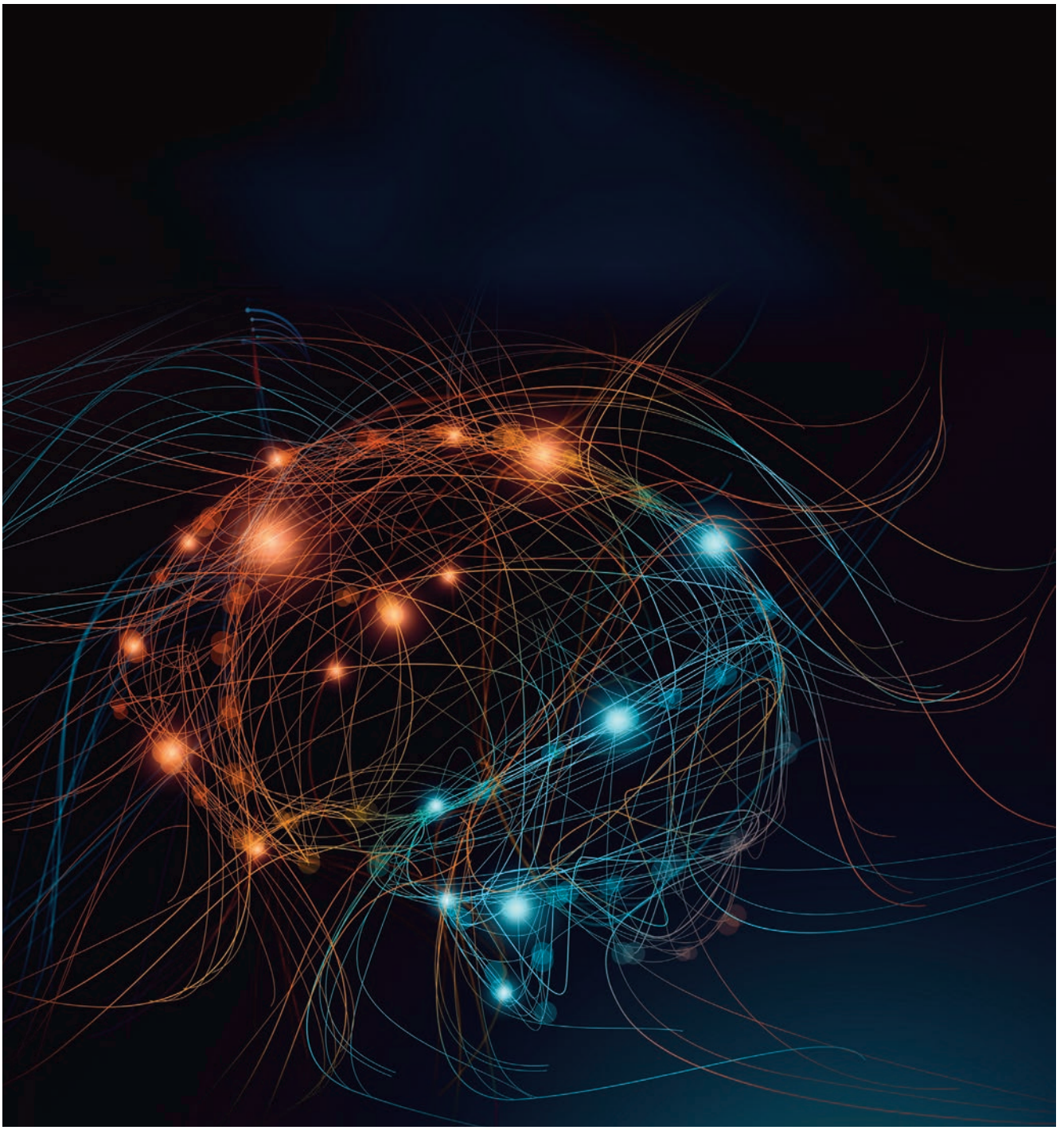
Caso práctico de interpretabilidad

"Los necios ignoran la complejidad. Los pragmáticos la sufren. Algunos pueden evitarla. Los genios la eliminan".

Alan Perlis[72]



Planteamiento

En esta sección se presenta un caso práctico de interpretabilidad en inteligencia artificial con el objetivo de ilustrar cómo se aplican las técnicas de XAI descritas en la sección anterior.

