IBM
Employee Attrition

CODER HOUSE

Modelo de predicción de abandono/rotación de empleados

by Agostina del Olmo

Tabla de Contenidos

Link a notebook Colab

Link a repositorio GitHub

Problema y Contexto Comercial

¿Existen patrones y relaciones entre los datos que nos permitan identificar perfiles de empleados que son más propensos a abandonar la empresa y aquellos que no lo son?

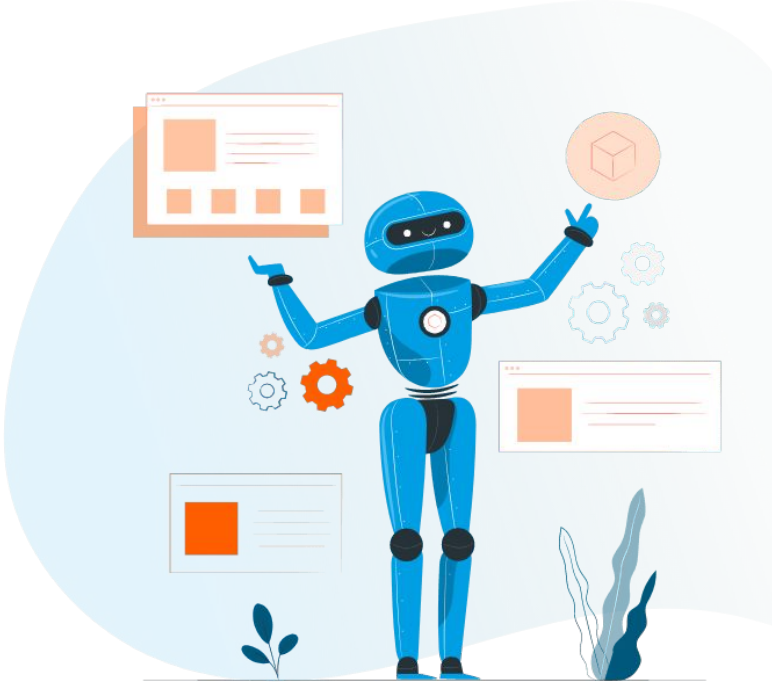
La rotación de empleados es un problema importante que puede enfrentar cualquier empresa, ya que puede afectar la rentabilidad y la productividad de la misma.

Este es un problema de **aprendizaje supervisado** ya que los datos que nos proporcionaron están etiquetados con información sobre si los empleados han abandonado la empresa o no.

Contexto Analítico

Se nos ha proporcionado un **conjunto de datos** que contiene información sobre la edad, salario, satisfacción laboral y otros **factores relacionados con la rotación de empleados en IBM**. Vamos a utilizar técnicas de aprendizaje automático para construir un modelo predictivo que pueda predecir la rotación de empleados en el futuro.

Este es un problema de aprendizaje supervisado, de clasificación, ya que los datos están etiquetados con información sobre si los empleados han abandonado la empresa o no.



Audiencia

La audiencia para este análisis es el **equipo de recursos humanos de la empresa**, ya que son los encargados de mejorar la retención de empleados y disminuir la rotación. También podría ser útil para la **gerencia** de IBM, por motivos similares.

Por el contenido técnico de este proyecto y las metodologías de trabajo utilizadas, también sería interesante mostrar este proyecto al **equipo de Data Science**, ya que posee el conocimiento y la experiencia necesaria para comprender los aspectos aquí involucrados.

Objetivo

Proporcionar una comprensión clara de los factores que influyen en la rotación de empleados en IBM es a lo que apunta este análisis.

Por otro lado, narrar este trabajo utilizando herramientas de storytelling, tiene como propósito principal explicar tanto el proceso de análisis de datos como el desarrollo del modelo y los resultados obtenidos. Al hacerlo, se busca brindar información valiosa a quienes recurren a este proyecto, permitiéndoles tomar decisiones fundamentadas para retener a los empleados más valiosos.

Contenido

El enfoque elegido para este storytelling se dirige a mostrar las relaciones y patrones significativos entre las variables involucradas, resaltando las implicaciones que tienen en la retención de empleados. A través de una narrativa que pretende ser cautivadora, **se busca transmitir la importancia de abordar este problema y proporcionar recomendaciones prácticas para mejorar la retención y el compromiso de los empleados en IBM.**

Preguntas a responder

- ¿Cuáles son los factores más influyentes en la decisión de los empleados de abandonar la empresa?
- ¿Cómo afectan los niveles de satisfacción laboral y compromiso de los empleados en su decisión de abandonar la empresa?
- ¿Hay alguna relación entre la edad y la rotación de los empleados?
- ¿Cómo influyen la carga de trabajo, el salario y otros factores en la rotación de los empleados?
- ¿Existen diferencias salariales significativas entre hombres y mujeres?

