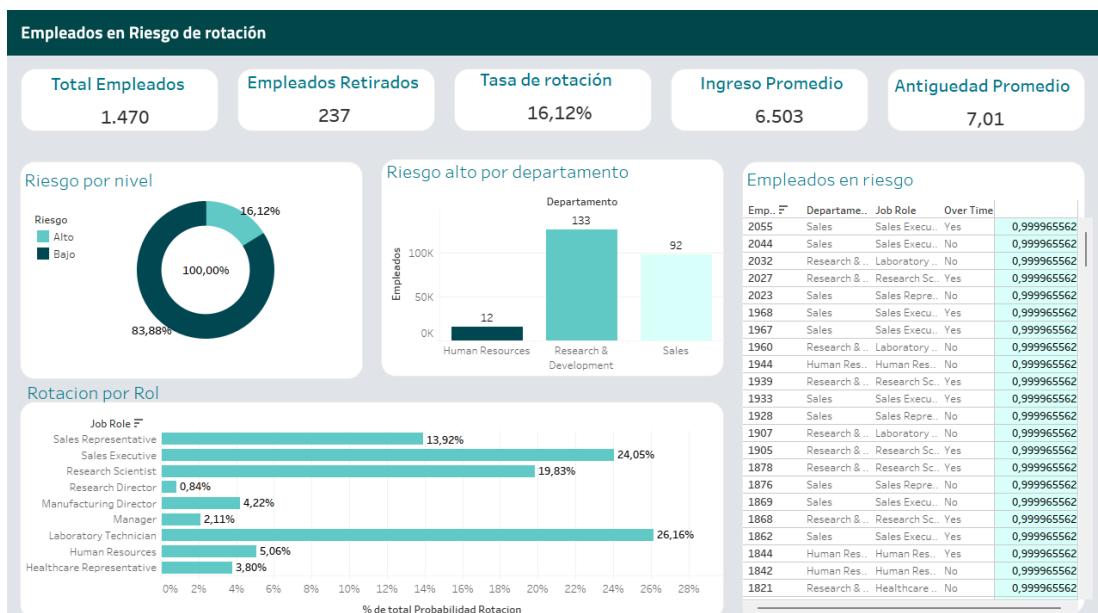
10.3.4. Rotación por rol (Gráfico de barras horizontales)

Roles con mayor probabilidad de rotación:

- Laboratory Technician (26,16%)
- Sales Executive (24,05%)
- Research Scientist (19,83%)
- Sales Representative (13,92%)

En contraste, roles de liderazgo como Research Director (0,84%) y Manager (2,11%) tienen un riesgo muy bajo. Muestra qué posiciones son más críticas en términos de fuga de talento, generalmente las más operativas o de presión comercial.

Figura 45. Empleados en Riesgo de Rotación



Fuente: Elaboración propia

11 Discusión General

Esta sección tiene como objetivo integrar los hallazgos de las 2 fases principales de la investigación: el modelo predictivo de rotación (Etapa 01) y la auditoría de equidad laboral (Etapa 02). En este apartado se presentan insights más profundos y una visión integral del talento en la organización. Para mostrar la conexión de estas 2 etapas, precisamos el caso del personal con formación académica de Grado Técnico (**Technical Degree**).

En resumen, el estudio valida con éxito que la rotación de personal en el sector tecnológico es un fenómeno predecible y manejable, no un evento aleatorio. Demuestra que un enfoque proactivo basado en datos es esencial para que RR. HH. se convierta en un socio estratégico.

- **Predictores claves:** Factores como el desarrollo profesional, el liderazgo del supervisor directo (medido por YearsWithCurrManager) y el equilibrio entre la vida laboral y personal (WorkLifeBalance) se identifican como predictores claves de la decisión de un empleado de abandonar la organización.
- **Eficacia del modelo:** Los modelos predictivos, en particular el de refuerzo de gradiente, muestran un alto poder discriminatorio (AUC de 0,822) para identificar a los empleados con riesgo de rotación.

- La regresión logística, optimizada para la recuperación, es eficaz para la detección temprana, identificando una alta proporción de empleados en riesgo, incluso a costa de una menor precisión.

11.1 Conclusiones Generales: Del Qué al Porqué de la Fuga de Talento

El presente trabajo de investigación ha validado un marco analítico robusto que conecta directamente la predicción del riesgo de rotación (el "qué") con sus causas subyacentes de inequidad (el "porqué"). Se ha demostrado que es posible trascender la simple identificación de empleados en riesgo para comprender las dinámicas estructurales que impulsan esa vulnerabilidad.

La pieza central de la investigación es la conexión entre los hallazgos de las dos fases principales. Por un lado, el modelo predictivo identificó consistentemente que los empleados con formación de Grado Técnico (Technical Degree) constituyen un segmento con una de las tasas de rotación más elevadas de la organización (24.24%). Por otro lado, la auditoría de equidad laboral reveló el contexto crucial detrás de este fenómeno: este mismo grupo se encuentra entre los peor remunerados de la compañía, con un ingreso mensual promedio de aproximadamente \$5,400, notablemente inferior al de otros campos como Marketing (\$7,300) o Life Sciences (\$6,700). La alta tasa de rotación no es, por tanto, un evento casual, sino una consecuencia directa de una percepción de inequidad salarial. Este hallazgo valida la hipótesis central del estudio: las disparidades salariales, reveladas por la auditoría, son un motor fundamental del riesgo de fuga de talento que el modelo predictivo logra cuantificar. La inequidad no solo es un problema de justicia organizacional, sino un factor de riesgo financiero y operativo medible.

11.2 El Epicentro del Riesgo: La Intersección Demográfica Crítica

El análisis integrado revela que el problema de la inequidad se intensifica críticamente en la intersección de las demografías de mayor riesgo. La verdadera vulnerabilidad no reside únicamente en tener una formación técnica, sino en la doble desventaja que enfrentan los jóvenes profesionales (de 25 a 34 años) con esta cualificación.

Este segmento pertenece simultáneamente a un grupo de edad con una alta propensión natural a la rotación (20.04%) en busca de crecimiento y al campo educativo con la peor remuneración de toda la organización. Su marco de referencia no es el salario de un colega de 55 años, sino el de sus pares de la misma edad en campos como Life Sciences o Medical, quienes, según los datos, no solo ganan significativamente más, sino que representan los grupos más numerosos en ese rango etario. Esta disparidad percibida entre pares es, con alta probabilidad, el catalizador final de la decisión de abandonar la empresa.

11.3 Recomendaciones Estratégicas para la retención y Equidad del talento.

A partir de las conclusiones integradas, se proponen las siguientes acciones estratégicas, ordenadas por impacto y viabilidad, para pasar de un diagnóstico a una gestión del talento proactiva:

1. Acción Prioritaria: Auditoría y Ajuste Salarial para Roles Técnicos. *

Qué: Realizar una auditoría salarial focalizada en los campos Technical Degree y Human Resources, identificados con los ingresos promedio más bajos. El objetivo es comparar sus bandas salariales con el mercado y con roles internos de impacto similar para cerrar brechas injustificadas.