El caso de estudio seleccionado aborda el problema de la retención de empleados en una organización, centrándose en comprender y explicar las causas que llevan a los empleados a abandonar su puesto de trabajo. La identificación de estos factores puede permitir a las organizaciones tomar medidas preventivas y desarrollar estrategias para mejorar la satisfacción laboral y la retención del talento.

En este caso práctico, se utilizará un conjunto de datos ficticios generados por IBM y publicados en Kaggle[73]. Este conjunto de datos contiene información sobre los empleados de una organización, incluyendo características demográficas, datos sobre su puesto de trabajo, y si han abandonado la empresa o no.

En el ejercicio que se plantea, la compañía presenta un nivel de abandono de empleados del 16%, un 6% por encima del promedio histórico, y está preocupada por conocer las causas para elaborar un plan de remediación.

Las principales variables presentes en el conjunto de datos incluyen:

- Nivel de educación (desde "secundaria" hasta "doctorado").
- Satisfacción con el ambiente laboral (desde "bajo" hasta "muy alto").
- Involucración en el trabajo (desde "bajo" hasta "muy alto").
- Satisfacción con el trabajo (desde "baja" hasta "muy alta").
- Calificación del rendimiento (desde "bajo" hasta "sobresaliente").
- Satisfacción con las relaciones laborales (desde "baja" hasta "muy alta").
- Equilibrio entre la vida personal y profesional (desde "malo" hasta "óptimo").
- Años desde la última promoción en el trabajo (variable numérica).
- Salario mensual (variable numérica).
- Años en el puesto de trabajo actual (variable numérica).
- Distancia al puesto de trabajo (variable numérica).
- Número de empresas en las que se ha trabajado (variable numérica).
- Rol en el puesto de trabajo actual (variable categórica, incluye *Manager*, *Director*, *Research Scientist*, entre otros).

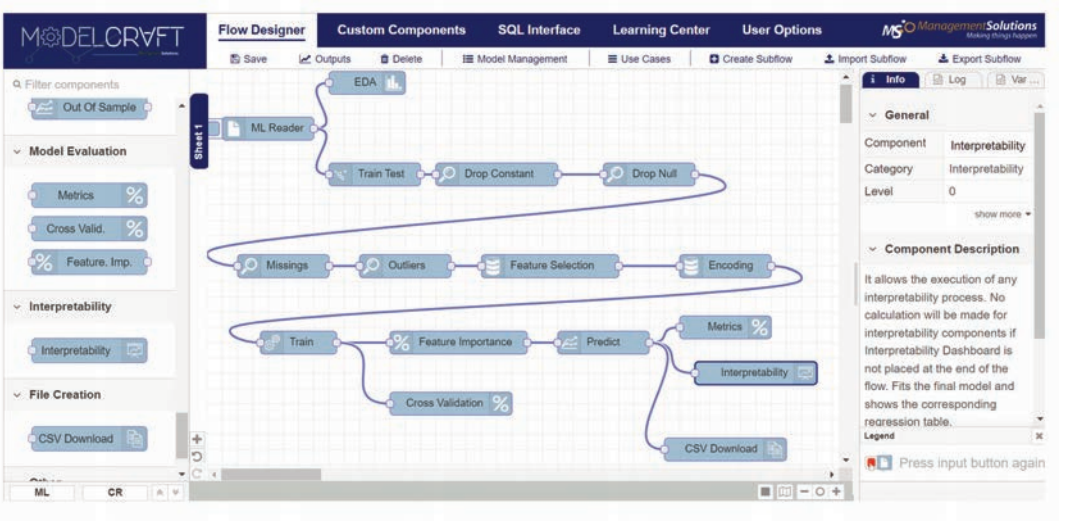
El enfoque del caso práctico será entrenar y validar diferentes modelos de inteligencia artificial para predecir el abandono de los empleados, utilizando las técnicas de XAI para analizar y comprender el comportamiento y las decisiones de los modelos seleccionados.

Para simplificar y agilizar el proceso, se ha empleado el sistema de modelización por componentes ModelCraft™, que contiene múltiples técnicas relevantes de AI y XAI. Este sistema permitirá realizar el estudio de forma eficiente y sin necesidad de escribir código.

[72] Alan Jay Perlis (1922-1990), informático estadounidense, doctor en Informática por el MIT y profesor de la Universidad de Purdue, la Universidad Carnegie Mellon y la Universidad de California en Berkeley, conocido por sus trabajos pioneros en lenguajes de programación y por ser el primer ganador del Turing Award.

