



UNIVERSIDAD TÉCNICA
FEDERICO SANTA MARÍA

DEPARTAMENTO
DE INFORMÁTICA



UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
VALPARAÍSO - CHILE



**“Análisis de deserción laboral en empresas mediante
técnicas de aprendizaje automatizado”**

Antonio Tognarelli Vargas

TESINA PARA OPTAR AL GRADO
DE MAGÍSTER EN TECNOLOGÍAS DE
LA INFORMACIÓN

Profesor Guía: José Luis Martí Lara



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título ☒ Tesis de Postgrado

Título del trabajo: Análisis de deserción laboral en empresas mediante técnicas de aprendizaje automatizado

Nombre del candidato(a): Antonio Alejandro Tognarelli Vargas

Carrera / Grado: Magister en tecnologías de la información

Campus: San Joaquín. **Departamento:** Informática.

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, José Luis Martí Lara, en mi calidad de profesor(a)

guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 26 / 10 / 2026

Firma:

Estudiante o Candidato(a):

Fecha: 26/10/2026

Firma:

Análisis de la Deserción laboral en empresas mediante técnicas de aprendizaje automatizado

Antonio Tognarelli Vargas¹

Antonio.flores@usm.cl

Resumen: El presente trabajo tiene como objetivo analizar y predecir la deserción laboral en empresas, mediante la aplicación de técnicas de análisis de datos y aprendizaje automático sobre información proveniente del área de Recursos Humanos. La investigación se basa en un conjunto de datos de acceso público, ampliamente utilizado en estudios académicos relacionado con la deserción laboral.

Para el desarrollo del estudio se adoptó la metodología **CRISP-DM**, permitiendo estructurar de manera sistemática las etapas de comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Se implementaron diversos algoritmos de aprendizaje supervisados con el propósito de identificar patrones asociados a la deserción laboral y evaluar su capacidad predictiva.

Los resultados obtenidos evidencian que el uso de modelos de aprendizaje automático constituye una herramienta eficaz para anticipar el riesgo de deserción laboral, facilitando la generación de información relevante para la toma de decisiones estratégicas. De este modo, el estudio aporta un enfoque analítico que contribuye a la gestión proactiva del capital humano, apoyando el diseño de estrategias orientadas a la retención del talento en las organizaciones

Palabras Clave: Deserción laboral, Aprendizaje Automatizado, CRISP-DM, Python, Recursos Humanos.

1 Introducción

1.1 Contexto y motivación

En los últimos años, la deserción laboral o rotación de personal se ha consolidado como uno de los principales desafíos para las organizaciones, debido a su impacto directo en los costos operativos, la productividad y el clima organizacional. Diversos estudios señalan que la rotación laboral se encuentra asociada a múltiples factores, entre los que destacan la insatisfacción laboral, percepción de falta de reconocimiento, sobrecarga de trabajo y ausencia de oportunidades de desarrollo profesional, los cuales influyen significativamente en la decisión de los trabajadores de abandonar la organización.

Este fenómeno ha generado una pérdida relevante de capital humano y experiencia organizacional, afectando especialmente a empresas que dependen de personal calificado y con conocimiento específico de sus procesos internos.

En Chile, según el Instituto Nacional de Estadísticas (INE, 2024), la Tasa de Rotación Laboral (TRL) alcanzó un **36,7 %**, llegando incluso a un **42,3 % en microempresas**. Estas cifras reflejan una problemática persistente en la gestión del talento humano, especialmente en sectores donde la demanda por personal calificado es alta [1]. Asimismo, el informe del Servicio Nacional de Capacitación y Empleo (SENCE, 2024) señala que las principales causas de rotación están asociadas a la falta de oportunidades de desarrollo, condiciones laborales percibidas como desfavorables y búsqueda de mejores expectativas salariales [1].

El contexto posterior a la pandemia evidenció aún más esta situación, destacando la necesidad de adoptar enfoques analíticos y tecnológicos que permitan **anticipar la deserción laboral** y **diseñar estrategias preventivas** para retener a los empleados más valiosos.

1.2 Definición del problema

Las organizaciones enfrentan actualmente un aumento sostenido en las tasas de rotación o deserción laboral, lo que genera costos asociados a la contratación, capacitación, pérdida de conocimiento institucional y disminución de la continuidad operativa. Sin embargo, muchos departamentos de Recursos Humanos no cuentan con herramientas predictivas que permitan identificar oportunamente a los empleados con mayor probabilidad de abandonar la organización.

La ausencia de mecanismos que anticipen este comportamiento limita la capacidad de las empresas para actuar de manera proactiva. Sin información basada en datos, las decisiones relacionadas con retención suelen ser reactivas y tardías, dificultando la implementación de estrategias efectivas para reducir la deserción laboral.

Dada esta necesidad, surge el desafío de desarrollar un modelo analítico que permita analizar información histórica de los empleados e identificar patrones asociados a la deserción laboral. Este enfoque posibilita evaluar variables como satisfacción laboral, desempeño, antigüedad, carga de trabajo y nivel salarial, con el fin de estimar la probabilidad de renuncia y apoyar la toma de decisiones en la gestión del capital humano.

1.3 Propuesta de solución e hipótesis

Frente a la necesidad de contar con herramientas analíticas que permitan a los departamentos de Recursos Humanos anticipar la deserción laboral, esta investigación propone el desarrollo de un modelo predictivo basado en técnicas de aprendizaje automático supervisado. La solución planteada consiste en analizar datos históricos de empleados con el fin de identificar patrones asociados a la rotación de personal y estimar la probabilidad de deserción de manera anticipada, proporcionando información relevante para la toma de decisiones estratégicas en la gestión del capital humano. Para el desarrollo de la solución propuesta se adopta la metodología CRISP-DM como marco de trabajo, la cual orienta de manera estructurada las etapas de análisis y modelado predictivo.

Hipótesis de investigación:

La aplicación de modelos de aprendizaje automático supervisado permite predecir la deserción laboral con un *Recall* superior a 0.90 sobre el conjunto de prueba, superando en al menos 20 puntos porcentuales a un modelo base de clasificación por mayoría. Este umbral se define considerando que la no detección temprana de un empleado en riesgo real de abandono constituye el error de mayor costo organizacional, por lo que la sensibilidad del modelo es la métrica crítica de evaluación.

1.4 Metodología aplicada

La metodología utilizada para el desarrollo de este estudio corresponde a **CRISP-DM** (*Cross Industry Standard Process for Data Mining*), un estándar ampliamente utilizado en proyectos de análisis de datos y modelado predictivo. Su estructura permite abordar el problema de manera sistemática, ordenada y replicable, alineando los objetivos del negocio con las decisiones técnicas del proceso analítico. Las seis fases de CRISP-DM, comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue, fueron aplicadas de manera iterativa durante el desarrollo del estudio, desde la identificación del problema de deserción laboral hasta la implementación del modelo predictivo final. La descripción detallada de cada fase y su aplicación práctica se presenta en el capítulo 3, correspondiente al desarrollo del proyecto.

1.5 Estructura del informe

El presente informe se organiza en los siguientes capítulos:

- **Capítulo 1: Introducción** Contextualiza la problemática, los objetivos, la hipótesis y la metodología aplicada.
- **Capítulo 2: Marco Teórico y Estado del Arte** Presenta los conceptos fundamentales, teorías y estudios relacionados con la deserción laboral y el uso de técnicas de aprendizaje automático en la gestión de Recursos Humanos.
- **Capítulo 3: Desarrollo del Proyecto** Describe la aplicación práctica del estudio, incluyendo la preparación de los datos, el modelado y la validación de los algoritmos utilizados.
- **Capítulo 4: Resultados Obtenidos** Expone los hallazgos cuantitativos y cualitativos, así como la comparación entre los modelos predictivos implementados.
- **Capítulo 5: Despliegue** Selección del modelo, valida la hipótesis.
- **Capítulo 6: Conclusiones** Sintetiza los principales resultados, plantea recomendaciones y proyecciones futuras.

2 Marco teórico y estado del Arte

En este capítulo se presentan los principales conceptos relacionados con la deserción laboral y la gestión del talento humano, así como los fundamentos teóricos que sustentan el uso de modelos predictivos basados en técnicas de aprendizaje automático. Además, se describen elementos clave como las bases de datos utilizadas y las clasificaciones de rotación, con el fin de contextualizar el problema y apoyar el desarrollo del modelo propuesto.

2.1 Deserción laboral

La deserción laboral o rotación se refiere al abandono voluntario de un empleado respecto de su puesto de trabajo dentro de una organización. Este fenómeno implica consecuencias económicas, operativas y sociales relevantes, debido a que conlleva pérdida de conocimiento, interrupción en los procesos internos y aumento en los costos asociados a la contratación y formación de nuevos trabajadores. Según Gutiérrez (2015), la deserción laboral afecta directamente la continuidad operativa y el cumplimiento de los objetivos organizacionales. Cuando las tasas de rotación aumentan, se intensifican las dificultades para retener talento, mantener equipos estables y garantizar una adecuada cohesión interna [2].

La deserción laboral o rotación puede clasificarse en distintos tipos, según el motivo o contexto del abandono. Los principales son [3]:

- **Deserción voluntaria:** Ocurre cuando el empleado renuncia por decisión propia, ya sea por una nueva oportunidad laboral, insatisfacción, falta de proyección o motivos personales.
- **Deserción involuntaria / Deseable:** Corresponde a desvinculaciones realizadas por la empresa, generalmente asociadas a bajo desempeño, incumplimientos o reestructuraciones internas.
- **Deserción funcional:** Se refiere a la salida de empleados cuyo desempeño no afecta significativamente la operación o puede generar beneficios indirectos.
- **Deserción disfuncional:** Implica la pérdida de trabajadores valiosos para la organización, generando impactos negativos en productividad y clima laboral.
- **Deserción inevitable:** Asociada a causas externas como jubilación, enfermedad o cambios vitales.
- **Deserción evitable:** ocurre cuando las razones de salida se relacionan con factores internos que la organización podría gestionar, como clima laboral, liderazgo o remuneraciones.