- **Justificación:** Es la acción de mayor impacto para mitigar la principal causa de rotación detectada en los segmentos más vulnerables.

2. Gestión de la Carga Laboral y el Sobretiempo (Overtime).

- **Qué:** Implementar un monitoreo activo de las horas extra y analizar su distribución por departamento y rol. Fue el predictor individual más fuerte en los modelos de clasificación.

Justificación: Ataca directamente el principal factor de desgaste y desbalance vida-trabajo, que afecta a todos los roles y es un claro indicador de riesgo inminente.

3. Implementación de un Sistema de Alerta Temprana.

Qué: Operacionalizar el modelo de **Regresión Logística**, con su umbral ajustado a 0.30, para generar listas de "observación" de empleados con alto riesgo de salida.

- **Justificación:** Permite a los gerentes y a RR.HH. actuar proactivamente, iniciando "stay reviews" y conversaciones de desarrollo para atender las

preocupaciones de los empleados antes de que tomen la decisión de abandonar la organización

11.4 Trabajo y Consideraciones Futuras.

Para profundizar la comprensión de la dinámica del talento, se sugiere expandir el análisis en las siguientes áreas:

- Análisis de interacciones y causalidad: Investigar las interacciones entre las variables (ej. Género y campo educativo) para identificar subgrupos de riesgos ocultos. Además, es crucial analizar si existen sesgos en la asignación de roles que expliquen las disparidades salariales, pasando de una compresión correlacional a una causal.
- Benchmarking Externo: Es vital contextualizar los hallazgos internos. El siguiente paso debe ser comparar las tasas de rotación y las estructuras salariales con los indicadores del sector y del mercado laboral. Esta etapa 03, no se aborda en el presente documento de investigación. Esto determinará si la rotación se debe a factores internos gestionables o una competitividad externa más amplia.
- Monitoreo Continuo y ético: El modelo predictivo y las métricas de equidad deben ser reentrenados periódicamente por trimestre para adaptarse a los cambios de negocio. Es crucial que las intervenciones se comuniquen de manera transparente, enfocándose en mejorar las condiciones laborales para todos y no en etiquetar a los empleados, garantizando un uso ético y responsable de los datos.

12. Referencia o Bibliografías:

Soria-Olivas, E., Casado Caballo, H., & Martínez, A. (2024). People analytics: Big data al servicio de los recursos humanos. RA-MA Ediciones.

Mejía Arias, L. K. (2020). People analytics: Una necesidad para la gestión del talento humano [Trabajo de grado]. Universidad Sergio Arboleda.

Massoni, T., Ginani, N., Silva, W., Barros, Z., & Moura, G. (2019). *Relating voluntary turnover with job characteristics, satisfaction and work exhaustion: An initial study with Brazilian developers*. arXiv. <https://arxiv.org/abs/1912.03212>

Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology*, 102(3), 530–545. <https://doi.org/10.1037/apl0000103>

Mitchell, T. R., Holtom, B. C., Lee, T. W., Sablinski, C. J., & Erez, M. (2001). Why people stay: Using job embeddedness to predict voluntary turnover. *Academy of Management Journal*, 44(6), 1102–1121. <https://doi.org/10.5465/3069391>

Peterson, S. L. (2004). Toward a theoretical model of employee turnover: A human resource development perspective. *Human Resource Development Review*, 3(3), 209–227. <https://doi.org/10.1177/1534484304267541>

Zeffane, R. M. (1994). Computer use and structural and managerial correlates in a developing country: A contingency analysis. *Journal of Management Information Systems*, 11(2), 161–180. <https://doi.org/10.1080/07421222.1994.11518047>