Hipótesis

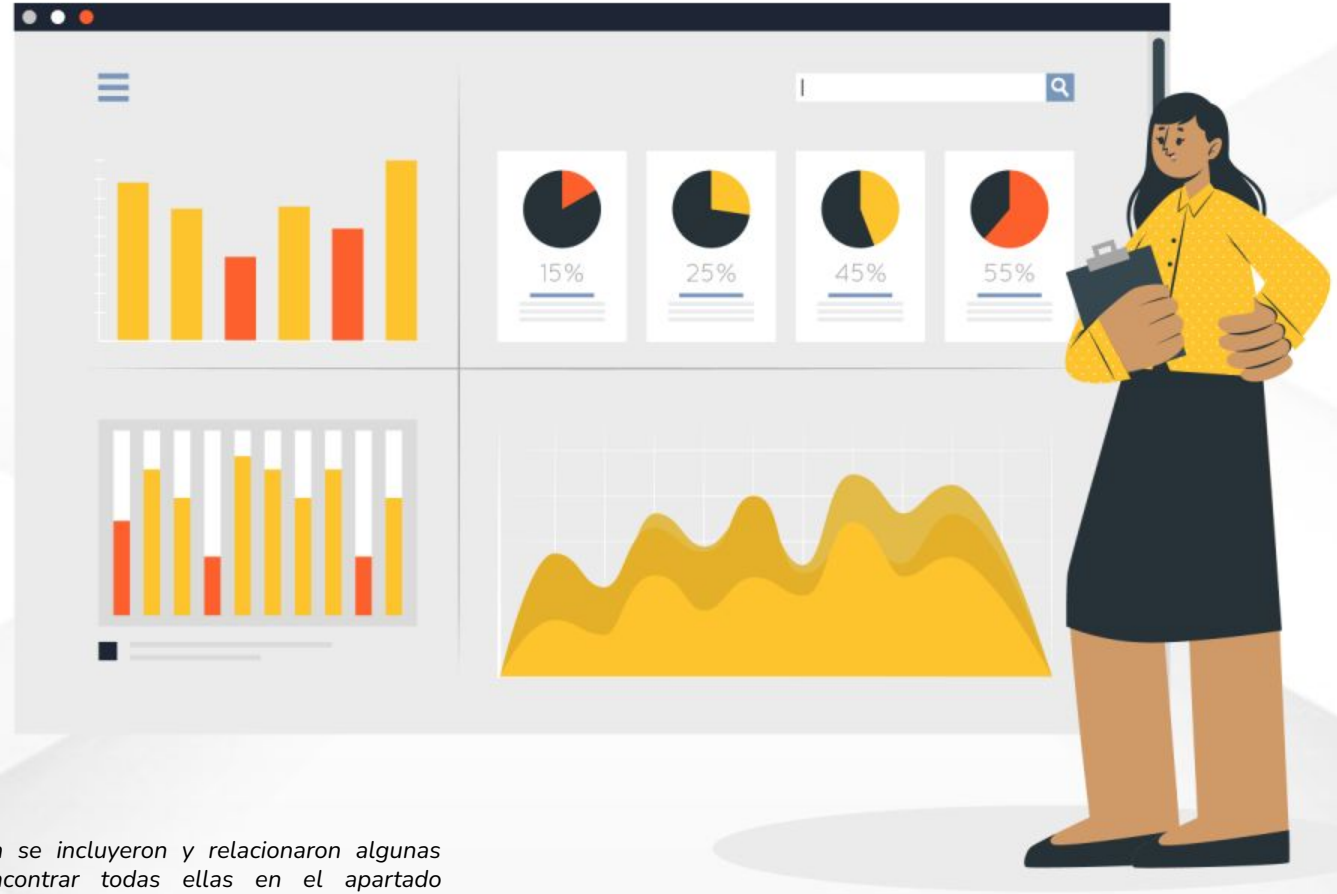
- Los empleados con niveles bajos de satisfacción laboral son más propensos a abandonar la empresa.
- Los empleados con mayores cargas de trabajo tienen más probabilidades de abandonar la empresa.
- Los empleados más jóvenes son más propensos a abandonar la empresa que los mayores.
- Los empleados que tienen que viajar largas distancias para llegar al trabajo tienen más probabilidades de abandonar la empresa.
- Los empleados que ganan más tienen más probabilidades de quedarse en la empresa.
- Existe una relación entre las oportunidades de promoción y el abandono de la empresa.

Descripción de los campos

Link a Imagen: [Descripción de los campos del dataset]

Campo	Descripción
Age Employee's	Edad del empleado
Gender Employee's	Género del empleado
BusinessTravel	Frecuencia de viajes de negocios de los empleados
DailyRate	Salario diario del empleado
Department	Departamento de la empresa al que pertenece el empleado
DistanceFromHome	Distancia en millas desde el hogar del empleado hasta el lugar de trabajo
Education	Nivel de educación alcanzado--> 1: Below College - 2: College - 3: Bachelor - 4: Master - 5: Doctor
EducationField	Área de estudio del empleado
EmployeeCount	Número total de empleados en la organización
EmployeeNumber	Identificador único para cada registro de empleado
EnvironmentSatisfaction	Satisfacción del empleado con su entorno laboral
HourlyRate	Tarifa por hora para los empleados
JobInvolvement	Nivel de compromiso requerido para el trabajo del empleado
JobLevel	Nivel de trabajo del empleado
JobRole	Papel del empleado en la organización
JobSatisfaction	Satisfacción del empleado con su trabajo
MaritalStatus	Estado civil del empleado
MonthlyIncome	Ingreso mensual del empleado
MonthlyRate	Tarifa salarial mensual para los empleados
NumCompaniesWorked	Número de empresas para las que trabajó el empleado
Over18	Si el empleado es mayor de 18 años
OverTime	Si los empleados trabajan horas extra
PercentSalaryHike	Tasa de aumento salarial para los empleados
PerformanceRating	La calificación de rendimiento del empleado
RelationshipSatisfaction	Satisfacción del empleado con sus relaciones interpersonales
StandardHours	Horas de trabajo estándar para los empleados
StockOptionLevel	Nivel de opción de compra de acciones del empleado
TotalWorkingYears	Número total de años trabajados por el empleado
TrainingTimesLastYear	Número de veces que los empleados asistieron a capacitación en el último año
WorkLifeBalance	Percepción de los empleados sobre su equilibrio entre el trabajo y la vida personal
YearsAtCompany	Número de años que los empleados han estado en la empresa
YearsInCurrentRole	Número de años que el empleado ha estado en su rol actual
YearsSinceLastPromotion	Número de años desde la última promoción del empleado
YearsWithCurrManager	Número de años que el empleado ha estado con su gerente actual
Attrition	Si el empleado abandonó la organización