Estas distinciones permiten comprender mejor la complejidad del fenómeno y la necesidad de contar con herramientas analíticas que faciliten anticipar cuáles empleados presentan mayor probabilidad de abandonar la organización.

2.2 Metodologías de trabajo

En el ámbito del análisis de datos se han desarrollado diversas metodologías orientadas a guiar proyectos de análisis de datos y construcción de modelos predictivos. Entre las más utilizadas se encuentran KDD, SEMMA y CRISP-DM, las cuales proporcionan estructuras sistemáticas para la exploración, preparación y modelado de datos.

- **KDD (*Knowledge Discovery in Databases*):** Propuesta por Fayyad et al. (1996), KDD describe el proceso completo para la obtención de conocimiento a partir de datos. Sus etapas incluyen selección de datos, preprocesamiento, transformación, minería de datos e interpretación de resultados. Si bien corresponde a una metodología amplia orientada al descubrimiento de patrones, su enfoque es principalmente técnico y no contempla explícitamente la alineación con los objetivos del negocio [4].
- **SEMMA (*Sample, Explore, Modify, Model, Assess*):** Desarrollada por *SAS Institute*, SEMMA se centra en el modelado estadístico y finalización con la evaluación del modelo. Su enfoque está orientado principalmente a los procesos técnicos del análisis, sin incorporar explícitamente fases asociadas a la comprensión del negocio ni al despliegue de resultados en un contexto organizacional [5].
- **CRISP-DM (*Cross Industry Standard Process for Data Mining*):** CRISP-DM es considerada una de las metodologías más utilizadas en proyectos de análisis predictivo debido a su estructura flexible, iterativa y replicable. Está compuesta por seis fases [6]:
 - Comprensión del negocio: Definición del contexto, los objetivos y las necesidades de la organización.
 - Comprensión de los datos: Análisis exploratorio y evaluación de la calidad de los datos.
 - Preparación de los datos: Selección, limpieza y transformación de variables relevantes.
 - Modelado: Aplicación de algoritmos y configuración de parámetros del modelo predictivo.
 - Evaluación: Análisis del rendimiento de los modelos y comparación de resultados.
 - Despliegue: Presentación y aplicación de los resultados dentro del contexto empresarial.

A diferencia de KDD y SEMMA, CRISP-DM integra explícitamente los objetivos organizacionales dentro del proceso analítico y contempla el despliegue de resultados como una etapa formal. Estas características la convierten en la metodología más adecuada para el desarrollo del presente estudio, dado que permite relacionar el proceso técnico de construcción del modelo predictivo con la toma de decisiones en el área de Recursos Humanos.

Cada fase puede iterarse en función de los resultados obtenidos, garantizando un proceso flexible y repetible, tal como se observa en la figura 1.

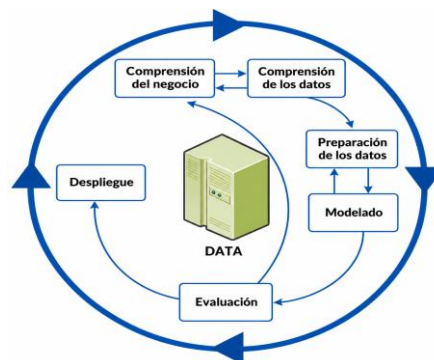


Figura 1. Fases del proceso CRISP-DM [7].

2.3 Aprendizaje automático (*Machine Learning*)

El aprendizaje automático es una rama de la inteligencia artificial que permite a los sistemas aprender de los datos y mejorar su desempeño sin necesidad de programación explícita. Mediante algoritmos especializados, es posible identificar patrones, realizar predicciones y apoyar la toma de decisiones en diversos contextos organizacionales.

Dentro del aprendizaje automático se distinguen distintos enfoques, entre los cuales destacan [8]:

- **Aprendizaje supervisado:**
Corresponde al enfoque en el cual los modelos se entrenan utilizando datos de entrada asociados a una variable objetivo conocida. Su finalidad es aprender una función que permita predecir dicha variable a partir de nuevas observaciones. En este enfoque se emplean técnicas de clasificación y regresión, siendo ampliamente utilizados en problemas donde se busca predecir comportamientos específicos, como la deserción laboral.
- **Aprendizaje no supervisado:**
En este enfoque los datos no cuentan con etiquetas, por lo que el sistema busca descubrir patrones o estructuras ocultas. Entre las técnicas más utilizadas se encuentran los métodos de agrupamiento, como *k-means*, el cual corresponde a un algoritmo de *clustering* basado en centroides, y los métodos de reducción de dimensionalidad, como análisis de componentes principales (PCA).
- **Aprendizaje por refuerzo:**
Consiste en entrenar a un agente que interactúa con un entorno, aprendiendo a partir de recompensas y penalizaciones. Este enfoque se utiliza comúnmente en optimización de procesos y toma de decisiones en tiempo real.

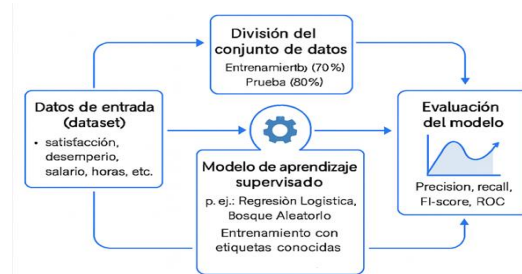


Figura 2. Estructura general del proceso de aprendizaje automático supervisado [8].

La Figura 2 presenta el flujo general de un proceso de aprendizaje automático supervisado. Este incluye las etapas fundamentales para construir un modelo predictivo, comenzando con la preparación del conjunto de datos, la división entre entrenamiento y prueba, el entrenamiento del algoritmo seleccionado y la posterior evaluación mediante métricas de desempeño como precisión, *recall*, *F1-score* y el área bajo la curva ROC.

2.3.1 Métricas de evaluación en modelos supervisados predictivos

En problemas de clasificación binaria, donde la variable objetivo representa dos clases mutuamente excluyentes como ocurre en la predicción de la deserción laboral (*left* = 0 / 1) la evaluación del rendimiento de los modelos predictivos requiere el uso de métricas específicas que permitan analizar su capacidad para discriminar correctamente entre ambas clases [9].

El rendimiento de un modelo de aprendizaje supervisado se entiende como su capacidad para generar predicciones correctas y útiles desde una perspectiva analítica y organizacional, considerando tanto la precisión global como el impacto asociado a los distintos tipos de error. En este contexto, las métricas de evaluación se basan en la comparación entre clases reales y las clases predichas por el modelo, permitiendo analizar su comportamiento frente a casos positivos y negativos.

En el contexto de la deserción laboral, donde la no detección de un empleado con riesgo real de abandono puede generar costos organizacionales significativos, la selección e interpretación adecuada de estas métricas adquiere especial relevancia para apoyar la toma de decisiones estratégicas en la gestión del talento humano.

Las métricas utilizadas en el presente estudio son [9]:

- **Accuracy** (Precisión global): Mide el porcentaje total de predicciones correctas. Es útil como primera aproximación global, pero puede resultar insuficiente cuando existe desbalance entre clases.
- **Precision** (Precisión positiva): Corresponde a la proporción de casos predichos como “renuncia” que realmente pertenecen a esa clase. Es importante cuando se busca evitar falsas alarmas o clasificaciones incorrectas de empleados considerados de riesgo.
- **Recall** (Sensibilidad o Tasa de Verdaderos Positivos): Mide la capacidad del modelo para identificar correctamente a los empleados que efectivamente renunciaron. Es una métrica crítica en escenarios donde no detectar un caso positivo puede tener un costo elevado para la organización.
- **F1-Score**: Es el promedio armónico entre *Precision* y *Recall*. Resulta especialmente útil cuando existe un leve desbalance de clases, ya que permite evaluar simultáneamente la calidad de las predicciones positivas y la capacidad de detección del modelo.
- **Matriz de Confusión**: Resume los aciertos y errores del modelo mediante cuatro elementos: verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Permite analizar detalladamente la naturaleza de los errores y apoyar decisiones sobre ajustes en umbrales o estrategias de clasificación.
- **Curva ROC y Área Bajo la Curva (AUC-ROC)**: La curva ROC representa la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos a distintos umbrales. Su área bajo la curva (AUC) indica la capacidad global del modelo para discriminar entre empleados que renuncian y los que permanecen. Valores cercanos a 1 indican un desempeño superior.

El uso conjunto de estas métricas permite evaluar no solo el rendimiento general de un modelo, sino también su estabilidad, precisión en la identificación de patrones relevantes y capacidad para prever correctamente la deserción laboral. Además, constituyen la base metodológica para la comparación de los algoritmos utilizados, donde se analizan los resultados obtenidos a partir del conjunto de datos real. En este sentido, las métricas de evaluación conforman un componente fundamental dentro del aprendizaje supervisado predictivo, ya que permiten determinar qué modelo presenta el mejor equilibrio entre desempeño, sensibilidad, interpretabilidad y capacidad discriminativa, aspectos esenciales para su aplicación en contextos organizacionales y de Recursos Humanos.

2.3.2 Algoritmos supervisados de clasificación

Los algoritmos supervisados de clasificación permiten aprender a partir de datos previamente etiquetados para asignar nuevas observaciones a una clase determinada. En problemas de clasificación binaria, como la predicción de deserción laboral, estos modelos permiten estimar si un empleado pertenece a la clase asociada a permanencia o deserción. Entre los principales enfoques se encuentra regresión logística, utilizada para clasificación lineal; *K-Nearest Neighbors (KNN)*, basado en la similitud entre observaciones; los Árboles de decisión, que generan reglas jerárquicas de clasificación; Bosque Aleatorio (*Random Forest*) [10], método de ensamble que combina múltiples árboles de decisión; y las Maquinas de Vectores de soporte (*SVM*), capaces de establecer fronteras de decisiones lineales o no lineales [9].