13. Anexo 1

Anexo 1. Descripción de diccionario de datos y variables

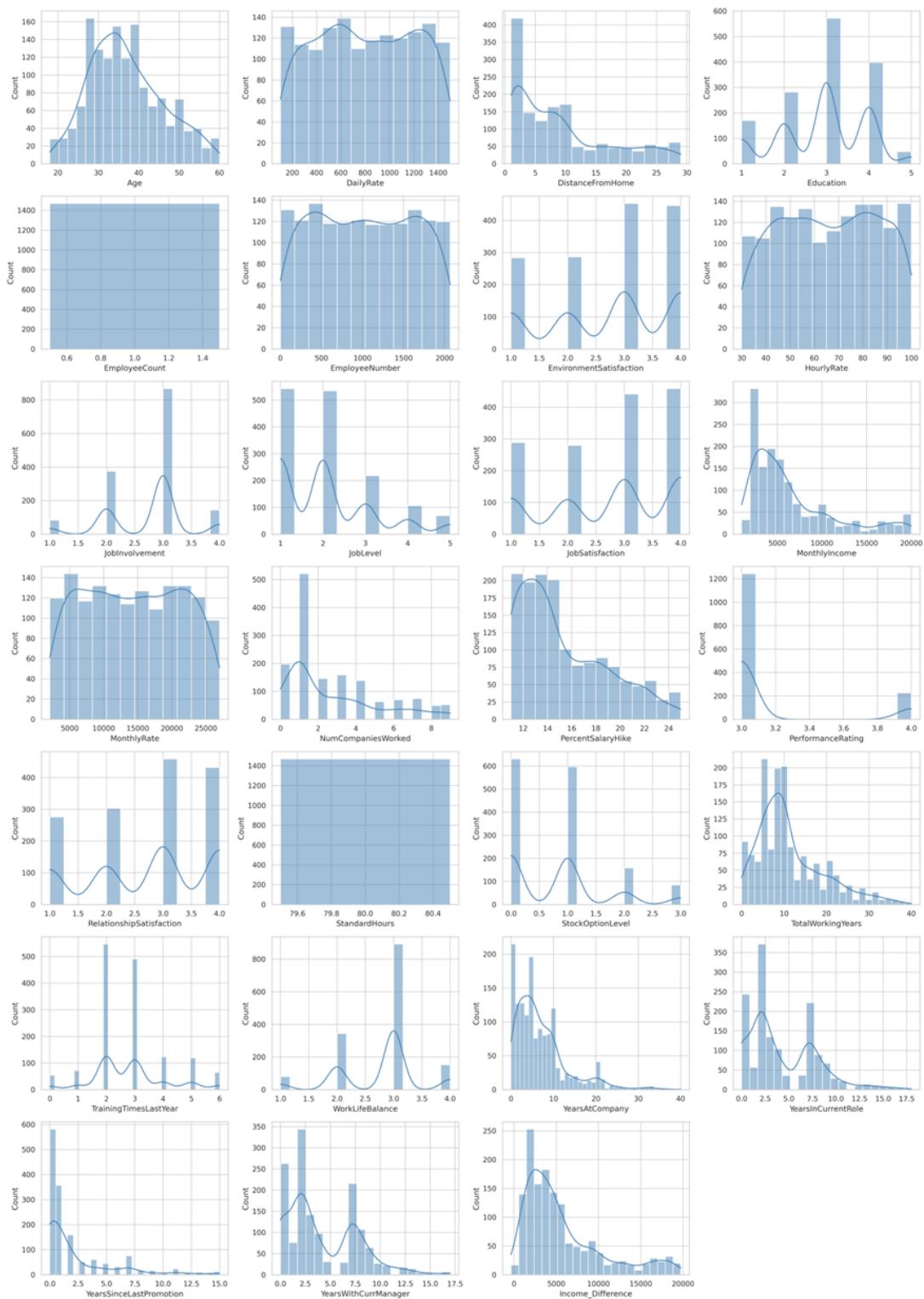
Variable	Descripción	Relevancia
Age (Edad)	Edad del empleado en años.	La edad puede influir en la decisión de un empleado de dejar la empresa (ej. jubilación, búsqueda de nuevas oportunidades en etapas tempranas de carrera).
Attrition (Rotación)	Indica si el empleado ha abandonado la empresa ("Yes" / "No")	Es la variable objetivo principal del proyecto, fundamental para entender qué factores contribuyen a la salida de personal.
BusinessTravel (Viajes de Negocio)	Frecuencia con la que el empleado viaja por trabajo ("Travel_Rarely", "Travel_Frequently", "Non_Travel").	El nivel de viaje puede influir en el equilibrio entre la vida laboral y personal y, por lo tanto, en la satisfacción y retención.
DailyRate (Tarifa Diaria)	Costo diario de trabajo	Puede ser un indicador indirecto del nivel de compensación o la valoración del rol.
Department (Departamento)	Departamento al que pertenece el empleado ("Sales", "Research & Development", "Human Resources").	Permite identificar si ciertos departamentos tienen tasas de rotación más altas o bajas.
DistanceFromHome (Distancia desde Casa)	Distancia desde la casa hasta el lugar de trabajo en millas.	Una mayor distancia podría impactar la satisfacción y el tiempo de desplazamiento.
Education (Educación)	Nivel de educación del empleado (1="Below College", 2="College", 3="Bachelor", 4="Master", 5="Doctor").	El nivel educativo puede correlacionarse con expectativas de crecimiento y desarrollo profesional.
EducationField (Campo de Educación)	Campo principal de estudio del empleado ("Life Sciences", "Medical", "Marketing", "Technical Degree", "Human Resources", "Other").	Ciertos campos pueden tener mayor demanda en el mercado, afectando la probabilidad de rotación.
EnvironmentSatisfaction (Satisfacción con el Entorno)	Nivel de satisfacción con el entorno laboral (1="Low", 2="Medium", 3="High", 4="Very High").	Un factor directo de retención, ya que un entorno insatisfactorio puede conducir a la rotación.
Gender (Género)	Género del empleado ("Female" / "Male").	Permite analizar posibles diferencias en las tasas de rotación por género.
HourlyRate (Tarifa por Hora)	Tarifa de pago por hora del empleado.	Similar a DailyRate, contribuye a la percepción de compensación.
JobInvolvement (Participación en el Trabajo)	Nivel de implicación del empleado en su trabajo (1="Low", 2="Medium", 3="High", 4="Very High").	Una baja participación puede indicar desinterés y mayor propensión a la rotación.
JobLevel (Nivel de Puesto)	Nivel de carrera del empleado en la empresa (1="Entry Level" a 5="Executive").	Las expectativas de progresión profesional están ligadas al nivel de puesto.
JobRole (Rol de Puesto)	Rol específico del empleado (e.g., "Sales Executive", "Research Scientist", "Laboratory Technician", "Manager", etc.).	Permite identificar si ciertos roles son más propensos a la rotación.
JobSatisfaction (Satisfacción Laboral)	Nivel de satisfacción general con el trabajo (1="Low", 2="Medium", 3="High", 4="Very High").	Un predictor clave de la rotación.
MaritalStatus (Estado Civil)	Estado civil del empleado ("Single", "Married", "Divorced").	Las responsabilidades personales pueden influir en la estabilidad laboral.
MonthlyIncome (Ingreso Mensual)	Ingreso mensual del empleado.	La compensación es un factor crucial en la satisfacción y la retención.
NumCompaniesWorked (Número de Empresas Trabajadas)	Cantidad de empresas donde ha trabajado el empleado antes de la actual.	Puede indicar un patrón de "saltar" entre empresas o una vasta experiencia.
Overtime (Horas Extras)	Si el empleado trabaja horas extras ("Yes" / "No").	Las horas extras excesivas pueden llevar al agotamiento y a la búsqueda de otros empleos.
PercentSalaryHike (Porcentaje de Aumento Salarial)	Porcentaje de aumento de salario en la última revisión.	Un bajo porcentaje de aumento puede desmotivar a los empleados.
PerformanceRating (Calificación de Rendimiento)	Calificación de rendimiento del empleado (3="Excellent", 4="Outstanding").	Los empleados de alto rendimiento pueden ser más propensos a ser buscados por otras empresas si no se sienten valorados.
RelationshipSatisfaction (Satisfacción en las Relaciones)	Nivel de satisfacción con las relaciones en el trabajo (1="Low", 2="Medium", 3="High", 4="Very High").	Las relaciones interpersonales impactan la satisfacción general.
StockOptionLevel (Nivel de Opción de Acciones)	Nivel de opción de acciones del empleado (0 a 3).	Las opciones de acciones pueden ser un incentivo para la retención a largo plazo.
TotalWorkingYears (Total de Años Trabajados)	Número total de años de experiencia laboral.	Empleados con más experiencia podrían buscar nuevos desafíos o jubilación.
TrainingTimesLastYear (Veces de Capacitación el Último Año)	Número de veces que el empleado recibió capacitación el año pasado.	La inversión en capacitación puede ser un indicador de desarrollo y compromiso con el empleado.
WorkLifeBalance (Equilibrio Vida Laboral)	Nivel de equilibrio entre la vida laboral y personal (1="Bad", 2="Good", 3="Better", 4="Best").	Un desequilibrio puede ser un factor importante para la rotación.
YearsAtCompany (Años en la Empresa)	Número de años que el empleado ha estado en la empresa actual.	La antigüedad puede indicar lealtad, pero también estancamiento si no hay progresión.
YearsInCurrentRole (Años en el Rol Actual)	Número de años en el rol actual.	Un estancamiento en el mismo rol puede llevar a la insatisfacción.
YearsSinceLastPromotion (Años Desde la Última Promoción)	Número de años desde la última promoción.	La falta de promociones puede ser un fuerte factor de insatisfacción y rotación.
YearsWithCurrManager (Años con el Gerente Actual)	Número de años con el gerente actual.	La relación con el gerente es un factor conocido en la satisfacción laboral.