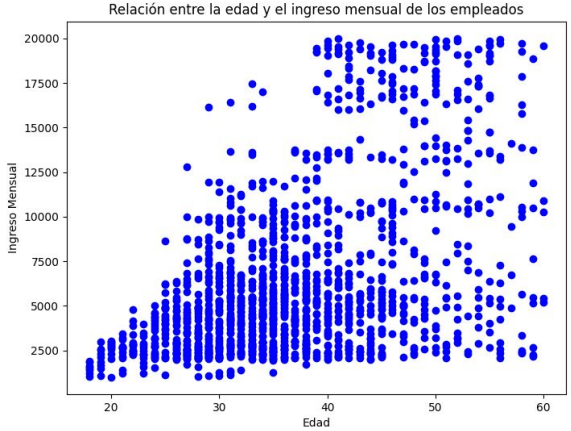
Análisis Exploratorio de Datos (EDA)

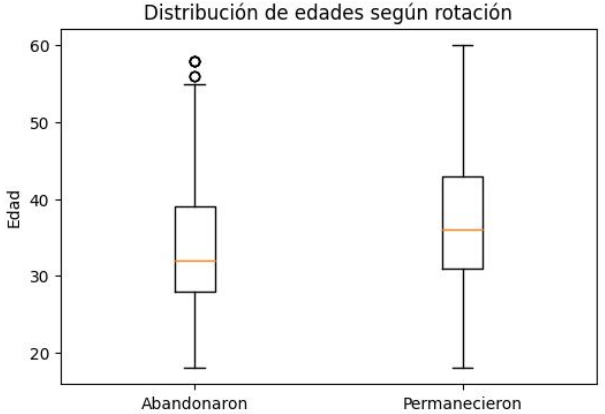


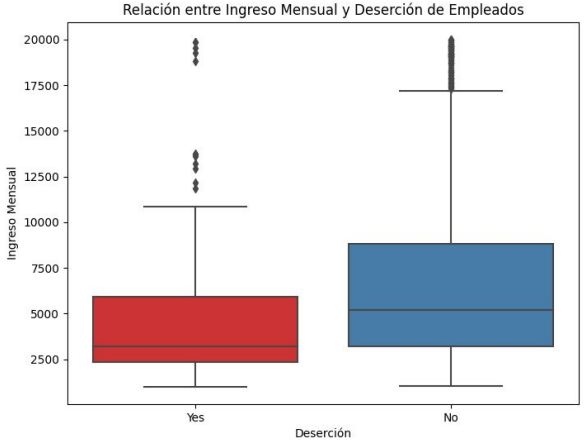
Nota: En esta presentación se incluyeron y relacionaron algunas visualizaciones, podrán encontrar todas ellas en el apartado correspondiente del notebook del proyecto.

Link a esta sección en el notebook

Relación entre edad, ingresos mensuales y deserción de los empleados



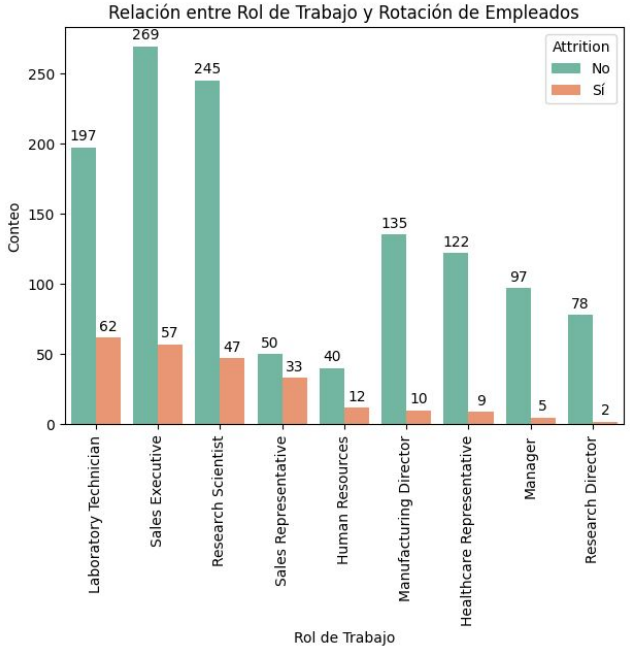


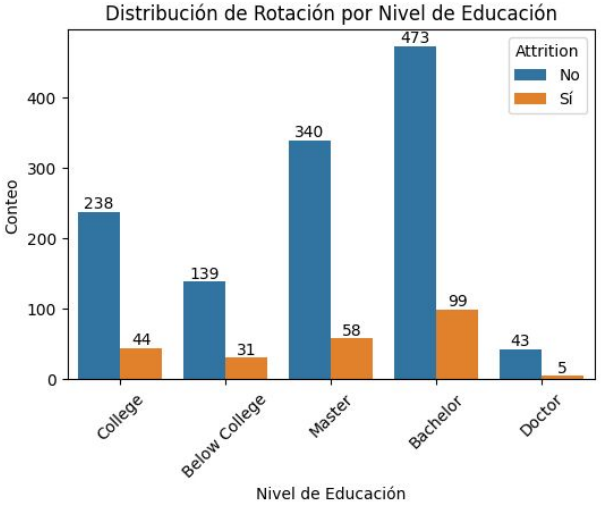


Se puede observar que los empleados que abandonaron la empresa tienen una edad media más baja que los empleados que permanecieron en la empresa. Esto sugiere que **los empleados más jóvenes tienen más probabilidades de abandonar la empresa**.

Además, existe una **relación positiva entre la edad y el ingreso mensual de los empleados**. Esto puede deberse, entre otros posibles factores, a que los empleados mayores tienen más experiencia y conocimientos, lo que les permite obtener salarios más altos; o que suelen ser más estables en sus carreras y menos propensos a buscar nuevas oportunidades.