2.3.3 Interpretación de la matriz de confusión

La matriz de confusión es una herramienta fundamental para evaluar el rendimiento de modelos supervisados de clasificación, ya que permite analizar de manera detallada los aciertos y errores del algoritmo en función de

las clases reales y las clases estimadas por el modelo. En problemas de clasificación binaria, como la predicción de la deserción laboral, Esta matriz se estructura a partir de cuatro categorías:

- **Verdaderos positivos (TP):** Empleados correctamente clasificados como desertores.
- **Verdaderos negativos (TN):** Empleados correctamente clasificados como permanentes.
- **Falsos positivos (FP):** Empleados clasificados como desertores, aunque permanecieron en la organización.
- **Falsos negativos (FN):** Empleados clasificados como permanentes, aunque finalmente desertaron.

Desde una perspectiva organizacional, los falsos negativos (FN) representan el error más crítico en estudios de deserción laboral, ya que implican no identificar a tiempo a empleados que presenta alto riesgo de abandono. Este tipo de error puede traducirse en pérdida de talento, problemas operativos y mayores costos de reemplazo, por ello, métricas como *Recall* y *F1-score*, que consideran explícitamente la detección de los casos positivos, adquieren especial relevancia en este tipo de problemas. En contraste, los falsos positivos (FP) implican clasificar erróneamente como en riesgo de deserción a empleados que finalmente permanecen en la organización. Si bien este error puede generar costos asociados a intervenciones innecesaria, su impacto suele ser menor en comparación con la no detección de un abandono real. La interpretación adecuada de la matriz de confusión permite comprender el comportamiento del modelo más allá de una métrica global de desempeño, facilitando el análisis del equilibrio entre los distintos tipos de error y apoyando la selección de los modelos que resulten más adecuados para contextos organizacionales orientados a la retención del talento humano.

2.4 Herramientas analíticas aplicadas a Recursos Humanos

El uso de herramientas basadas en datos ha adquirido un papel fundamental en la transformación de los procesos de gestión del talento dentro de las organizaciones. Estas tecnologías permiten analizar grandes volúmenes de información, identificar patrones relevantes y apoyar la toma de decisiones estratégicas orientadas a mejorar la eficiencia operativa y la experiencia de los empleados. En este contexto, la incorporación de soluciones de analítica avanzada y técnicas de aprendizaje automático ha fortalecido la capacidad de los departamentos de Recursos Humanos para comprender mejor el comportamiento laboral, anticipar riesgos y diseñar intervenciones más efectivas. La utilización de tecnologías basadas en datos contribuye a optimizar procesos de contratación, mejorar la comprensión de los patrones de desempeño y aumentar el compromiso de los empleados. Asimismo, estas herramientas facilitan la toma de decisiones más informadas y apoyan la implementación de estrategias orientadas a disminuir la deserción laboral y fortalecer la retención del talento. En conjunto, estos beneficios evidencian el impacto positivo que pueden tener las herramientas analíticas en la gestión de Recursos Humanos, justificando la relevancia de desarrollar modelos predictivos que permitan anticipar el riesgo de deserción laboral, como el propuesto en este estudio [11].

2.5 Tecnologías y herramientas para el desarrollo del modelo predictivo

En el ámbito de la analítica de datos y el aprendizaje automático existen diversas herramientas tecnológicas que apoyan el desarrollo de modelos predictivos y la visualización de resultados. Estas herramientas abarcan desde entornos de programación hasta plataformas de análisis visual, y su elección depende del contexto del problema, los objetivos del proyecto y el nivel de control requerido sobre el proceso. Entre las herramientas comúnmente utilizadas se encuentran lenguajes de programación como *Python* [12] y *R* [13], *plataformas de análisis como RapidMiner* [14] y *Orange3* [15], y *soluciones orientadas a la visualización y resultados, tales como Tableau* [16], *Flourish* [17] y *Power Bi* [18]. Estas herramientas permiten abordar distintas etapas del ciclo analítico, tales como la exploración de datos, el modelado predictivo y la presentación de información para la toma de decisiones. La literatura reporta un uso extendido de estas tecnologías en proyectos de ciencia de datos y minería de datos, destacando su aporte en términos de flexibilidad, escalabilidad y facilidad de uso, dependiendo del enfoque metodológico adoptado.

Python: es uno de los lenguajes de programación más utilizados en ciencia de datos debido a su versatilidad, sintaxis clara y extenso ecosistema de librerías especializadas. Su capacidad para manejar grandes volúmenes de datos y ejecutar algoritmos de aprendizaje automático de manera eficiente lo convierte en una herramienta fundamental para estudios analíticos [12].

Bibliotecas de interés

En el ecosistema de Python, la literatura destaca diversas bibliotecas ampliamente utilizadas en proyectos de análisis de datos y aprendizaje automático, entre las cuales se encuentran [19]:

- **Pandas:** Se utiliza para la manipulación, limpieza y exploración estructurada del conjunto de datos, permitiendo realizar operaciones eficientes sobre *DataFrames*.
- **Scikit-learn:** Paquete central para la implementación de algoritmos de aprendizaje supervisado, división del *dataset*, preprocesamiento, construcción de modelos y evaluación mediante métricas estándar.
- **Matplotlib y Seaborn:** Herramientas de visualización que permiten generar gráficos exploratorios, curvas ROC, matrices de confusión y comparaciones de desempeño entre modelos.
- **NumPy:** Es utilizada para operaciones numéricas y manejo eficiente de arreglos, facilitando el cálculo de métricas y transformaciones internas del modelo.

Jupyter Notebook: Es un entorno de trabajo interactivo de código abierto, se utiliza en el mundo de la ciencia de datos y análisis de datos, este permite integrar en un mismo documento código ejecutable, gráficos y texto explicativo. Esta característica facilita la trazabilidad de cada etapa del análisis, mejora la claridad metodológica y favorece la replicabilidad del proceso [20].

2.6 Trabajos relacionados

Diversos estudios han investigado la aplicación del aprendizaje automático para predecir la deserción laboral en distintos sectores y contextos organizacionales. En términos generales, la literatura coincide en que los modelos supervisados de clasificación, tales como Regresión Logística, Árboles de Decisión, Bosques Aleatorios y Máquinas de Soporte Vectorial, presentan un alto rendimiento para anticipar comportamientos asociados al abandono laboral.

Uno de los estudios más citados en esta área es el de **Saradhi y Palshikar (2011)** [21], quienes desarrollaron un modelo de predicción de rotación en una empresa del sector *retail* utilizando algoritmos como Árboles de Decisión, *Naïve Bayes* y Regresión Logística. Los autores demostraron que los modelos supervisados pueden identificar patrones de comportamiento que permiten anticipar la salida de empleados con una alta precisión, resaltando la importancia de variables como desempeño, remuneración y antigüedad.

De manera complementaria, **Fallucchi et al. (2020)** [8] evaluaron múltiples algoritmos de aprendizaje automático —incluyendo *Random Forest*, SVM y redes neuronales— para predecir la deserción laboral en un conjunto de datos de Recursos Humanos. Los resultados mostraron que los modelos basados en ensambles, particularmente *Random Forest*, lograron el mejor rendimiento al manejar interacciones complejas y variables heterogéneas propias del ámbito organizacional.

En el contexto latinoamericano, **Sierra Buriticá (2022)** [22] realizó un estudio aplicado a una empresa del sector de software en Colombia, comparando modelos como *Naïve Bayes*, Regresión Logística, Árbol de Decisión y *Random Forest*. Este último destacó nuevamente por su capacidad para identificar variables determinantes en la deserción, tales como satisfacción laboral, compensación y oportunidades de crecimiento, validando su aplicabilidad en escenarios empresariales reales.

Asimismo, **Živković et al. (2024)** [23] estudiaron la intención de rotación utilizando principalmente Árboles de Decisión como técnica predictiva. Los resultados mostraron que este algoritmo es especialmente útil para problemas en los que la interpretabilidad es relevante, ya que permite visualizar reglas de decisión asociadas a factores como compromiso organizacional y satisfacción en el puesto de trabajo.

Finalmente, **Rana et al. (2024)** [24] compararon algoritmos como KNN, Regresión Logística, *Naïve Bayes* y *Random Forest* en la predicción de deserción laboral. El estudio evidenció que los modelos de clasificación pueden representar una herramienta estratégica para los departamentos de Recursos Humanos, facilitando la identificación temprana de empleados con mayor probabilidad de abandonar la organización.

En conjunto, estos estudios demuestran que las técnicas de aprendizaje automático supervisado constituyen un enfoque eficaz para anticipar la deserción laboral en distintos sectores. Además, refuerzan la pertinencia del uso de modelos como *Random Forest* en el presente estudio, dado su desempeño superior en la predicción y su capacidad para analizar la importancia relativa de las variables que influyen en la rotación del personal. Estos antecedentes demuestran la efectividad de los modelos supervisados para anticipar la deserción en distintos contextos, reforzando la pertinencia de su aplicación en el ámbito laboral.

3 Desarrollo del Proyecto

El presente capítulo detalla el proceso completo de construcción del modelo predictivo de deserción laboral, siguiendo la metodología CRISP-DM como marco estructural. Cada sección describe las fases del ciclo analítico, desde la comprensión del problema organizacional hasta la evaluación del modelo final. Este enfoque permite documentar de manera ordenada y reproducible las decisiones técnicas adoptadas durante el desarrollo, garantizando la coherencia metodológica del estudio.