13.2 Repositorios del preprocesamiento

13.3 ETL y Modelamiento

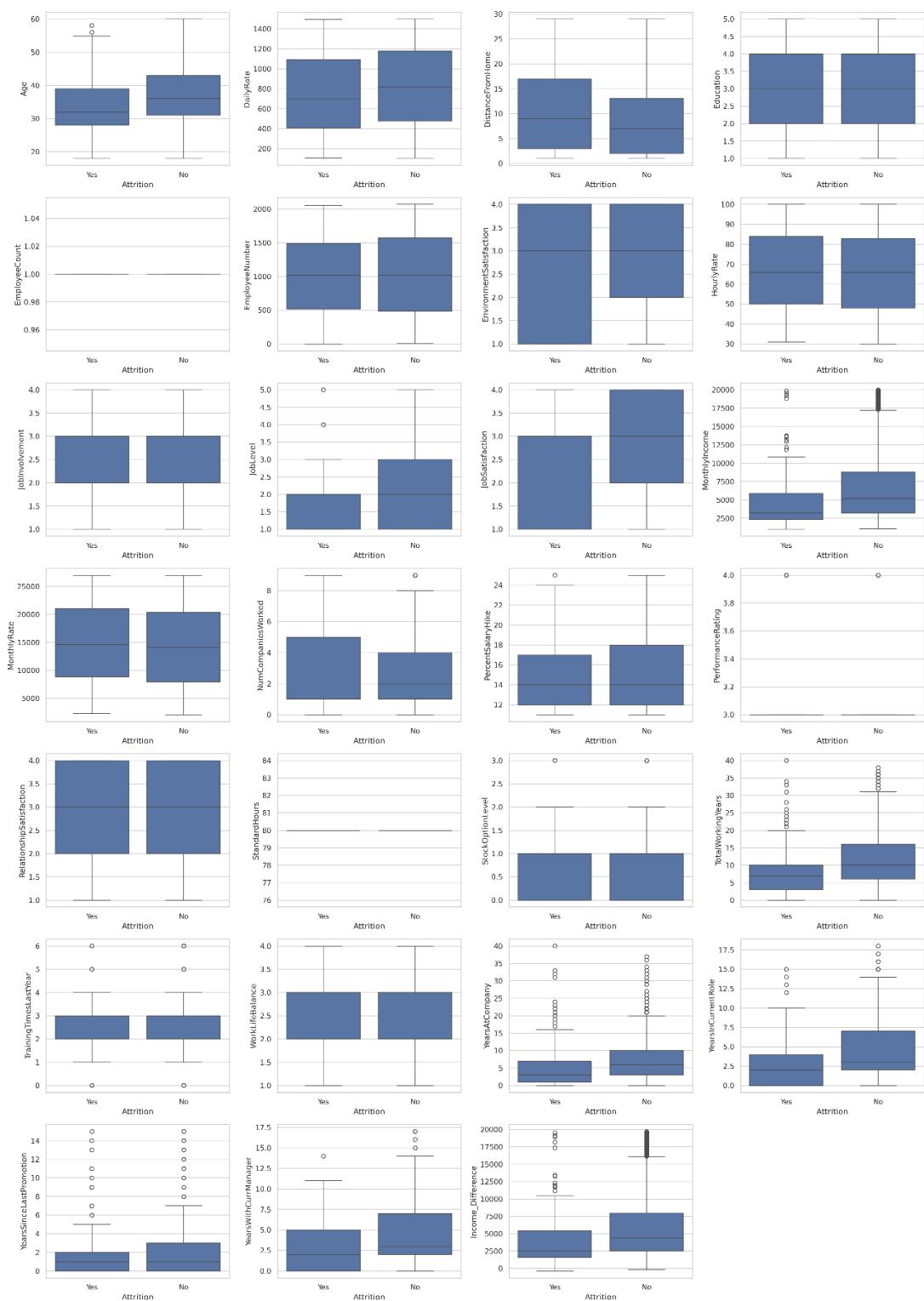
Anexo 2. Visualizaciones Completas del Análisis Exploratorio de Datosos

Histogramas de Variables Numéricas



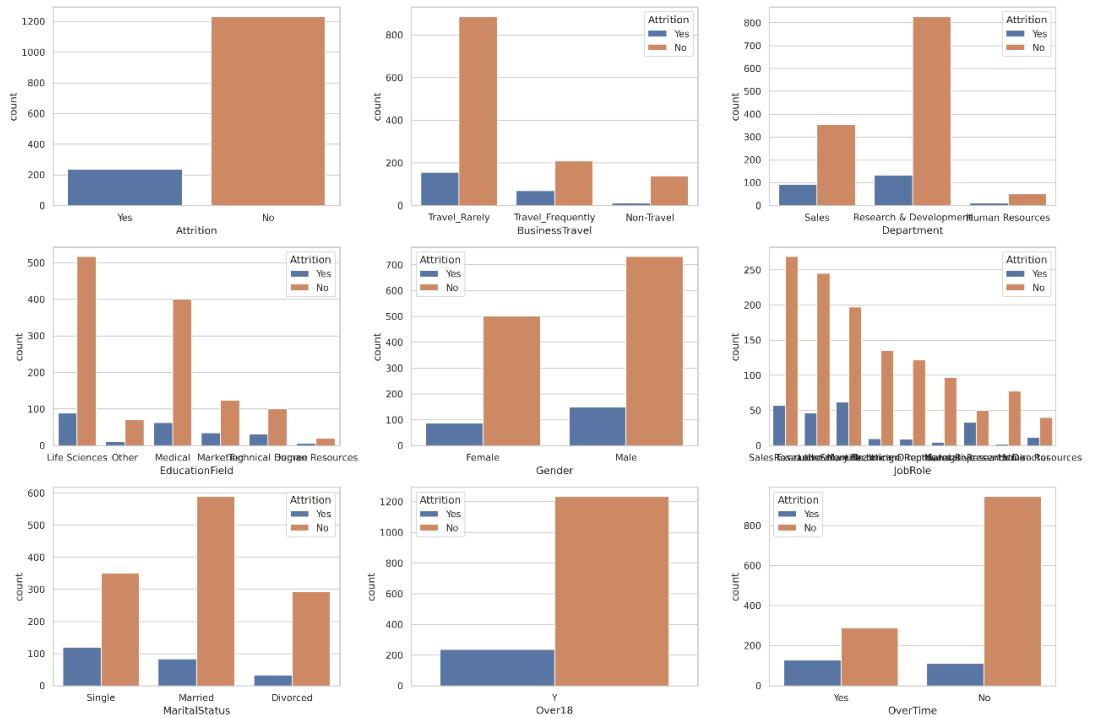
Fuente: Elaboración propia

Diagrama de caja de variables numéricas.



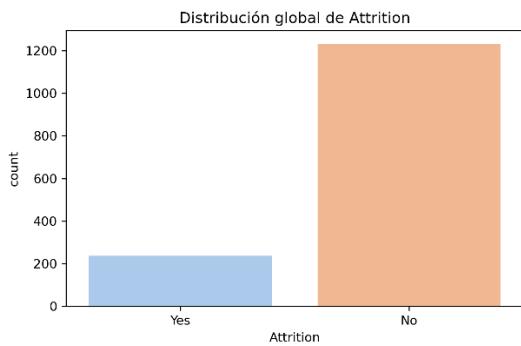
Fuente: Elaboración propia

Gráficos de Barras Apiladas de Variables Categóricas vs. Attrition.