[73] Kaggle (2017). IBM HR Analytics Employee Attrition & Performance.

Figura 13. Flujo de modelización en ModelCraft™.



A lo largo del desarrollo del caso práctico, se aplicarán las técnicas de interpretabilidad SHAP, LIME y PDP para analizar los modelos seleccionados y comprender qué variables influyen en la decisión de los empleados de abandonar su puesto de trabajo. Además, se explorará cómo estas variables interactúan entre ellas y cómo afectan a diferentes segmentos de la población de empleados.

Al finalizar el caso práctico, se evaluará la efectividad y las limitaciones de las técnicas de interpretabilidad utilizadas. También se discutirá cómo la combinación de modelos de inteligencia artificial y módulos de interpretabilidad puede mejorar la capacidad predictiva y la comprensión de los modelos, facilitando así la toma de decisiones basadas en datos en el ámbito empresarial.

Proceso de modelización

El proceso de modelización se realiza en tres fases: ingeniería de datos, modelización y análisis de interpretabilidad del modelo.

1. Ingeniería de datos

La ingeniería de datos es la fase inicial en la que se prepara y procesa el conjunto de datos para su uso en la creación de modelos de inteligencia artificial. En este caso, se realizan las siguientes acciones:

- Definición del ámbito de análisis: en este caso, se toma como población a todos los empleados que han sido baja en los últimos dos años.
- Limpieza de datos: se verifica la calidad de los datos y se eliminan o corrigen registros con información faltante o inconsistente.
- Transformación de variables: se convierten las variables categóricas en numéricas mediante técnicas como el *one-hot encoding* o el *ordinal encoding*. Además, se normalizan o estandarizan las variables numéricas cuando es necesario.
- Selección de variables: se identifican las variables más relevantes para predecir el abandono de los empleados utilizando técnicas de selección de variables como la correlación de Pearson, la importancia de las características en modelos basados en árboles o la eliminación recursiva de características.
- Construcción de variables: se generan nuevas variables a partir de las ya existentes para analizar si son mejores predictores del abandono de empleados, tales como la "satisfacción total", que se ha construido como suma de las puntuaciones de las variables "Satisfacción con el ambiente", "Satisfacción con el trabajo", "Calificación del rendimiento", "Equilibrio entre la vida personal y la laboral", "Involucración en el trabajo" y "Satisfacción con las relaciones laborales".
- División del conjunto de datos: se divide el conjunto de datos en dos subconjuntos: entrenamiento y prueba. El subconjunto de entrenamiento se utiliza para ajustar y optimizar los modelos de inteligencia artificial, mientras que el subconjunto de prueba se emplea para evaluar el rendimiento y la capacidad predictiva de los modelos.

2. Desarrollo del modelo

En esta fase, se entrenan y validan diferentes modelos de inteligencia artificial utilizando el subconjunto de entrenamiento. En concreto, se ajustan y comparan varios de los algoritmos de aprendizaje automático más comunes, como regresión logística, árboles de decisión, máquinas de vectores de soporte, redes neuronales y *random forest,* para seleccionar el modelo con el mejor rendimiento.

Para evitar el sobreentramiento y optimizar los hiperparámetros de los modelos, se emplean técnicas de validación cruzada y búsqueda en cuadrícula o aleatoria. Asimismo, se ha tenido especialmente en cuenta la complejidad del modelo durante el entrenamiento a la hora de seleccionar un algoritmo determinado, para facilitar su interpretación.

Para ello, se ha generado un flujo de desarrollo de modelos en ModelCraft™ (Fig. 13).

Para seleccionar el modelo con la mejor capacidad predictiva, se evalúa su rendimiento en el subconjunto de prueba utilizando métricas como la precisión, la sensibilidad, la especificidad y el área bajo la curva ROC (AUC-ROC). Estas métricas permiten evaluar la efectividad del modelo seleccionado en términos de su capacidad para predecir correctamente el abandono de empleados en datos no vistos previamente.

Considerado todo lo anterior, el algoritmo random forest arroja resultados de rendimiento superiores, aunque plantea un desafío de interpretabilidad a la hora de comprender sus predicciones. Este modelo ha considerado 300 árboles de decisión y ha arrojado una precisión del 75% y una sensibilidad del 84%. Por tanto, se trata de predicciones muy fiables y en pocas ocasiones se obtienen falsos negativos. Esto es relevante para este caso de estudio, en el que la compañía previsiblemente querría reducir lo máximo posible este tipo de error.