Enfoque metodológico y propuesta de solución

El desarrollo del presente estudio se basa en la metodología CRISP-DM, ampliamente utilizada en proyectos de análisis predictivo y minería de datos debido a su estructura iterativa, flexible y orientada a la toma de decisiones en contextos organizacionales. Los fundamentos conceptuales de esta metodología y la descripción general de sus fases fueron presentados previamente en el capítulo 2.2. En el contexto de esta investigación, CRISP-DM se adopta como marco metodológico para estructurar la construcción del modelo predictivo de deserción laboral, permitiendo desarrollar el proceso analítico de manera sistemática y coherente con los objetivos definidos. La propuesta de solución consiste en construir un modelo capaz de estimar la probabilidad de deserción laboral a partir de variables asociadas al desempeño, satisfacción y características contractuales de los empleados. Este enfoque metodológico permite que el proceso de análisis avance de forma ordenada desde la identificación del problema organizacional hasta la evaluación del modelo predictivo, siguiendo el flujo descrito en el capítulo 3.1.1. De este modo, se garantiza la alineación entre las etapas y las decisiones técnicas adoptadas durante el desarrollo del proyecto.

3.1.1 Flujo metodología adoptado

El presente estudio adopta la metodología CRISP-DM como marco estructural para el desarrollo del modelo predictivo. La figura 3 presenta como fue implementado el caso, en ella se muestran las seis fases aplicadas para la construcción del modelo predictivo de deserción laboral y las principales herramientas tecnológicas utilizadas en cada etapa.

El proceso se desarrolló de manera iterativa permitiendo retroalimentar las fases anteriores a partir de los resultados obtenidos durante el modelado y evaluación.

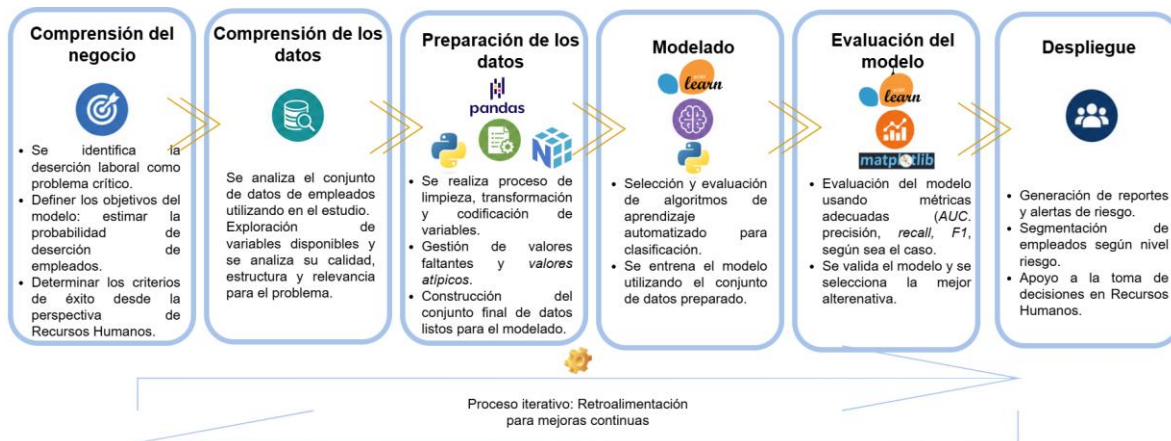


Figura 3. Flujo de la metodología y tecnologías adoptadas. Fuente: Elaboración propia

3.2 Comprensión del negocio

La deserción laboral constituye uno de los desafíos más relevantes para las organizaciones modernas, ya que impacta directamente en la continuidad operativa, la productividad y la capacidad competitiva de la empresa. A nivel de gestión de personas, la rotación implica la pérdida de talento, conocimiento institucional y experiencia acumulada, elementos fundamentales para sostener el desempeño organizacional en el largo plazo. La literatura distingue entre rotación voluntaria cuando el empleado renuncia por decisión propia y rotación involuntaria asociada a desvinculaciones por necesidades de la empresa. La rotación voluntaria es especialmente crítica, pues generalmente involucra a trabajadores con competencias difíciles de reemplazar o con experiencia clave para la operación. La salida de un trabajador genera una serie de costos directos e indirectos que incluyen procesos de reclutamiento y selección, inducción y capacitación, adaptación al puesto, disminución temporal del rendimiento y redistribución de tareas. Además, puede provocar sobrecarga laboral en los equipos, afectar el clima organizacional y debilitar la cohesión interna. Desde una perspectiva estratégica, una alta rotación dificulta la retención de talento, afecta la continuidad del conocimiento técnico y debilita la capacidad de la organización para ejecutar planes de largo plazo. Asimismo, compromete la eficiencia operativa y la estabilidad de los equipos, incidiendo en la calidad del servicio y en la satisfacción de los clientes internos y externos. Frente a este escenario, resulta indispensable que el área de Recursos Humanos cuente con herramientas que permitan anticipar la posibilidad de deserción antes de que ocurra. Las metodologías tradicionales, como encuestas de clima laboral o análisis retrospectivos, suelen ofrecer información tardía o parcial, lo que limita la capacidad de intervención temprana. El uso de técnicas de análisis predictiva y aprendizaje automático permite transformar datos históricos en información accionable, identificando patrones y factores de riesgo asociados al comportamiento de los empleados. Por su aporte al análisis de riesgo individual y la generación de conocimiento organizacional, un modelo predictivo se convierte en una herramienta clave para gestionar de manera proactiva la deserción laboral y fortalecer la toma de decisiones estratégicas en la organización.

3.3 Comprensión de los datos

En esta fase se analiza en detalle el conjunto de datos utilizado para el desarrollo del modelo predictivo. El estudio emplea el *HR Analytics – Employee Turnover Dataset* [25], el cual contiene información estructurada sobre características laborales y comportamentales de los empleados.

Este *dataset* está compuesto por **14.999 registros** y **10 variables** cuantitativas y categóricas, que reflejan aspectos relevantes de la dinámica laboral, tales como nivel de satisfacción, evaluación de desempeño, carga de

trabajo mensual, antigüedad en la empresa, promociones recientes, ocurrencia de accidentes, número de proyectos asignados, departamento y nivel salarial. La variable objetivo *left* indica si el empleado renunció (1) o permaneció en la organización (0). El conjunto de datos fue diseñado con fines académicos y de investigación, con el propósito de facilitar la construcción y validación de modelos de predicción de deserción laboral. Su formato estandarizado y libre de valores faltantes permite realizar análisis comparativos y aplicar algoritmos de aprendizaje automático sin requerir procesos extensivos de limpieza. A pesar de sus ventajas metodológicas, es importante señalar que este *dataset* no representa la complejidad total de una organización real, ya que no incorpora factores cualitativos asociados al entorno laboral, como liderazgo, clima organizacional o satisfacción intrínseca. Por ello, su uso en este estudio se orienta principalmente a la validación metodológica del enfoque predictivo, más que a la representación exacta de un escenario empresarial específico.

3.3.1 Identificación y clasificación de variables

Con el propósito de comprender la estructura del conjunto de datos utilizado en este estudio, se realizó una clasificación de las variables según su función dentro del modelo predictivo y su naturaleza (numérica o categórica). Esta diferenciación es esencial para definir las técnicas de preparación de datos, así como los algoritmos de aprendizaje supervisado que se aplicarán posteriormente.

El *dataset* empleado contiene una combinación de variables cuantitativas y cualitativas que describen distintos aspectos del comportamiento laboral de los empleados. La única variable dependiente corresponde a la deserción laboral, mientras que el resto de las variables actúan como predictores del fenómeno. La Tabla 1 resume la clasificación general de las variables.

Tabla 1. Clasificación de variables del conjunto de datos

Tipo de Variable	Nombre	Descripción
Dependiente (objetivo, y)	<i>left</i>	1 si el empleado renunció, 0 en caso contrario
Independiente (numérica, x)	<i>satisfaction_level</i>	Nivel de satisfacción del empleado (0 - 1)
Independiente (numérica, x)	<i>last_evaluation</i>	Última evaluación de desempeño (0 - 1)
Independiente (numérica, x)	<i>number_project</i>	Número de proyectos asignados
Independiente (numérica, x)	<i>average_monthly_hours</i>	Promedio de horas trabajadas al mes
Independiente (numérica, x)	<i>time_spend_company</i>	Años de antigüedad en la empresa
Independiente (categórica, x)	<i>salary</i>	Nivel de salario (bajo, medio, alto)
Independiente (numérica, x)	<i>Work_accident</i>	Indica si el empleado ha tenido un accidente laboral (1 - 0)
Secundaria (categórica, x)	<i>promotion_last_5years</i>	Promoción en los últimos 5 años
Secundaria (categórica, x)	<i>department</i>	Departamento del empleado

Variables predictoras (x) y variable objetivo (y):

En el contexto del aprendizaje supervisado, la variable *left* se define como la **variable dependiente (y)**, ya que representa el resultado que el modelo debe predecir. Su valor indica el estado final del empleado respecto a la organización y constituye el objetivo del proceso de modelado. Las variables restantes actúan como **variables independientes (x)** o predictoras.

Estas variables corresponden a características asociadas al desempeño, carga laboral, satisfacción, historial de promociones y nivel salarial del empleado. Su función es proporcionar información relevante que permita al modelo estimar la probabilidad de deserción laboral. Diferenciar correctamente entre variables **x** e **y** permite estructurar adecuadamente el proceso de modelado, definir las transformaciones necesarias y asegurar la coherencia metodológica dentro del marco CRISP-DM.

Consideraciones sobre el alcance del *dataset*

Aunque la clasificación anterior permite estructurar adecuadamente el proceso de modelado, es importante reconocer que el conjunto de datos utilizado posee un carácter académico y no incluye ciertas variables cualitativas habitualmente relevantes en estudios reales de rotación laboral, como clima organizacional, liderazgo, satisfacción intrínseca o percepción de justicia interna. La ausencia de estos factores limita parcialmente la capacidad explicativa del modelo en contextos empresariales reales. Por ello, los resultados obtenidos deben interpretarse como una validación metodológica del enfoque predictivo, más que como una representación exhaustiva del fenómeno de deserción en organizaciones específicas.