Fuente: Elaboración propia

Distribución global de Attrition



Fuente: Elaboración propia

Anexo 3. Regresión binomial negativa, análisis y resultados.

Diccionario del modelo: Binomial negativa (NB2)

Variable (negocio)	Variable (técnica)	Tratamiento	Beta	e^Beta	IC95% e^Beta (robusto) - Bajo	IC95% e^Beta (robusto) - Alto	p-valor (robusto)	Signif.	Interpretación
Años en el rol actual	YearsInCurrentRole	Regresor estandarizado (media 0, $sd = 1$); Interpretación por 1 sd .	0.6253529021550202	1.868905381806006	1.3995571548 66713e-15	2495651795282417.0	0.97192605415 796		Un aumento de 1 desviación estándar en Años en el rol actual multiplica la media esperada por aprox. 1.869 (IC95% 0.0-2495651795282417.0).
Años en la empresa x Años en el rol actual	YearsAtCompany*YearsInCurrentRole	Regresor estandarizado (media 0, $sd = 1$); Interpretación por 1 sd .	-0.5318443677491262	0.5875203661666311 80037e-17	8.2043477438 39661e-13	4207283643212937.0	0.97722070228 03279		Un aumento de 1 desviación estándar en Años en la empresa x Años en el rol actual multiplica la media esperada por aprox. 0.588 (IC95% 0.0-4207283643212937.0).
Parámetro de log_alpha sobre-dispersión (log alpha)		Estimado en escala log; no estandarizado.	-0.6899183194179694 0.3711820660531364	0.2907698635 1837242644231.599	1837242644231.599 02046	0.97971330317			Mayor alpha implica mayor sobre-dispersión: $exp(\beta)$ es alpha en escala original.
Años en la empresa	YearsAtCompany	Regresor estandarizado (media 0, $sd = 1$); Interpretación por 1 sd .	0.8272425376662467	2.287003710947764	3.1305417678 11279e-29	1.670760642029612e+29	0.98053658629 59585		Un aumento de 1 desviación estándar en Años en la empresa multiplica la media esperada por aprox. 2.287 (IC95% 0.0-1.670760642029612e+29).
Años en la empresa x Edad	YearsAtCompany*Age	Regresor estandarizado (media 0, $sd = 1$); Interpretación por 1 sd .	0.1678842846013178	1.182799734585781 47735e-25	4.0253389217 47735e-25	3.475521538267823e+24	0.99533998175 44994		Un aumento de 1 desviación estándar en Años en la empresa x Edad multiplica la media esperada por aprox. 1.183 (IC95% 0.0-3.475521538267823e+24).
Años en la empresa x Años totales de experiencia	YearsAtCompany*TotalWorkingYears	Regresor estandarizado (media 0, $sd = 1$); Interpretación por 1 sd .	-0.257200703534344	0.7732130116069289 59023e-43	3.1281935096 59023e-43	1.91119366328278e+42	0.99587943795 43134		Un aumento de 1 desviación estándar en Años en la empresa x Años totales de experiencia multiplica la media esperada por aprox. 0.773 (IC95% 0.0-1.91119366328278e+42).
Años con el gerente actual	YearsWithCurrentManager	Regresor estandarizado (media 0, $sd = 1$); Interpretación por 1 sd .	0.1183557020888965	1.125644434241895	7.2506272746 85925e-23	1.74753899812709e+22	0.99637764706 12444		Un aumento de 1 desviación estándar en Años con el gerente actual multiplica la media esperada por aprox. 1.126 (IC95% 0.0-1.74753899812709e+22).
Ingreso mensual	MonthlyIncome	Regresor estandarizado (media 0, $sd = 1$); Interpretación por 1 sd .	0.0802705201766632	1.083580158325563 5513e-17	2.2264104032 99209	5.273717540127301e+16 99209	0.99673299059		Un aumento de 1 desviación estándar en Ingreso mensual multiplica la media esperada por aprox. 1.084 (IC95% 0.0-5.273717540127301e+16).
Attrition = Yes	Attrition_Yes	Regresor estandarizado (media 0, $sd = 1$); Interpretación por 1 sd .	0.0901180685742967 7	1.094303478929717 62101e-20	1.4141962194 51711	8.467708282045643e+19	0.99692258252		Un aumento de 1 desviación estándar en Attrition = Yes multiplica la media esperada por aprox. 1.094 (IC95% 0.0-8.467708282045643e+19).
MonthlyRate	MonthlyRate	Regresor estandarizado (media 0, $sd = 1$); Interpretación por 1 sd .	0.0503412997754627 6	1.051629956201073 23922e-12	3.1671461445 93526	349186780247.463 93526	0.99703239327		Un aumento de 1 desviación estándar en MonthlyRate multiplica la media esperada por aprox. 1.052 (IC95% 0.0-349186780247.463).
Años en la empresa x Nivel de puesto	YearsAtCompany*NivelJobLevel	Regresor estandarizado (media 0, $sd = 1$); Interpretación por 1 sd .	-0.0647921216812637 2	0.9372622796345692 51213e-20	7.1346429694 84548	1.231260743651981e+19	0.99769830125		Un aumento de 1 desviación estándar en Años en la empresa x Nivel de puesto multiplica la media esperada por aprox. 0.937 (IC95% 0.0-1.231260743651981e+19).
Edad	Age	Regresor estandarizado (media 0, $sd = 1$);	0.0458807556366788 5	1.046949560679447 56844e-18	8.2221281543 24071	1.33311396031461e+17 24071	0.99817824892		Un aumento de 1 desviación estándar en Edad multiplica la media esperada por aprox. 1.047 (IC95% 0.0-1.33311396031461e+17).