3. Análisis de interpretabilidad

En esta última fase, se aplican técnicas de interpretabilidad para analizar y comprender el comportamiento y las decisiones del modelo seleccionado. En concreto, los objetivos del análisis son:

- Entender qué variables son más importantes en la toma de decisiones para la compañía a nivel global, para lo que se ha empleado la comparación por importancia de cada variable.
- Entender cómo impactan cambios en las variables más importantes para distintos rangos de población.
- Entender los resultados del modelo en casos particulares donde se observa una determinada probabilidad de abandono.

En este caso práctico, se utilizan las técnicas de SHAP, LIME y PDP para explicar cómo el modelo toma decisiones y cómo las variables de entrada influyen en las predicciones.

SHAP permite obtener resultados de interpretabilidad global, que dan una interpretación de la importancia de cada variable, y LIME permite realizar un análisis intuitivo de interpretabilidad local que permita explicar el resultado del modelo para cada empleado partiendo de modelos lineales más sencillos. Como complemento, los gráficos PDP permiten visualizar cómo cambios en cada variable impactan en la predicción del modelo.

Con ello, se ha obtenido la siguiente distribución de la importancia de cada variable (Fig. 14).

En este caso, se observa que la variable con más importancia en la predicción de abandono (15,65%) es la "satisfacción total", un indicador sintético definido como una media ponderada de seis elementos (ambiente de trabajo, adecuación de funciones y áreas al puesto, rating interno, conciliación familiar, relación con compañeros y supervisores, y cargo y responsabilidad del empleado).

Este resultado es intuitivo y demuestra que la variable "satisfacción total" está bien diseñada. Sin embargo, las siguientes tres variables por importancia (antigüedad en la empresa, salario y distancia desde casa al puesto de trabajo) han demostrado tener una elevada influencia en el abandono de los empleados, que colectivamente duplica la del indicador "satisfacción total".

Figura 14. Interpretabilidad global del modelo random forest mediante SHAP, donde se utilizan los valores de Shapley para la obtención de la importancia de las variables.

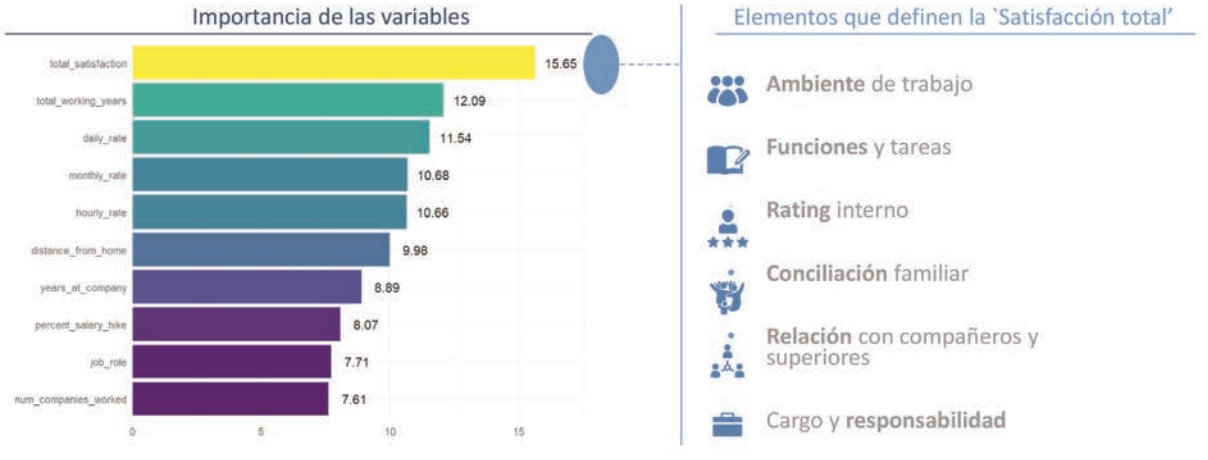
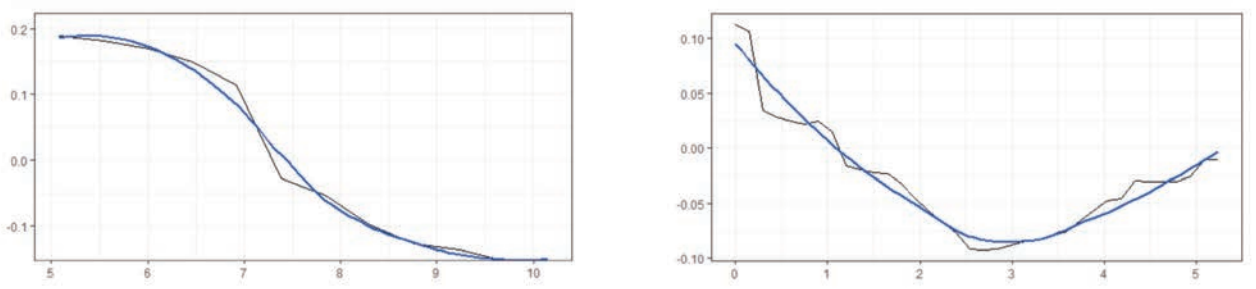