3.3.2 Comprensión inicial del *dataset*

Se realizó un análisis exploratorio orientado a identificar patrones, tendencias y relaciones entre las variables disponibles. Este análisis permitió identificar asociaciones relevantes entre la satisfacción laboral, la carga de trabajo y la probabilidad de deserción. La matriz de correlación evidenció que la variable *satisfaction_level* presenta la relación negativa más significativa con *left* ($r = -0.39$), lo que indica que niveles más bajos de satisfacción se asocian directamente con un mayor riesgo de renuncia. Este comportamiento coincide con la evidencia empírica reportada en estudios sobre rotación voluntaria, en los cuales la satisfacción laboral emerge como uno de los predictores más influyentes del abandono organizacional.

Asimismo, variables como *average_monthly_hours* y *last_evaluation* mostraron correlaciones positivas entre sí y con *number_project*, sugiriendo que una mayor carga laboral y un desempeño elevado suelen estar asociados a una mayor cantidad de proyectos asignados. Aunque estas relaciones no presentan una correlación directa fuerte con la deserción, permiten comprender mejor la dinámica de trabajo y las exigencias operacionales dentro de la organización. Por otro lado, *time_spend_company* presentó una correlación positiva leve con *left*, lo que sugiere que una mayor antigüedad puede estar asociada a un mayor desgaste acumulado, aumentando eventualmente la probabilidad de renuncia. Si bien estas correlaciones no son elevadas, su comportamiento es consistente con la literatura que señala que ciclos prolongados sin movilidad o cambios internos pueden influir en la intención de abandono. Adicionalmente, se constató un leve desbalance en la variable objetivo (*left*), característico en problemas de clasificación binaria. Aunque no representa un desbalance severo.

Este aspecto fue considerado al interpretar las métricas de desempeño para evitar sesgos en la evaluación de los modelos predictivos.

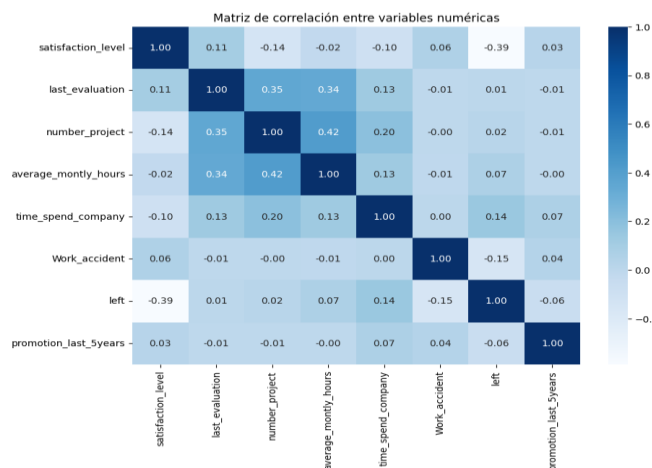


Figura 4. Matriz de correlación entre variables. Fuente: Elaboración propia

En la Figura 4 se presenta la matriz de correlación correspondiente al análisis exploratorio. Los patrones observados permitieron fortalecer la comprensión del comportamiento interno del *dataset* y orientar de forma

adecuada las decisiones tomadas en las etapas de preparación y modelado dentro del marco metodológico CRISP-DM.

3.4 Preparación de los datos

La fase de preparación de los datos tiene como objetivo garantizar que el conjunto de datos esté en condiciones adecuadas para el modelado.

En esta etapa se ejecutaron los siguientes procesos:

- Verificación de integridad del *dataset* (sin valores nulos ni duplicados).
- Aplicación de las transformaciones definidas previamente, incluyendo la codificación de variables categóricas.
- Construcción de la matriz de características (**x**) y el vector objetivo (**y**).
- Revisión del leve desbalance de clases en la variable *left*.

Como resultado de este proceso, se obtuvo un conjunto de datos limpio, estructurado y preparado para la fase de modelado. La configuración experimental utilizada para el entrenamiento y evaluación de los algoritmos, incluyendo la división en conjuntos de entrenamiento y prueba, se describe en la sección 3.5.4.

Codificación de variables categóricas

Para permitir el uso de algoritmos de aprendizaje supervisado, fue necesario transformar las variables categóricas del *dataset* en valores numéricos. Este procedimiento es fundamental, ya que los modelos de clasificación no pueden operar directamente con datos textuales. Las variables categóricas presentes fueron *salary* y *department*, las cuales se codificaron mediante un esquema de asignación numérica simple.

Las Tablas 2 y 3 presentan el resultado del proceso de codificación aplicado a las variables categóricas del *dataset*. Esta transformación permitió convertir los valores textuales en representaciones numéricas compatibles con los algoritmos de clasificación utilizados en el estudio. Con este paso, El conjunto de datos quedó completamente preparado para avanzar hacia la fase siguiente dentro del marco metodológico CRISP-DM.

Tabla 2. Codificación de salario

Categoría	Código
<i>High</i>	0
<i>Low</i>	1
<i>Medium</i>	2

Tabla 3. Codificación de departamento

Categoría	Código
<i>It</i>	0
<i>RandD</i>	1
<i>Accounting</i>	2
<i>Hr</i>	3
<i>Management</i>	4
<i>Marketing</i>	5
<i>Product mmg</i>	6
<i>Sales</i>	7
<i>Support</i>	8
<i>Technical</i>	9

3.4.1 Relevancia técnica de las variables en modelos supervisados predictivos

Dentro del proceso de modelado, algunos algoritmos supervisados, particularmente Bosque Aleatorio, generan una estimación de la importancia relativa de cada variable predictora. Esta medida representa el aporte de cada atributo en la reducción del error durante el proceso de entrenamiento del modelo.

Desde una perspectiva metodológica, la importancia de variables no constituye un resultado final en sí mismo, sino una herramienta técnica que permite interpretar el comportamiento del modelo y comprender que

características del trabajador influyen con mayor peso en la predicción de la deserción laboral. En este estudio, la estimación de la importancia de variables no fue utilizada como un criterio de filtrado previo al modelado, y se utilizó exclusivamente como un mecanismo de análisis e interpretación posterior. Esta decisión se adoptó con el objetivo de evaluar el desempeño predictivo del modelo considerando el conjunto completo de variables disponibles, evitando introducir sesgos derivados de una selección temprana de atributos y permitiendo una comparación consistente entre los distintos algoritmos implementados. Los valores de importancia obtenidos se presentan y analizan en el capítulo 4, una vez entrenado el modelo definitivo, como apoyo a la interpretación de los resultados y la comprensión del fenómeno deserción laboral desde una perspectiva analítica y organizacional. Para complementar la interpretación de los modelos de clasificación, se consideraron dos métodos de análisis de importancia de variables: *Gini Importance* y *Permutation Importance*. El primero estima la importancia de cada variable a partir de su contribución a la reducción de impureza en los árboles, mientras que el segundo evalúa la disminución del desempeño predictivo al permutar aleatoriamente sus valores. Ambas técnicas fueron aplicadas posteriormente al modelo de clasificación seleccionado, con el propósito de analizar la contribución relativa de las variables a sus predicciones.

3.5 Modelado

La fase de modelado corresponde al proceso en que se seleccionan, entrenan y comparan distintos algoritmos de aprendizaje automático supervisado con el fin de construir un modelo predictivo capaz de estimar la probabilidad de deserción laboral. De acuerdo con la metodología CRISP-DM, esta etapa se desarrolla después de comprender el negocio, analizar los datos y preparar el conjunto final de variables, permitiendo aplicar de manera práctica los algoritmos seleccionados y evaluar su desempeño.

Antes de detallar los algoritmos utilizados. Este flujo resume las etapas operativas aplicadas en el entorno Python, desde la carga de datos hasta la evaluación final del modelo, y sirve como guía para comprender la secuencia de pasos que estructuran el proceso. Se presenta en la Figura 5 el flujo general del proceso de modelado predictivo implementado en este estudio.

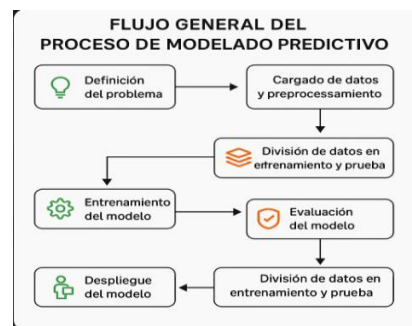


Figura 5. Flujo general del proceso del modelo predictivo. [7]

3.5.1 Selección de algoritmos

La selección se realizó considerando la diversidad de técnicas de clasificación y su utilización en estudios previos sobre deserción laboral revisados en la sección 2.6. Este criterio permite comparar métodos con características predictivas diferentes bajo un mismo conjunto de datos, facilitando una evaluación consistente de su desempeño.

- **Regresión Logística (Logistic Regression):** Modelo lineal de referencia para clasificación binaria. Permite interpretar la influencia de cada variable y sirve como línea base para comparar el desempeño del resto de los algoritmos.
- **K-Nearest Neighbors (KNN):** Algoritmo basado en la similitud entre casos. Adecuado para capturar patrones locales, aunque sensible a la normalización y a valores atípicos.

- **Árbol de Decisión** (*Decision Tree Classifier*): Enfoque basado en una familia de modelos de tipo árbol que generan reglas jerárquicas para la toma de decisiones. Su principal ventaja radica en su alta interpretabilidad y en su capacidad para identificar interacciones no lineales entre variables.
- **Bosque Aleatorio** (*Random Forest*): Ensamble de múltiples árboles que mejora la estabilidad del modelo, reduce el sobreajuste y captura relaciones complejas entre los atributos.
- **Máquinas de Vectores de Soporte (SVM)**: Algoritmo robusto y eficaz en espacios de alta dimensionalidad. Permite separar clases incluso en fronteras no lineales mediante funciones *kernel*.