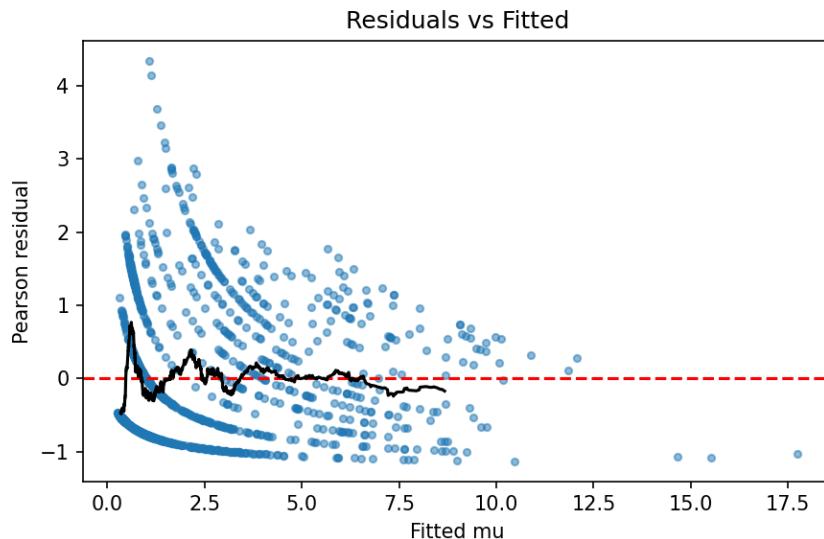
Interpretación por 1 sd.									
Rol de puesto = Research Scientist	JobRole_Research Scientist	Regresor estandarizado (media 0, sd 1); interpretación por 1 sd.	0.0381179291566558 4	1.038853736800033	9.1203042309 49337e-21	1.183312594771913e+20	0.99870921642 85234	Un aumento de 1 desviación estándar en Rol de puesto = Research Scientist, multiplica la media esperada por aprox. 1.039 (IC95% 0.0-1.183312594771913e+20).	
Estado civil = Married	MaritalStatus_Married	Regresor estandarizado (media 0, sd 1); interpretación por 1 sd.	0.0477836123875236 3	1.048943652352486	4.5072377611 94972e-26	2.441146538315451e+25	0.99872063979 32094	Un aumento de 1 desviación estándar en Estado civil = Married multiplica la media esperada por aprox. 1.049 (IC95% 0.0-2.441146538315451e+25).	
Nivel de puesto	JobLevel	Regresor estandarizado (media 0, sd 1); interpretación por 1 sd.	- 0.0205360582572319 2	0.979673370523228	1.4740361213 28261e-14	65111017228498.99	0.99899095968 50575	Un aumento de 1 desviación estándar en Nivel de puesto multiplica la media esperada por aprox. 0.98 (IC95% 0.0-65111017228498.99).	
Satisfacción con el entorno	EnvironmentSatisfaction	Regresor estandarizado (media 0, sd 1); Interpretación por 1 sd.	0.0341548375926379 4	1.034744811720562	1.3314017141 69573e-25	8.04187657254746e+24	0.99906803730 64886	Un aumento de 1 desviación estándar en Satisfacción con el entorno multiplica la media esperada por aprox. 1.035 (IC95% 0.0-8.041876572544746e+24).	
RelationshipSatisfaction	RelationshipSatisfaction	Regresor estandarizado (media 0, sd 1); interpretación por 1 sd.	0.0307407205475716 2	1.031218095558128	2.1290560628 11547e-25	4.994752271587578e+24	0.99915421807 65003	Un aumento de 1 desviación estándar en RelationshipSatisfaction multiplica la media esperada por aprox. 1.031 (IC95% 0.0-4.994752271587578e+24).	
Nivel educativo	Education	Regresor estandarizado (media 0, sd 1); interpretación por 1 sd.	0.0161403678788636 4	1.01627132724525	2.8172705635 15739e-25	3.665985879936049e+24	0.99955360909 51205	Un aumento de 1 desviación estándar en Nivel educativo multiplica la media esperada por aprox. 1.016 (IC95% 0.0-3.665985879936049e+24).	
EmployeeNumber	EmployeeNumber	Regresor estandarizado (media 0, sd 1); interpretación por 1 sd.	0.0094402890357855	1.009484989114152	1.1057010779 39697e-22	9.216414486505376e+21	0.99970803939 7798	Un aumento de 1 desviación estándar en EmployeeNumber multiplica la media esperada por aprox. 1.009 (IC95% 0.0-9.216414486505376e+21).	
Años totales de experiencia	TotalWorkingYears	Regresor estandarizado (media 0, sd 1); interpretación por 1 sd.	- 0.0075449785161979 78	0.9924834133839081	1.1328549979 73905e-22	8.695052125857835e+21	0.99976646532 68259	Un aumento de 1 desviación estándar en Años totales de experiencia multiplica la media esperada por aprox. 0.992 (IC95% 0.0-8.695052125857835e+21).	

Principales Efectos (etiquetas de Negocio)

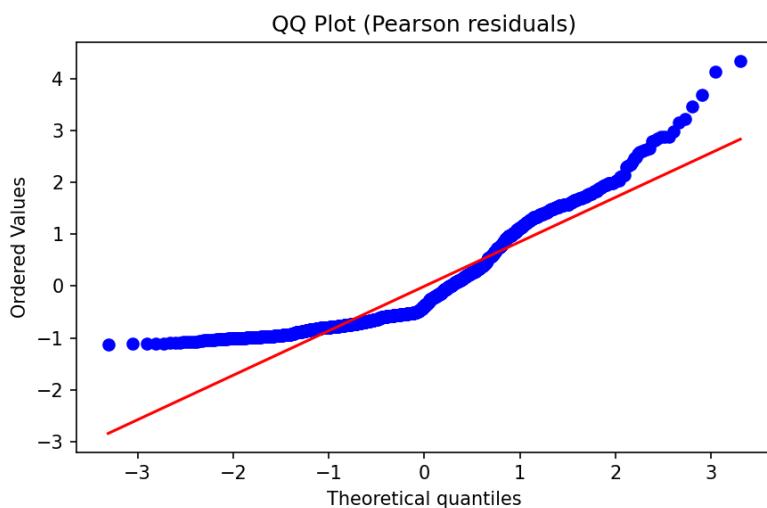
Variable (negocio)	e^Beta	IC95% e^Beta (robusto) - Bajo	IC95% e^Beta (robusto) - Alto	p-valor (robusto)	Signif.
Años en el rol actual	1.86890538180	1.399557154866713e-15	2495651795282417.0	0.97192605415796	6006
Años en la empresa	0.58752036616	8.204347743880037e-17	4207283643212937.0	0.9772207022803279	x Años en el rol actual
Parámetro de sobre-dispersión (log alpha)	0.68991831941	2.590769863539661e-13	1837242644231.599	0.9797133031702046	79694
Años en la empresa	2.28700371094	3.130541767811279e-29	1.670760642029612e+29	0.9805365862959585	7764
Años en la empresa	1.18279973458	4.025338921747735e-25	3.475521538267823e+24	0.9953399817544994	x Edad
Años en la empresa	0.77321301160	3.128193509659023e-43	1.91119366328278e+42	0.9958794379543134	x Años totales de experiencia
Años con el gerente actual	1.12564443424	7.250627274685925e-23	1.74753899812709e+22	0.9963776470612444	1895
Ingreso mensual	1.08358015832	2.22641040325513e-17	5.273717540127301e+16	0.9967329905999209	5569
Attrition = Yes	1.09430347892	1.414196219462101e-20	8.467708282045643e+19	0.9969225825251711	9717
MonthlyRate	1.05162995620	3.167146144523922e-12	349186780247.463	0.9970323932793526	1073