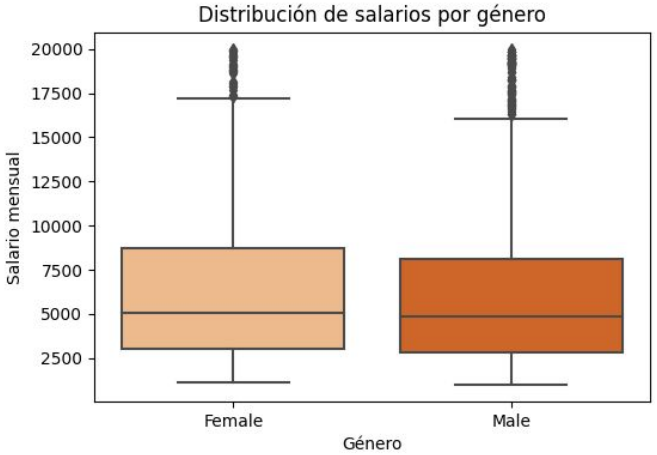
Relación entre Rol, Nivel de Educación y Rotación

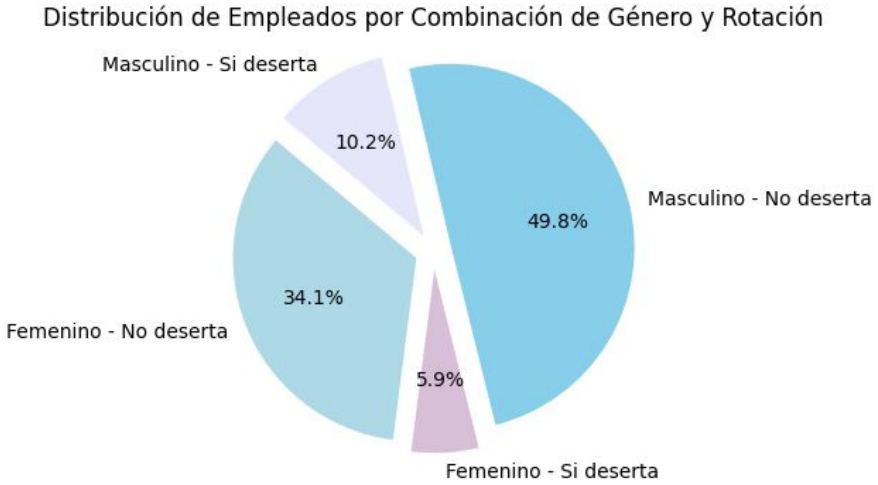




- Se puede observar que el **dpto. de ventas y de RRHH son los que tienen la tasa de rotación más alta**. Esto puede indicar que los empleados en estos departamentos pueden estar experimentando mayores niveles de estrés, carga de trabajo y/o insatisfacción laboral, lo que lleva a una mayor tasa de rotación.
- Además, vemos que aquellos empleados con grados 3 (Bachelor) y 4 (Master) son los más propensos a desertar.
- Los 3 puestos o funciones en los que los empleados tienen más probabilidades de dimitir son:
    - Técnico de laboratorio
    - Ejecutivo de ventas
    - Científico de investigación

Relación entre género, salario y rotación de personal



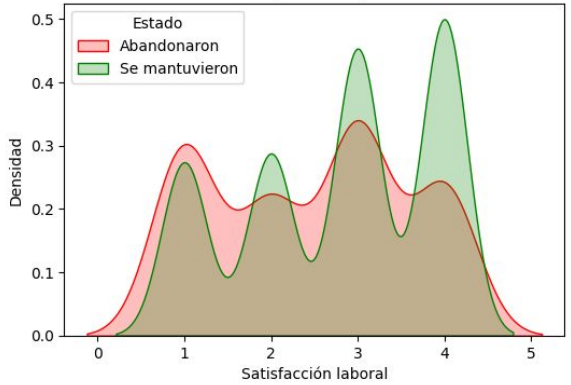


Podemos observar que para empleados del género masculino parece haber una mayor dispersión de los datos relacionados a salarios. Vemos que algunos hombres ganan salarios significativamente más altos que la mayoría de los demás hombres y mujeres en la empresa.

Sin embargo, podemos ver que tanto mujeres como hombres presentan empleados que desertan en proporciones similares, lo que sugiere que **la rotación de empleados no parece estar relacionada directamente con el género en esta organización.**

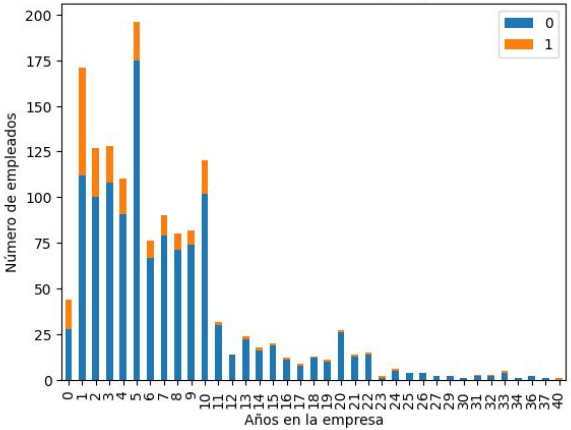
Satisfacción laboral y duración en la empresa de acuerdo a la rotación

Distribución de la satisfacción laboral en empleados que abandonaron vs los que se mantuvieron



Vemos que los empleados que abandonaron la empresa presentan una mayor densidad de valores bajos de satisfacción laboral en comparación con los empleados que se mantuvieron. **Esto sugiere que la satisfacción laboral puede ser un factor importante en la decisión de abandonar la empresa.**

Distribución de empleados por duración de empleo y abandono de la empresa



Se observa que **la tasa de abandono es más alta en los primeros años de trabajo (1-5 años),** y luego disminuye a medida que los empleados permanecen más tiempo en la empresa. Los empleados que han estado en la empresa durante más de 10 años tienen la tasa de abandono más baja.