3.5.2 Proceso de entrenamiento

Para asegurar comparabilidad y rigor metodológico, todos los modelos fueron entrenados siguiendo el mismo procedimiento:

- División del conjunto de datos en **70% para entrenamiento** y **30% para prueba**, utilizando *train_test_split*.
- Estratificación por la variable objetivo para mantener la proporción original de empleados que renunciaron y permanecieron.
- Normalización de variables cuando el algoritmo lo requiere (KNN y SVM).
- Entrenamiento de cada modelo con los datos preparados.
- Ajuste inicial de *hiperparámetros* mediante validación cruzada simple.
- Generación de predicciones sobre el conjunto de prueba.
- Obtención de métricas estándar para evaluar el rendimiento.

Este procedimiento uniforme garantiza que las diferencias observadas entre los modelos se deben a su funcionamiento interno y no a variaciones en la forma de entrenamiento.

3.5.3 Métricas de evaluación

Los modelos fueron evaluados mediante *Accuracy*, *Precisión*, *Recall*, *F1-score*, *Matriz de confusión* y *AUC-ROC*, cuyas definiciones se presentan en el capítulo 2.3.1. De acuerdo con la hipótesis de investigación, *Recall* se estableció como métrica principal, debido a que permite evaluar la capacidad del modelo para identificar empleados que efectivamente presentan deserción laboral. Las métricas restantes se utilizaron de manera complementaria para comparar el desempeño global de los algoritmos.

3.5.4 Configuración experimental y parametrización de los algoritmos

El propósito de asegurar la trazabilidad metodológica, la reproducibilidad del proceso analítico y la comparabilidad de los resultados obtenidos, todos los algoritmos fueron entrenados y evaluados utilizando la misma participación del conjunto de los datos, el mismo conjunto de variables predictoras y mismas variables objetivo. La implementación se realizó mediante la biblioteca *Scikit-learn* de Python.

División del conjunto de datos

El conjunto de datos se dividió en subconjuntos de entrenamiento y prueba mediante la función *train_test_split*, utilizando un 70 % de las observaciones para entrenamiento y un 30 % para prueba (*test_size* = 30). Esta configuración se adoptó con el propósito de disponer de una cantidad suficiente de datos para el entrenamiento de los modelos y, simultáneamente, reservar un conjunto independiente para evaluar su capacidad de generalización sobre datos no utilizados durante el entrenamiento. Con el fin de preservar la distribución de la variable objetivo (*left*) en ambos subconjuntos, se empleó el parámetro *stratify* = *y*, evitando alteraciones en la proporción de las clases. Asimismo, se estableció *random_state* = 42 para garantizar la reproducibilidad del experimento y permitir que todos los algoritmos fueran evaluados bajo las mismas condiciones.

Preprocesamiento y normalización de los datos

Previo al proceso de entrenamiento, las variables categóricas *salary* y *department* fueron transformadas a valores numéricos mediante uso de *labelEncoder*, como se muestra en el capítulo 3.4. Esto permitiendo su incorporación en los algoritmos de clasificación supervisada. Para aquellos modelos sensibles a la escala de las variables, específicamente KNN y SVM, se aplicó un proceso de normalización utilizando *StandardScaler*. Este procedimiento fue ajustado exclusivamente sobre el conjunto de entrenamiento y posteriormente aplicado al conjunto de pruebas, con el fin de evitar la fuga de información (*data leakage*) y mantenerlas validez del proceso de evaluación.

Configuración de los algoritmos supervisados

Los algoritmos seleccionados fueron entrenados utilizando configuraciones estándar, priorizando la estabilidad del modelo y evitando un ajuste exhaustivo de *hiperparametros* que pudiera introducir sesgos en la comparación de desempeño entre los distintos enfoques. A continuación, se detallan las configuraciones utilizadas para cada modelo:

- **Regresión Logística:** El modelo fue entrenado utilizando un máximo de 1000 iteraciones (*max_iter* = 1000) y una semilla fija (*random_state* = 42) para garantizar la reproducibilidad del experimento.
- **K-Nearest Neighbors (KNN):** Se configuro utilizando cinco vecinos (*n_neighbors* = 5). Debido a la sensibilidad de este algoritmo a la escala de las variables, los datos fueron normalizados previamente mediante *StandardScaler*. El resto de los parámetros se mantuvo con la configuración predeterminada de *Scikit-learn*.
- **Árbol de decisión:** El clasificador de árbol de decisión fue entrenado utilizando el criterio de impureza Gini por defecto, sin aplicar restricciones explícitas de profundidad ni poda inicial. Se utilizo *random_state* = 42 para asegurar consistencia en los resultados.
- **Random Forest:** Se configuro con 100 árboles de decisión (*n_estimators* = 100) y una semilla fija (*random_state* = 42). Los demás parámetros permanecieron con la configuración predeterminada de *Scikit-learn*.
- **Máquinas de vectores de soporte (SVM):** El modelo SVM fue implementado utilizando un *kernel radial (RBF)*, adecuado para capturar relaciones no lineales entre las variables predictoras. Se habilitó el parámetro *probability* = true con el fin de obtener probabilidades de clasificación y calcular métricas como el área bajo la curva *ROC (AUC-ROC)*. Previamente al entrenamiento, los datos fueron normalizados mediante *StandardScaler*.

Todas las configuraciones (descritas en la tabla 4), se mantuvieron constantes durante el proceso de entrenamiento y evaluación, permitiendo una comparación objetiva, coherente y metodológicamente consistente entre los distintos modelos supervisados analizados en el estudio.

Tabla 4. Configuración experimental de los algoritmos supervisados

Algoritmo	Parámetros utilizados	Preprocesamiento
Regresión Logística (<i>Logistic regression</i>)	<i>max_iter</i> = 1000, <i>random_state</i> = 42	No aplica
KNN (<i>K-nearest neighbors</i>)	<i>n_neighbors</i> = 5	Normalización (<i>StandardScaler</i>)
Árbol de Decisión (<i>Decision tree</i>)	<i>random_state</i> = 42	No aplica
Bosque aleatorio (<i>Random Forest</i>)	<i>n_estimators</i> = 100, <i>random_state</i> = 42	No aplica
SVM (<i>Support vector machine</i>)	Kernel = RBF, <i>probability</i> = true, <i>random_state</i> = 42	Normalización (<i>StandardScaler</i>)

Adicionalmente, se definió un modelo base de clasificación por mayoría como referencia mínima de desempeño. Este modelo asigna a todas las observaciones la clase mas frecuente del conjunto de entrenamiento y se utiliza para comparar el *Recall* del modelo supervisado seleccionado, de acuerdo con el criterio establecido en la hipótesis de investigación.

3.6 Herramientas tecnológicas utilizadas

La implementación del modelo se realizó en *Python*, utilizando las bibliotecas descritas en el capítulo 2.5 para la preparación de datos, modelado, evaluación y visualización de resultados. El proceso de desarrollo en *jupyter notebook*, permitiendo integrar código, resultados y graficar dentro de un entorno reproducible. Estas herramientas facilitaron la implementación concreta de las etapas definidas por la metodología CRISP-DM.

4 Evaluación

En este capítulo se presentan los resultados obtenidos durante la fase de evaluación de la metodología CRISP-DM, correspondiente al análisis del desempeño de los modelos de aprendizaje automático supervisados implementados en este estudio. Los algoritmos fueron comparados utilizando las métricas de evaluación definida en el capítulo 3. De acuerdo con la hipótesis de investigación, el análisis se centra principalmente en la métrica *Recall*, por ser el indicador que mide la capacidad del modelo para identificar correctamente a los empleados que efectivamente presentan riesgo de deserción laboral. Dado que la omisión de estos casos representa el mayor costo para la organización, esta métrica constituye el criterio principal para aceptar o rechazar la hipótesis planteada.

4.1 Matriz de confusión

La Tabla 5 presenta los resultados obtenidos por cada algoritmo sobre el conjunto de prueba. De acuerdo con la hipótesis de investigación, el análisis se centra principalmente en la métrica *Recall*, ya que fue definida como el criterio principal para evaluar la capacidad de los modelos de identificar correctamente a los empleados que efectivamente presentan riesgo de deserción laboral.

Tabla 5. Comparación de métricas entre modelos supervisados

Modelo	<i>Recall</i>	<i>Accuracy</i>	<i>Precision</i>	<i>F1-Score</i>
Bosque Aleatorio (<i>Random Forest</i>)	0.968254	0.990889	0.993295	0.980615
<i>KNN</i>	0.922502	0.956222	0.896552	0.909342
<i>SVM</i>	0.913165	0.961556	0.924386	0.918741
Árbol de Decisión (<i>Decision Tree</i>)	0.967320	0.974889	0.929982	0.948284
Regresión Logística (<i>Logistic Regression</i>)	0.257703	0.765556	0.514925	0.343497

Los resultados evidencian diferencias relevantes entre algoritmos evaluados. Regresión Logística presenta el menor rendimiento global, principalmente debido a su bajo valor de *Recall*, lo que indica una limitada capacidad para identificar correctamente a los empleados que efectivamente abandonaron la organización. Este comportamiento la hace menos adecuada para el contexto de la deserción laboral, donde la omisión de casos positivos representa un costo significativo para la organización. Por lo contrario, Bosque Aleatorio destaca como la mejor alternativa, al presentar valores superiores al 98% en *Accuracy* y *Precision*, junto con un *Recall* cercano al 97%. Este desempeño equilibrado refleja una alta capacidad predictiva y una adecuada detección de empleados en riesgo de abandono. En conjunto, estos resultados posicionan a Bosque Aleatorio como el modelo más robusto y consistente para el fenómeno de deserción laboral analizado en este estudio.