Gráficos de diagnóstico y sensibilidad

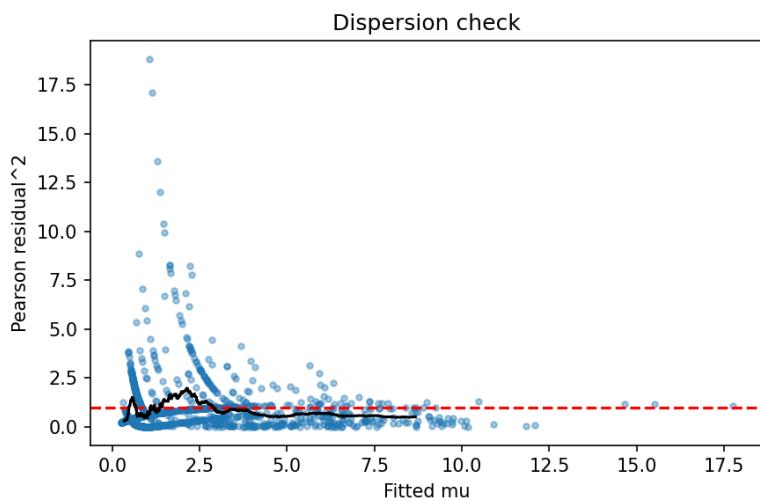
nb2_residuals_vs_fitted



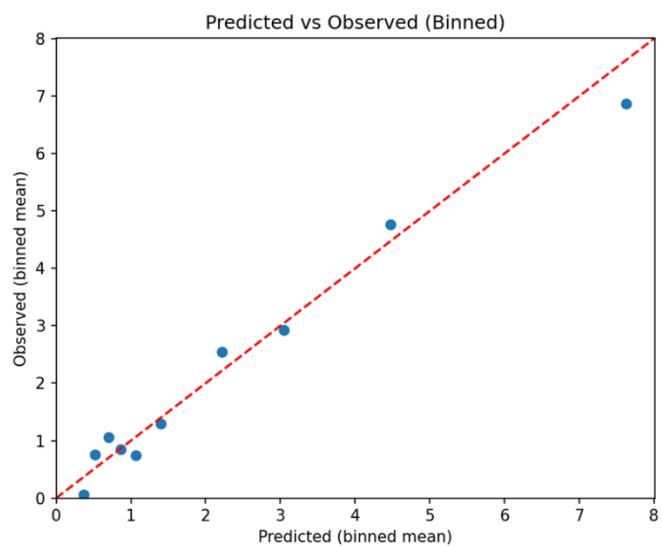
nb2_qq_plot



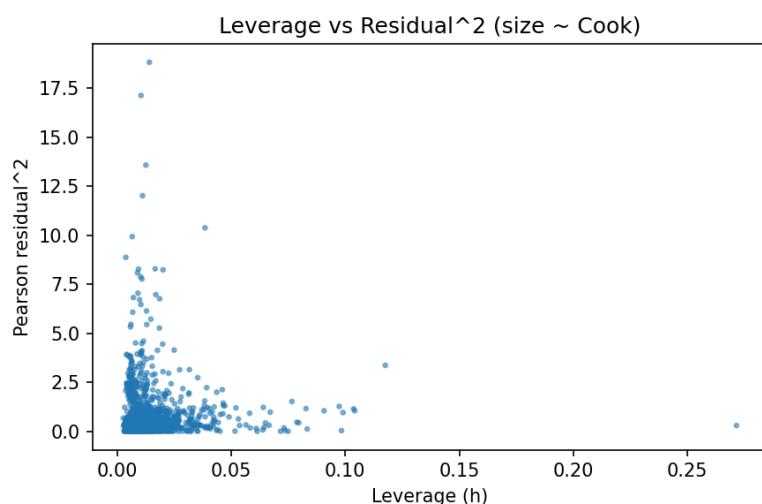
nb2_dispersion_plot



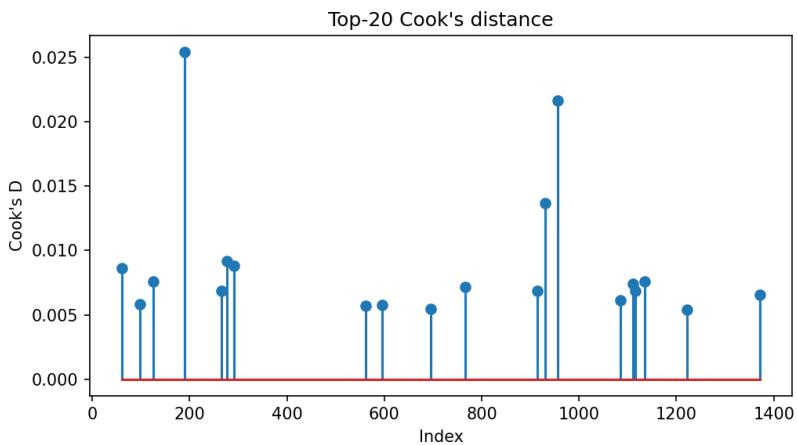
nb2_pred_vs_obs_binned



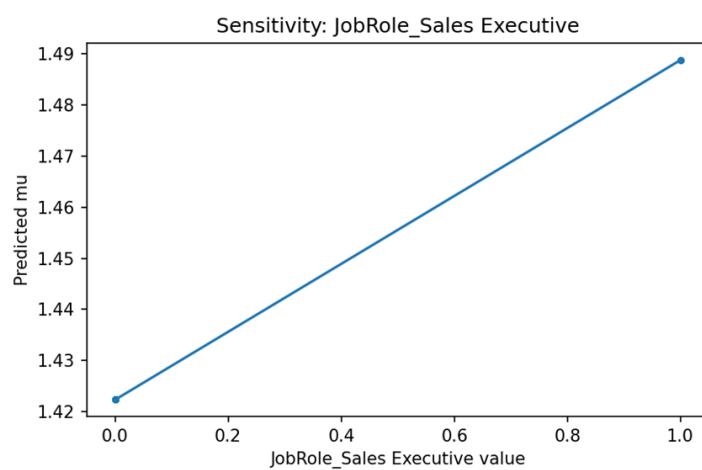
nb2_leverage_vs_residual2



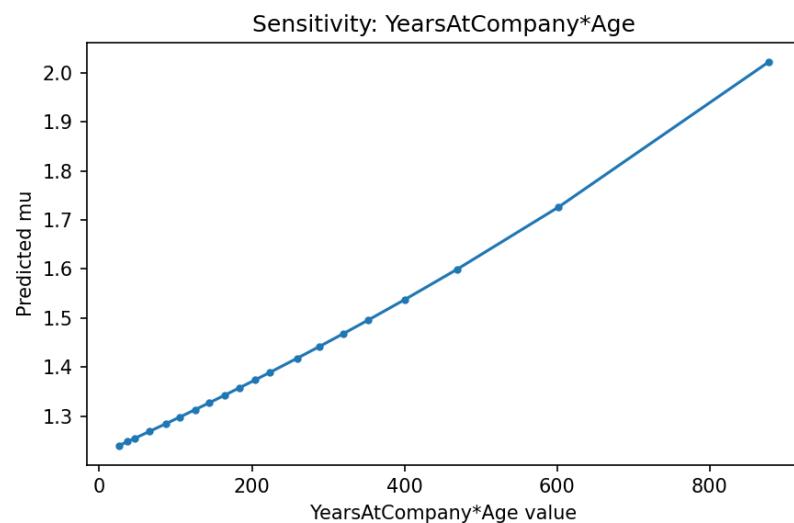
nb2_cooks_distance



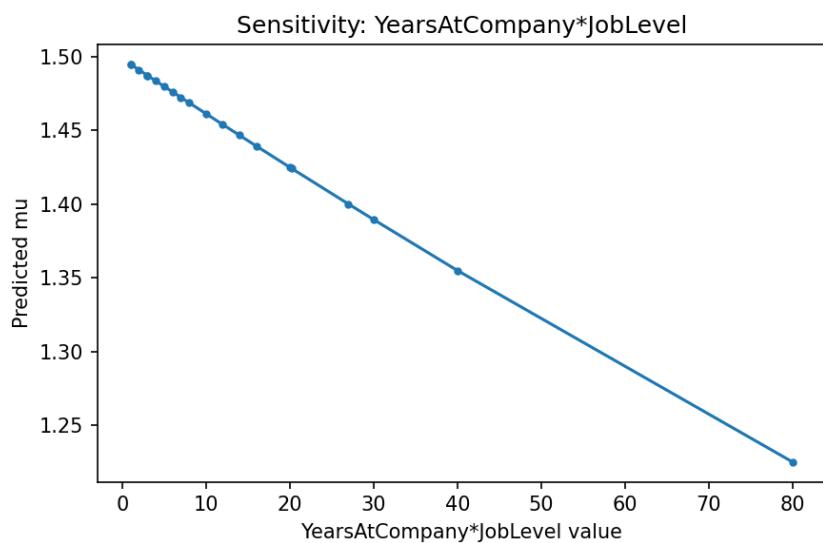
nb2_sensitivity_JobRole_Sales_Executive



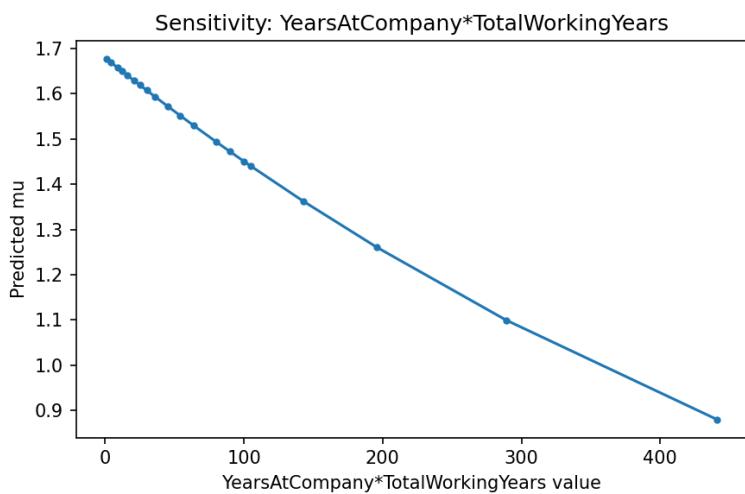
nb2_sensitivity_YearsAtCompany_Age



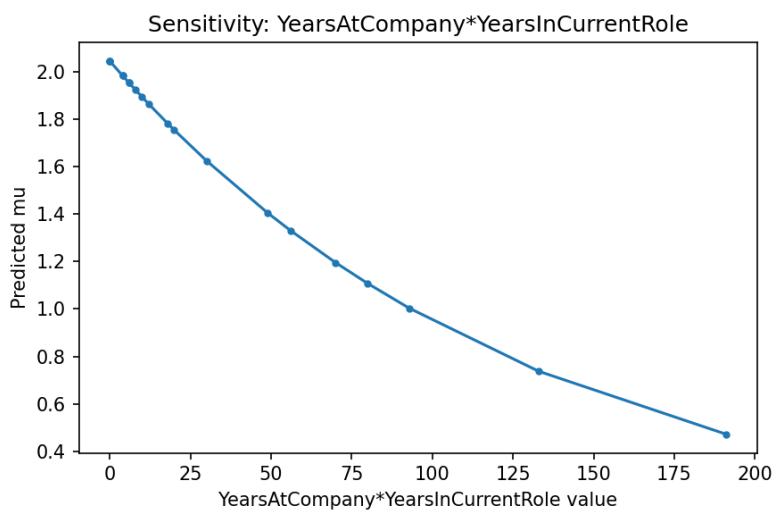
nb2_sensitivity_YearsAtCompany_JobLevel



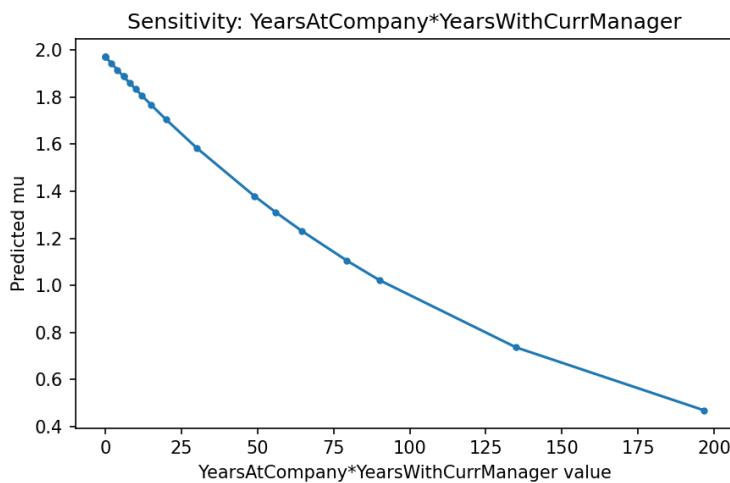
nb2_sensitivity_YearsAtCompany_TotalWorkingYears



nb2_sensitivity_YearsAtCompany_YearsInCurrentRole



nb2_sensitivity_YearsAtCompany_YearsWithCurrManager



Anexo 4. Código Fuente de análisis elaborado en el trabajo

https://github.com/adofredy/TFM-UNIR/tree/main/Codigo_TMF_FINAL

TFM-UNIR / Código_TMF_FINAL / ⌂

Adolfo Huanacuni first commit

Name
..
HR_Employee_Attrition_clean_with_diff.csv
NB2_final.ipynb
Regresion minimos cuadrados ordinariosOLS.ipynb
WA_Fn-UseC_-HR-Employee-Attrition(1).csv
WA_Fn-UseC_-HR-Employee-Attrition.csv
attrition_demografico.ipynb
auditoria_equidad_1.ipynb
auditoria_equidad_unico_final.ipynb
comparativo_tres_modelos_ols (4) (1).ipynb
generate_model_results_FINAL.ipynb