Correlación de variables

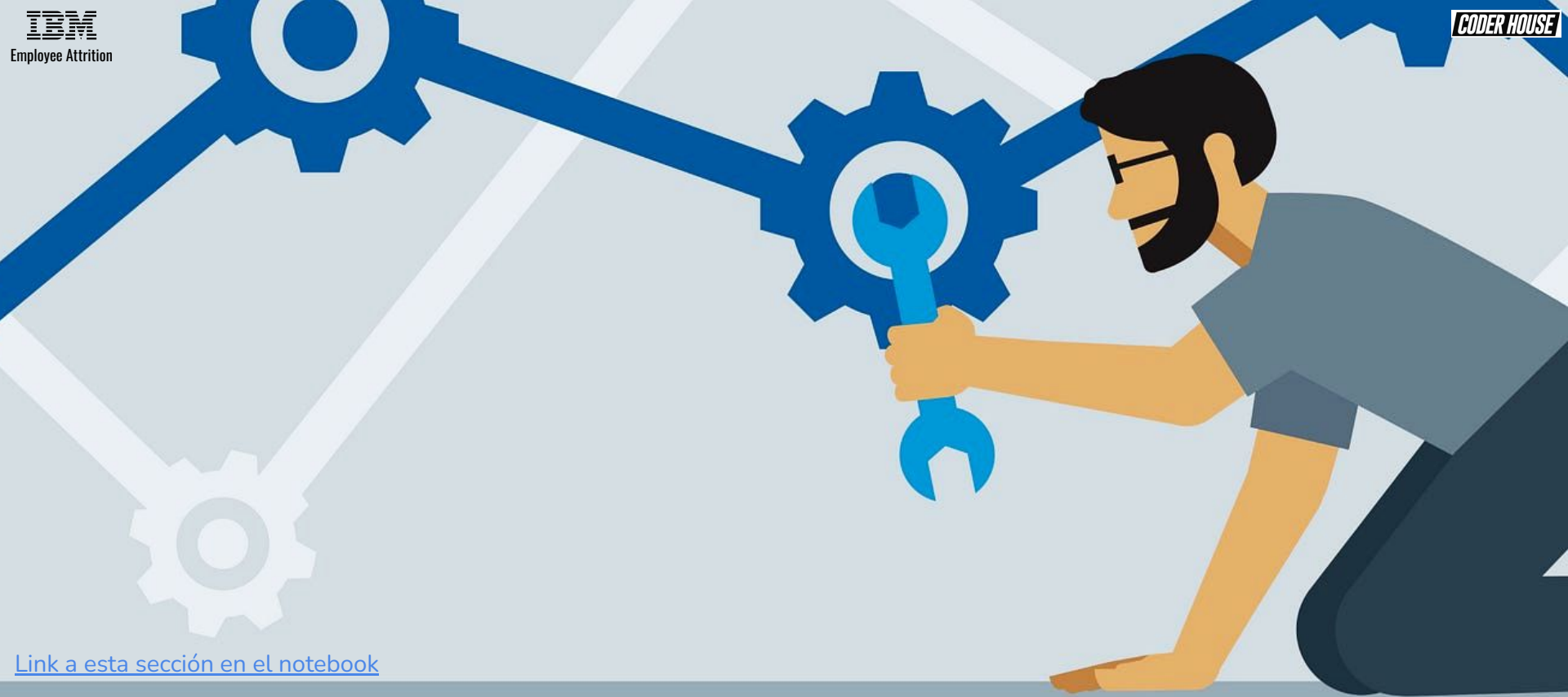
Feature 1	Feature 2	Correlation
MonthlyIncome	JobLevel	0.950300
JobLevel	MonthlyIncome	0.950300
TotalWorkingYears	JobLevel	0.782208
JobLevel	TotalWorkingYears	0.782208
PerformanceRating	PercentSalaryHike	0.773550
PercentSalaryHike	PerformanceRating	0.773550
TotalWorkingYears	MonthlyIncome	0.772893
MonthlyIncome	TotalWorkingYears	0.772893
YearsWithCurrManager	YearsAtCompany	0.769212
YearsAtCompany	YearsWithCurrManager	0.769212
YearsInCurrentRole	YearsAtCompany	0.758754
YearsAtCompany	YearsInCurrentRole	0.758754
YearsInCurrentRole	YearsWithCurrManager	0.714365
YearsWithCurrManager	YearsInCurrentRole	0.714365
TotalWorkingYears	Age	0.680381
Age	TotalWorkingYears	0.680381
JobRole	Department	0.662431
Department	JobRole	0.662431
StockOptionLevel	MaritalStatus	-0.662577
MaritalStatus	StockOptionLevel	-0.662577

Las correlaciones identificadas pueden estar indicando que:

- los empleados tienden a permanecer más tiempo en la empresa cuando están **satisfechos con su puesto** y tienen una **buena relación con su gerente**.
- **escalar de posición** o crecer profesionalmente dentro de la empresa influye en la decisión de los trabajadores de seguir siendo parte de IBM.
- los empleados casados tienden a tener un nivel de opciones sobre acciones más bajo en comparación con los solteros.
- los empleados que **reciben aumentos salariales tienen un rendimiento calificado como alto**.
- a medida que los empleados acumulan más años de experiencia laboral, es más probable que alcancen niveles de empleo más altos, lo que a su vez se asocia con un mayor ingreso mensual. Esto puede generar un **sentido de estabilidad y perspectivas de avance,** además podrían sentir que se les reconoce y recompensa económicamente por su experiencia, lo que podría influir en la decisión de los empleados de quedarse en la empresa.