Para reforzar la selección del modelo final, se calculó la mejora porcentual de Bosque aleatorio respecto del segundo mejor algoritmo en cada métrica. Este análisis permite cuantificar el valor agregado del modelo seleccionado y demostrar su superioridad desde un enfoque comparativo:

- *Accuracy*: +1,65 % frente a *Decision Tree*
- *Precision*: +6,3 % frente a *KNN*
- *Recall*: +0,97 % frente a *Decision Tree*
- *F1-Score*: +3,41 % frente a *Decision Tree*

Estos resultados evidencian que Bosque aleatorio no solo lidera, sino que supera consistentemente al resto de los algoritmos en todas las métricas críticas, demostrando un rendimiento más estable, preciso y balanceado. La Figura 6 presenta la comparación visual de las métricas *Accuracy*, *Precision*, *Recall* y *F1-Score*. Esta representación facilita la identificación de patrones, destacando la superioridad del modelo Bosque Aleatorio frente a las demás alternativas.

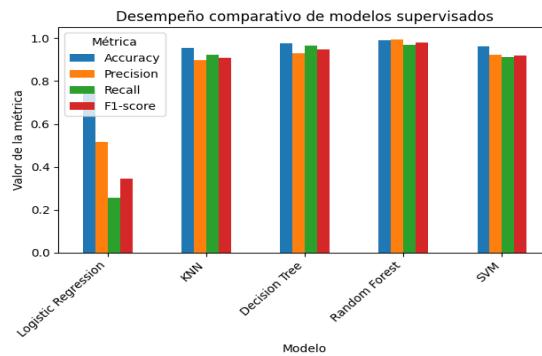


Figura 6. Comparación de modelos supervisados. Fuente: Elaboración propia

4.2 Curva ROC y AUC

La capacidad discriminativa de cada algoritmo se evaluó mediante la curva ROC (*Receiver Operating Characteristic*) y el área bajo la curva (AUC), métricas que permiten analizar el desempeño de los modelos de forma independiente del umbral de clasificación. La tabla 6 presenta los valores de AUC obtenidos para cada uno de los modelos evaluados.

Tabla 6. Métricas curva ROC – AUC

<i>Modelo</i>	<i>ROC-AUC</i>
Bosque aleatorio (<i>Random Forest</i>)	0.989975
<i>KNN</i>	0.978385
<i>SVM</i>	0.977326
Árbol de decisión (<i>Decision Tree</i>)	0.972287
Regresión logística (<i>Logistic Regression</i>)	0.801621

La figura 7 presenta las curvas ROC correspondientes a los modelos evaluados. De acuerdo con los valores de AUC presentados en la tabla 6 el modelo Bosque Aleatorio obtiene el mayor valor ($AUC \approx 0.99$), evidenciando la mejor capacidad para discriminar entre empleados que renuncian y aquellos que permanecen en la organización.

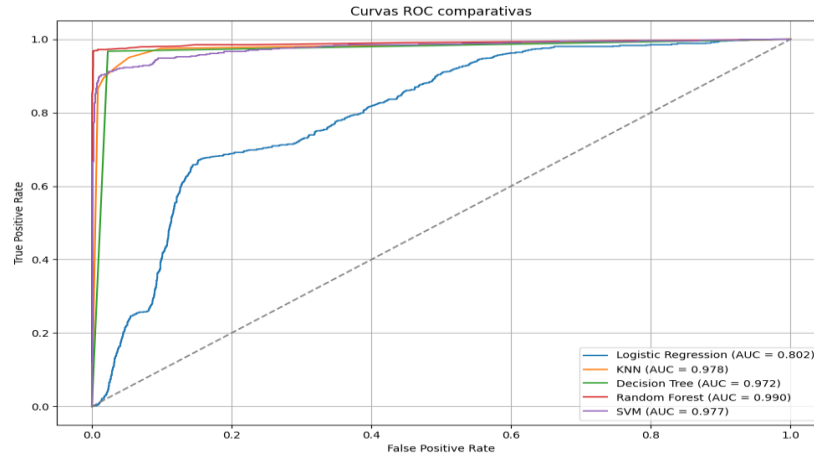


Figura 7. Curvas ROC comparación de los modelos supervisados. Fuente: Elaboración propia

Los modelos KNN, SVM y Árbol de Decisión presentan también una elevada capacidad discriminativa, mientras que Regresión Logística exhibe un desempeño considerablemente inferior, consistente con los resultados observados en métricas como *Recall* y *F1-score*.

4.3 Validación de la Hipótesis

A partir de los resultados obtenidos, los cinco modelos fueron comparados utilizando *Recall* como métrica principal. Bosque Aleatorio alcanzó el mayor valor (0,968254), por lo que fue seleccionado para su comparación con el modelo base de clasificación por mayoría, cuyo *Recall* fue 0,0000. El modelo seleccionado superó el umbral 0,90 definido en la hipótesis y presentó una diferencia de 96,83 puntos porcentuales respecto del modelo base, superior al mínimo establecido de 20 puntos porcentuales. En consecuencia, se cumplen ambos criterios, permitiendo aceptar la hipótesis de investigación para las condiciones experimentales del presente estudio.

```

=== Modelo Base: Clasificación por Mayoría ===
Accuracy: 0.7620
Precision: 0.0000
Recall: 0.0000
F1-score: 0.0000

=== Validación de la hipótesis ===
Recall modelo base: 0.0000
Recall Random Forest: 0.9683
Diferencia: 96.83 puntos porcentuales

Recall > 0.90: CUMPLE
Mejora >= 20 pp: CUMPLE

Resultado: SE CUMPLEN los criterios establecidos en la hipótesis.

```

Figura 8. Resultado de la validación de hipótesis. Fuente: Elaboración propia

5 Despliegue

En el contexto de la predicción de deserción laboral, la estabilidad y robustez del modelo adquieren especial relevancia, dado que los datos organizacionales suelen presentar diversidad, ruido y relaciones complejas entre variables asociadas al comportamiento humano. Por esta razón, más allá del rendimiento numérico obtenido, resulta necesario analizar la conducta del modelo desde una perspectiva interpretativa y aplicada al dominio del problema.

A partir de los resultados obtenidos en la evaluación comparativa, el modelo Bosque Aleatorio evidenció un desempeño consistente frente a los demás algoritmos considerandos, manteniendo valores elevados y equilibrados en métricas críticas como *Accuracy*, *Recall* y *F1-Score*. Esta consistencia sugiere una mayor estabilidad relativa del modelo al momento de generalizar patrones asociados a la deserción laboral. Desde el punto de vista metodológico, esta estabilidad se explica por la naturaleza del Bosque Aleatorio como método de ensamblaje, el cual combina múltiples árboles de decisión entrenados sobre diferentes subconjuntos de datos y variables. Este enfoque reduce la dependencia de patrones particulares, atenúa el impacto del ruido y permite capturar relaciones no lineales frecuentes en fenómenos organizacionales, como la interacción entre satisfacción laboral, carga de trabajo, antigüedad y desempeño [10]. En contraste, modelos como Regresión Logística presentan limitaciones al asumir relaciones lineales entre variables, mientras que algoritmos como KNN muestra una mayor sensibilidad al escalamiento de los datos [9]. Estas diferencias se reflejan en una mayor variabilidad del desempeño y en una menor capacidad para identificar correctamente casos de deserción laboral, especialmente en escenarios donde la detección temprana de empleados desertores es prioridad. Cabe señalar que la variable deserción (*Left*) fue definida explícitamente como variable dependiente u objetivo del estudio, representando el evento de deserción laboral, y no fue utilizada como variable independiente en ningún momento del proceso de modelado. En consecuencia, la estabilidad observada en el modelo se asocia a su capacidad para generalizar la relación entre las variables predictoras y el fenómeno de deserción laboral, y no un sesgo derivado de la definición de la variable objetivo.

Desde una perspectiva organizacional, esta estabilidad y robustez refuerzan la utilidad práctica del algoritmo Bosque Aleatorio como herramienta de apoyo a la gestión de Recursos Humanos, al ofrecer predicciones consistentes que pueden ser utilizadas para la identificación temprana de empleados en riesgo de deserción y la toma de decisiones preventivas basadas en evidencia.

5.1.1 Interpretación de resultados y utilidad para Recursos Humanos

Los resultados del análisis de importancia de variables muestran una concordancia general entre *Gini Importance* y *Permutation Importance* (figura 9), destacando *satisfaction_level* como una de las variables con mayor aporte predictivo. Asimismo, *number_project*, *average_monthly_hours*, *time_spend_company* y *last_evaluation* presentan una contribución relevante en las predicciones realizadas por el modelo de clasificación. Desde la perspectiva de Recurso Humanos, estos resultados permiten identificar las características con mayor relevancia predictiva y orientar el análisis hacia aquellos factores que podrían ser considerados en la gestión preventiva de la deserción laboral. No obstante, la importancia asignada por el modelo representa una contribución predictiva y no implica necesariamente una relación casual con la decisión de abandonar la organización.