Figura 15. Gráficos PDP para las variables "satisfacción total" y "antigüedad en la empresa".



Para entender cómo influye cada variable individualmente, se han estudiado los PDP (fig. 15).

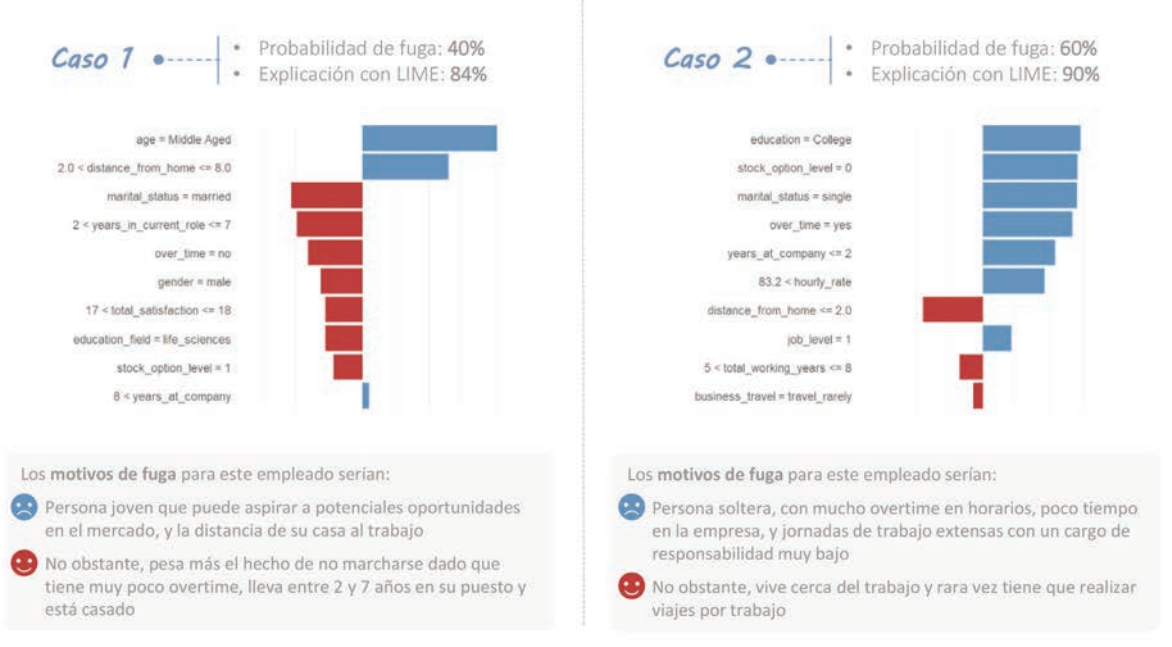
En la antigüedad en la empresa, se observa que a los tres años la tendencia se invierte: los empleados de antigüedad intermedia son, en promedio, los menos propensos a abandonar la compañía. Para la satisfacción total se observa una tendencia intuitiva: una mayor satisfacción reportada en las encuestas internas resulta en una menor tasa de abandono.

Para complementar el análisis anterior, se ha empleado LIME para analizar caso a caso los valores de las variables que influyen en la probabilidad de abandono de determinados empleados. La fig. 16 muestra dos empleados con distintas probabilidades de abandono obtenidas mediante el modelo. LIME muestra una métrica de explicabilidad que representa cómo de bueno es el ajuste lineal que ha obtenido mediante el modelo subrogado local para explicar estas predicciones.