Conclusiones y Recomendaciones post análisis exploratorio

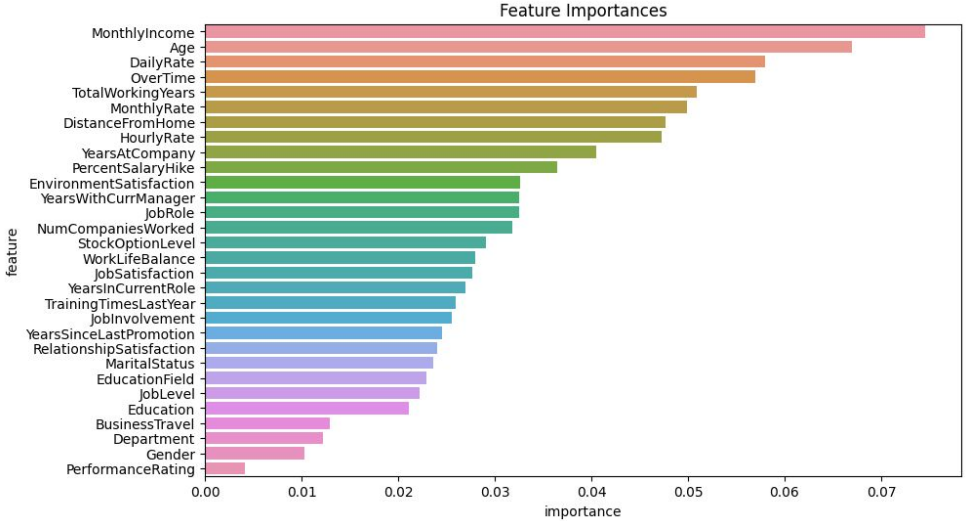
- La edad y la satisfacción laboral parecen ser factores importantes en la rotación de empleados. Los empleados más jóvenes y los que tienen niveles más bajos de satisfacción laboral tienden a abandonar la empresa con más frecuencia. Por lo tanto, **se recomienda que la empresa se centre en mejorar factores que influyen en la satisfacción laboral y las oportunidades de desarrollo para los empleados jóvenes.**
- **Se recomienda a la empresa que analice más detenidamente los problemas específicos en los departamentos con tasas de rotación más altas** y tome medidas para abordarlos.
- Se observa que la tasa de abandono es más alta en los primeros años (1-5 años) de trabajo, por lo tanto, **se recomienda a la empresa que preste especial atención a la retención de los empleados durante los primeros años de su empleo, mediante la implementación de programas de capacitación y desarrollo.**
- Es fundamental para una empresa mantener a empleados con niveles de educación como licenciaturas o maestrías, ya que estos profesionales suelen aportar un alto valor a la organización. P**ara fomentar su retención, es necesario revisar y mejorar los beneficios y oportunidades que se les ofrecen.**



Link a esta sección en el notebook

Feature Selection y Feature Engineering

Random Forest para selección de variables



El uso de la variable "MonthlyIncome" como la más importante **indica que el nivel de ingresos de los empleados puede desempeñar un papel significativo en su decisión de renunciar o no.**

Otros factores importantes son la edad de los empleados, el tiempo que han estado en la compañía, el salario diario y si realizan horas extra. Estos **indican que la experiencia, la estabilidad laboral y la carga de trabajo pueden ser factores determinantes en la retención de empleados.**

Correlación:

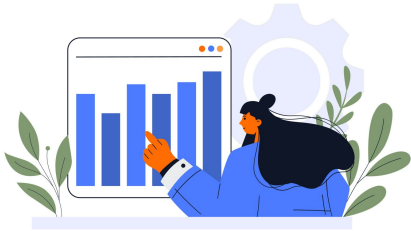
Por otro lado, analizando la correlación entre variables, se pudo observar que muchas de ellas tenían un alto valor de correlación, por lo que conservé en el dataset las que consideré más relevantes para el fin de este proyecto.

Detección de Outliers:

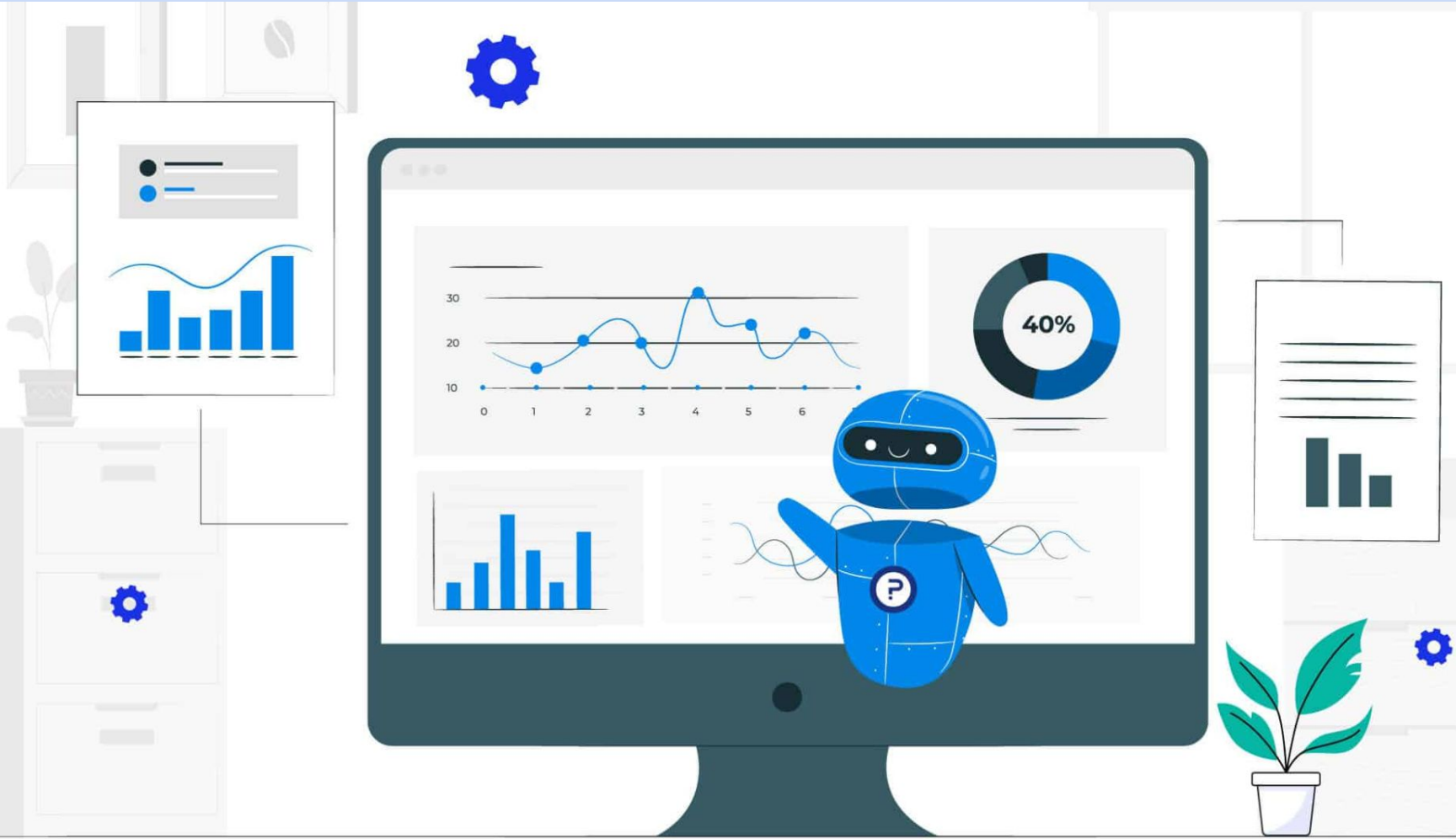
- Se aplica la detección de outliers utilizando el algoritmo Isolation Forest. Esto ayuda a identificar observaciones inusuales en los datos que pueden sesgar el modelo.

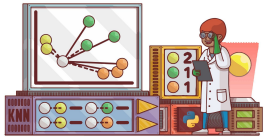
Análisis de Componentes Principales (PCA):

- Se utiliza el análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos.
- Se muestra un gráfico que representa la varianza explicada acumulativa en función del número de componentes principales. Esto ayuda a determinar cuántos componentes principales retener. En este caso, se opta por retener las primeras 18 componentes principales, que explican aproximadamente el 81.6% de la varianza total de los datos originales.
- Se aplica PCA a los datos sin outliers (previamente detectados con Isolation Forest) para obtener un conjunto de datos de dimensiones reducidas que conserva la mayoría de la información relevante.



Entrenamiento y Evaluación de Modelos de Machine Learning





Modelo KNN

Métricas:

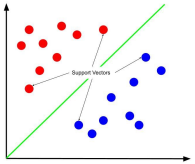
Precisión del modelo: 86.73%

Matriz de Confusión:
Verdaderos Negativos: 252 - Falsos Positivos: 3
Falsos Negativos: 36 - Verdaderos Positivos: 3

Informe de clasificación:

	precision	recall	f1-score	support
0	0.88	0.99	0.93	255
1	0.50	0.08	0.13	39

Resumen:
La precisión general es buena, pero la capacidad para predecir rotación es limitada. Bajo recall y F1-score para la clase de rotación indican dificultades en identificar a empleados en riesgo.



Modelo SVM

Métricas:

Precisión del modelo: 86.73%

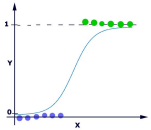
Matriz de Confusión:
Verdaderos Negativos: 255 - Falsos Positivos: 39
Falsos Negativos: 0 - Verdaderos Positivos: 0

Informe de clasificación:

	precision	recall	f1-score	support
0	0.87	1.00	0.93	255
1	0.00	0.00	0.00	39

Resumen:
La falta de recall y F1-score para la clase de rotación indica que el modelo no es capaz de identificar adecuadamente a los empleados en riesgo de rotación.

Nota: Clase 0 = No Renuncia/Rotación – Clase 1 = Renuncia/Rotación



Modelo LogisticRegression

Métricas:

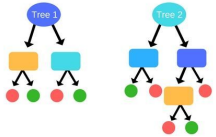
Precisión del modelo: 89.80%

Matriz de Confusión:
Verdaderos Negativos: 254 - Falsos Positivos: 1
Falsos Negativos: 29 - Verdaderos Positivos: 10

Informe de clasificación:

	precision	recall	f1-score	support
0	0.90	1.00	0.94	255
1	0.91	0.26	0.40	39

Resumen:
Muestra una alta precisión general y un buen rendimiento en la clasificación de la clase 0 . Sin embargo, el rendimiento en la clasificación de la clase 1 es menos satisfactorio. En general, el modelo es efectivo para predecir la no rotación, pero se puede mejorar en la predicción de la rotación.



Modelo RandomForestClassifier

Métricas:

Precisión del modelo: 86.73%

Matriz de Confusión:
Verdaderos Negativos: 255 - Falsos Positivos: 0
Falsos Negativos: 39 - Verdaderos Positivos: 0

Informe de clasificación:

	precision	recall	f1-score	support
0	0.87	1.00	0.93	255
1	0.00	0.00	0.00	39

Resumen:
Muestra una alta precisión general, pero enfrenta un problema de underfitting en la clasificación de la clase de rotación. La falta de recall y F1-score para la clase de rotación indica que el modelo no es capaz de identificar adecuadamente a los empleados en riesgo de rotación.