Es destacable que las causas de abandono más relevantes en estos dos casos no corresponden necesariamente a las variables más influyentes a nivel global. Si bien se puede observar que la satisfacción total contribuye a explicar la probabilidad de abandono del empleado en el caso 1, no parece tener un impacto significativo en el caso 2, donde la probabilidad de abandono es mayor.

Esto refleja las dificultades existentes a la hora de interpretar este modelo, generalizable a modelos similares: pese a que la satisfacción total puede explicar de manera notable la probabilidad de abandono en promedio, esta conclusión es una generalización; se dan casos individuales y de colectivos en los que el abandono se explica en mayor medida por otras variables.

Figura 16. Interpretabilidad local del modelo random forest mediante LIME



Conclusiones del caso práctico

Del caso práctico de interpretabilidad en inteligencia artificial presentado se pueden extraen diversas conclusiones y lecciones aprendidas que pueden ser de utilidad en futuras aplicaciones de modelos de AI y XAI:

- **Aplicación del modelo:** la correcta aplicación e interpretación del modelo en este caso puede permitir anticipar y prevenir el abandono de empleados. Entre los usos que se pueden dar al modelo, destaca la capacidad de crear distintos perfiles con propensión al abandono e identificar las características de estos empleados con antelación para tomar las medidas adecuadas, lo que a largo plazo puede contribuir a reducir el nivel de rotación de la empresa.

- **Elección del modelo:** el proceso de modelización ha demostrado la importancia de comparar y validar diferentes algoritmos de aprendizaje automático para seleccionar el modelo con la mejor capacidad predictiva. En este caso, el modelo de *random forest* resultó ser el más adecuado para predecir el abandono de empleados.

- **Importancia de la interpretabilidad:** la aplicación de técnicas de interpretabilidad, como SHAP, LIME y PDP, ha proporcionado una comprensión más profunda de cómo el modelo toma decisiones y cómo las variables de entrada influyen en las predicciones. Esta información es crucial para validar la aplicabilidad del modelo en el contexto real y para garantizar que las predicciones se basen en características relevantes y significativas.

- **Variables influyentes:** el análisis de interpretabilidad ha permitido identificar las variables más relevantes para predecir el abandono de empleados. Estas variables pueden ser útiles para desarrollar estrategias de retención y mejorar la satisfacción laboral. Además, la comprensión de cómo estas variables interactúan entre ellas y cómo afectan a diferentes segmentos de la población de empleados puede enriquecer el análisis y facilitar la toma de decisiones basada en datos.

- **Implementación práctica:** el caso práctico demuestra la viabilidad y utilidad de aplicar técnicas de AI y XAI en un escenario realista, utilizando datos ficticios pero representativos de una situación empresarial. Este enfoque puede adaptarse a otros contextos y problemas empresariales, aprovechando las ventajas de la inteligencia artificial y la interpretabilidad para mejorar la toma de decisiones y obtener resultados más eficientes y efectivos.

- **Limitaciones:** al mismo tiempo, este caso de uso ha puesto de manifiesto las limitaciones y dificultades en la aplicación de las técnicas de interpretabilidad *post-hoc*. Es importante reconocer que los métodos de interpretabilidad no son infalibles y que, en ocasiones, pueden presentar resultados aproximados o parciales. Por lo tanto, es fundamental aplicar un enfoque crítico y riguroso al interpretar y validar los resultados de las técnicas de interpretabilidad.

- **Combinación de modelos de AI y módulos de interpretabilidad:** este caso práctico muestra cómo la integración de modelos de AI y módulos de interpretabilidad puede mejorar la capacidad predictiva y la comprensión de los modelos. Esto facilita la adopción de soluciones basadas en AI en la toma de decisiones empresariales.

- **Continuidad en el análisis de interpretabilidad:** por último, cabe destacar que el análisis de interpretabilidad no debe ser un ejercicio aislado aplicado durante el desarrollo de los modelos, sino que debe realizarse de manera continuada, reproducible y fiable a lo largo de toda la vida del modelo.

En conclusión, este caso práctico de interpretabilidad en inteligencia artificial ha proporcionado una experiencia valiosa en la aplicación de técnicas de AI y XAI en un contexto empresarial, y ha mostrado el potencial de la AI y la interpretabilidad para mejorar la toma de decisiones, al tiempo que ha revelado las limitaciones y dificultades asociadas con estas técnicas y la necesidad de un enfoque crítico y riguroso al interpretar y validar los resultados de la AI.