Nota: Clase 0 = No Renuncia/Rotación – Clase 1 = Renuncia/Rotación

Modelo XGBoost

Métricas:

Precisión del modelo: 88.44%

Matriz de Confusión:
Verdaderos Negativos: 252 - Falsos Positivos: 3
Falsos Negativos: 31 - Verdaderos Positivos: 8

Informe de clasificación:

	precision	recall	f1-score	support
0	0.89	0.99	0.94	255
1	0.73	0.21	0.32	39

Resumen:
Muestra una precisión general sólida y un rendimiento mejorado en la clasificación de la clase Rotación en comparación con algunos de los modelos anteriores. Aunque el recall y el F1-score para la clase 1 aún son modestos, indican una mejora en la identificación de empleados en riesgo de rotación en comparación con otros modelos.

Nota: Clase 0 = No Renuncia/Rotación – Clase 1 = Renuncia/Rotación

Optimización de Hiperparámetros

Random Grid Search
(modelo RandomForestClassifier)

Métricas:

Precisión del modelo: 87.07%

Matriz de Confusión:
Verdaderos Negativos: 250 - Falsos Positivos: 5
Falsos Negativos: 33 - Verdaderos Positivos: 6

Informe de clasificación:

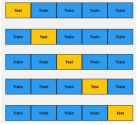
	precision	recall	f1-score	support
0	0.88	0.98	0.93	255
1	0.55	0.15	0.24	39

Resumen:
Si bien muestra una mejora con respecto al modelo RandomForestClassifier, aún enfrenta un desafío en la clasificación de la clase de rotación con un recall bajo, lo que indica dificultades en identificar adecuadamente a los empleados en riesgo de rotación. La precisión y el F1-score para la clase de rotación son moderados.

Resumen de Evaluaciones y Métricas de los Modelos de ML

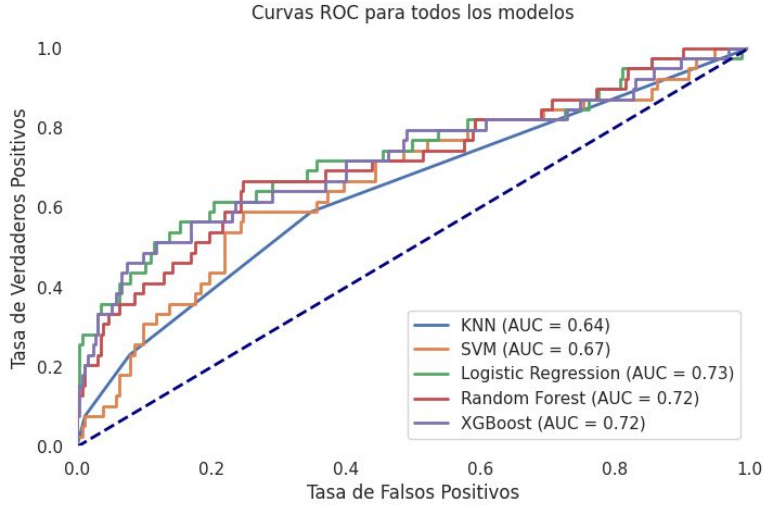
Métricas generales:

	accuracy	precision	f1 score	recall
KNN	0.86735	0.50000	0.13333	0.07692
SVM	0.86735	0.00000	0.00000	0.00000
Logistic Regression	0.89796	0.90909	0.40000	0.25641
Random Forest	0.86735	0.00000	0.00000	0.00000
XGBoost	0.88435	0.72727	0.32000	0.20513
Random Grid Search	0.87075	0.54545	0.24000	0.15385



Resultados de la K-Fold Cross Validation para los 5 modelos:

Modelo **KNN** → Precisión Promedio: **84.12%**
Modelo **SVM** → Precisión Promedio: **84.78%**
Modelo **LogisticRegression** → Precisión Promedio: **86.11%**
Modelo **RandomForestClassifier** → Precisión Promedio: **84.78%**
Modelo **XGBoost** → Precisión Promedio: **84.97%**

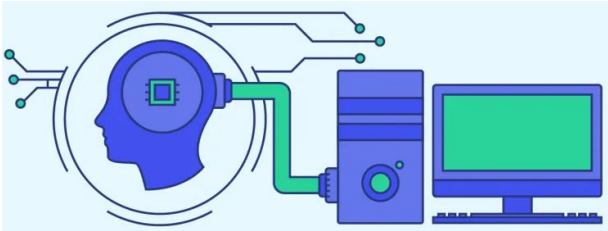


Elección del Modelo

A pesar de que el modelo Random Forest mejorado con Random Grid Search muestra una precisión superior en comparación con los otros modelos, es el modelo Logistic Regression el que sobresale en términos de exactitud (accuracy), puntaje F1 (F1 score) y exhaustividad (recall).

Además, es importante destacar la eficiencia en el tiempo de ejecución del modelo Logistic Regression en comparación con el Random Forest con Random Grid Search, lo que lo convierte en una elección aún más atractiva para aplicaciones prácticas.

Podemos destacar que el modelo XGBoost se acerca muchísimo a las métricas correspondientes al modelo de LogisticRegression, por lo que también es un buen candidato a ser elegido como mejor modelo. En resumen, **el modelo LogisticRegression se posiciona como la solución más eficaz y eficiente para nuestro problema de clasificación.**



Conclusiones generales

En resumen, los datos y análisis de los mismos indican que **la satisfacción laboral es un factor crítico en la rotación de empleados**, mientras que la edad no parece ser determinante. También se evidencia una **brecha salarial** de género que requiere atención. Además, l**a carga de trabajo y las oportunidades de promoción influyen en la decisión de los empleados** de quedarse o abandonar la empresa.

La implementación de **herramientas de Machine Learning y Data Science** proporciona a las empresas una **ventaja competitiva al prever y abordar posibles rotaciones, lo que conduce a una mayor estabilidad laboral y ahorro de costos.** La retención de empleados valiosos y la mejora de la satisfacción laboral son esenciales para un entorno de **trabajo saludable y la retención de talento clave.**

Agostina del Olmo.
Sep, 2023. Data Science. 39990.