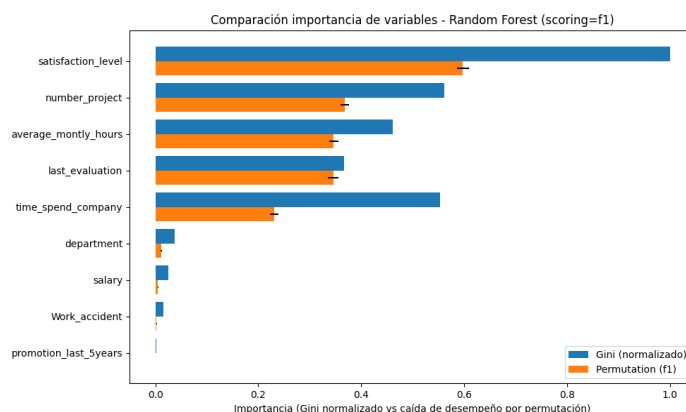


Figura 9. Gráfico de variables de importancia Bosque Aleatorio. Fuente: Elaboración propia

5.1.2 Impacto organizacional de los errores del modelo

En el contexto de la predicción de deserción laboral, no todos los errores predictivos tienen el mismo impacto desde una perspectiva organizacional. En particular, el Falso negativo constituye el error más crítico, ya que corresponde a aquellos casos en los que un empleado efectivamente renuncia, no es identificado previamente por el modelo como un riesgo. Este tipo de error implica que la organización pierde la oportunidad de implementar acciones preventivas, tales como planes de retención, ajustes en la carga laboral o intervenciones tempranas, generando una pérdida directa del capital humano. Desde esta perspectiva, la evaluación de los modelos no se limita únicamente al desempeño global, sino que prioriza la capacidad de detectar correctamente a los empleados que abandonan la organización, lo cual se refleja principalmente en métricas como *Recall* y en el análisis de las matrices de confusión. De acuerdo con los resultados obtenidos en la evaluación comparativa, el Bosque Aleatorio presentó un desempeño superior en términos de *Recall*, lo que se traduce en una menor proporción de Falsos negativos en comparación con los demás algoritmos evaluados. Este comportamiento indica que el modelo logra identificar de manera más efectiva a los empleados con alta probabilidad de deserción, reduciendo el riesgo de omitir casos críticos. Asimismo, el equilibrio observado entre *Precision* y *Recall* en el Bosque Aleatorio permite minimizar, simultáneamente, los Falsos negativos y los Falsos positivos, evitando tanto la pérdida de talento sin advertencias como la generación excesiva de alertas innecesarias. Este balance resulta fundamental para su aplicación práctica en contextos organizacionales, donde los recursos destinados a acciones de retención suelen ser limitados.

En conjunto, estos resultados refuerzan la idoneidad de Bosque Aleatorio como una herramienta de apoyo a la gestión de Recursos Humanos; su comportamiento frente a los errores en la estimación del riesgo de deserción laboral se alinea con los objetivos estratégicos de prevención del abandono y la toma de decisiones basadas en evidencia.

5.2 Conclusión de la evaluación y decisión del modelo final

La selección del modelo predictivo se fundamentó en la comparación cuantitativa y visual del desempeño de los algoritmos evaluados, así como en el análisis del impacto organizacional asociado a los errores de predicción. En este sentido, la evaluación no se limitó únicamente a métricas globales de desempeño, sino que consideró explícitamente la capacidad de los modelos para identificar correctamente a los empleados con mayor riesgo de abandono.

Los resultados presentados en la tabla 5 y en la figura 6 evidencian que Bosque Aleatorio supera de manera consistente al resto de los algoritmos en métricas clave como *Accuracy*, *Precision*, *Recall* y *F1-score*, mostrando un equilibrio robusto entre desempeño predictivo y estabilidad. Asimismo, el análisis de las curvas ROC (figura 7 y tabla 6) confirman su superior capacidad discriminativa, al aproximarse de manera más cercana al comportamiento ideal del modelo predictivo. Desde una perspectiva organizacional, este desempeño resulta especialmente relevante debido a su impacto sobre los errores críticos del modelo. Tal como se analizó en la sección 5.1.2, Bosque Aleatorio presenta una menor proporción de Falsos negativos, reduciendo el riesgo de que empleados con alta probabilidad de deserción no sean identificados oportunamente. Esta característica es fundamental en contextos de gestión del talento humano, donde la omisión de casos críticos puede traducirse en pérdida de conocimiento, costos de reemplazo y deterioro del clima organizacional. Adicionalmente, la naturaleza de ensamble del modelo Bosque Aleatorio le permite capturar relaciones no lineales y mitigar el efecto del ruido presente en los datos organizacionales, otorgándole mayor robustez frente a variaciones en la información de entrada. Estas propiedades refuerzan su efectividad como herramienta de apoyo a la toma de decisiones en Recursos Humanos. En conjunto, los resultados obtenidos permiten concluir que Bosque Aleatorio presenta el mejor equilibrio entre capacidad predictiva, estabilidad y utilidad práctica, siendo seleccionado como el modelo final para la estimación del riesgo de deserción laboral en el conjunto de empleados analizados.

5.2.1 Despliegue del modelo predictivo para estimación del riesgo de deserción

Una vez seleccionado el modelo Bosque Aleatorio como resultado del proceso evaluación, se procedió a su despliegue con el objetivo de estimar el riesgo de deserción laboral en el conjunto de empleados analizado. Esta fase corresponde a la etapa de Despliegue de la metodología CRISP-DM y se orienta a la aplicación práctica del modelo seleccionado en un contexto organizacional. El modelo fue aplicado sobre el *dataset* completo, generando para cada empleado una probabilidad estimada de deserción laboral. Estas probabilidades permiten cuantificar el nivel de riesgo individual y facilitan la interpretación de los resultados desde una perspectiva operativa. Con el fin de apoyar la toma de decisiones en Recursos Humanos, las probabilidades obtenidas fueron categorizadas en niveles de riesgo bajo, medio, alto, utilizando umbrales previamente definidos en función del comportamiento del modelo y del contexto del problema.

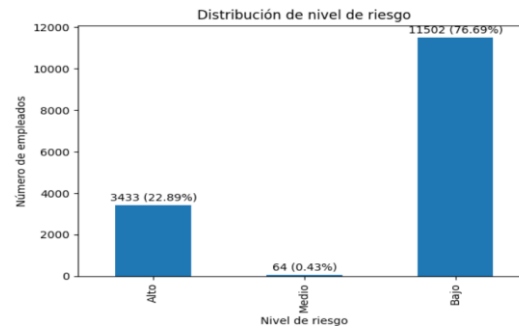


Figura 10. Distribución general de niveles de riesgo deserción laboral o rotación. Fuente: Elaboración propia

La figura 10 presenta la distribución general de los niveles de riesgos estimados en la organización. Se observa que la mayor proporción de empleados se concentran en nivel de riesgo bajo, mientras que un subconjunto relevante presenta niveles de riesgo medio y alto. Esta segmentación permite focalizar los esfuerzos de gestión del talento humano en aquellos grupos con mayor probabilidad de deserción, optimizando el uso de recursos organizacionales y priorizando acciones preventivas. Con el objetivo de profundizar el análisis, la figura 11 presenta un extracto de los primeros empleados identificados con riesgo alto de deserción, incluyendo variables relevantes asociados a su perfil laboral. En estos casos se observan patrones coherentes con los resultados del análisis previo, tales como niveles reducidos de satisfacción laboral, cargas de trabajo elevadas, ausencia de promociones recientes y salarios ubicados en rangos bajos o medios. Estos patrones concuerdan con las variables identificadas como más influyentes en el análisis de importancia del modelo (figura 9), reforzando la consistencia y validez del enfoque predictivo implementado.

```

=== Primeros empleados con riesgo ALTO (ordenado y decodificado) ===
department salary  satisfaction_level  last_evaluation  number_project  average_monthly_hours  time_spend_company  promotion_last_5years  Work_accident  prob_desercion  nivel_riesgo
sales      low      0.38             0.53             2                157                  3                    0                0              1.00           Alto
sales      medium  0.80             0.86             5                262                  6                    0                0              1.00           Alto
sales      medium  0.11             0.88             7                272                  4                    0                0              1.00           Alto
sales      low      0.72             0.87             5                223                  5                    0                0              0.99           Alto
sales      low      0.37             0.52             2                159                  3                    0                0              1.00           Alto

```

Archivo generado: lista_empleados_riesgo_alto.csv

Figura 11. Extracto de empleados en riesgo de deserción laboral.

Desde una perspectiva aplicada, este despliegue permite transformar la salida técnica del modelo en información accionable, apoyando la toma de decisiones estratégicas en Recursos Humanos. La identificación temprana de empleados con mayor riesgo de deserción facilita la implementación de intervenciones focalizadas, tales como programas de retención, ajuste en la carga laboral o acciones orientadas a mejorar la satisfacción y el compromiso organizacional. En síntesis, el despliegue de Bosque Aleatorio permite operacionalizar los resultados obtenidos en la fase de evaluación, convirtiendo el modelo predictivo en una herramienta efectiva para la estimación del riesgo de deserción laboral, y cerrando de manera coherente el ciclo analítico propuesto por la metodología CRISP-DM.

5.3 Impacto para la gestión de Recursos Humano RRHH

Los resultados obtenidos en la etapa de evaluación permiten a la organización incorporar un enfoque predictivo en la gestión del riesgo de deserción laboral, apoyando en la información generada por el modelo de aprendizaje automático supervisado. Al ser entrenado con datos históricos de empleados, el modelo aprende patrones asociados a la deserción y los utiliza para estimar la probabilidad individual de abandono en empleados activos. A partir de estimaciones, el modelo de clasificación entrega una probabilidad individual de deserción para cada empleado, la cual fue segmentada en tres niveles de riesgo: bajo, para probabilidades inferiores a 0,65; medio, para probabilidades iguales o superiores a 0,65 e inferiores a 0,80; y alto, para probabilidades iguales o superiores a 0,80. Esta segmentación permite transformar la salida probabilística del modelo en categorías interpretables para Recursos Humanos y priorizar acciones preventivas según el nivel de riesgo. Las predicciones generadas por el modelo facilitan el diseño de acciones concretas de gestión del talento, tales como:

- Implementación de programas de reconocimiento y compensación dirigidos a empleados riesgo alto estimado.
- Monitoreo continuo de indicadores asociados a satisfacción laboral, carga de trabajo y desempeño.
- Definición de estrategias de retención personalizadas basadas en los factores que el modelo identifica como influyentes en la deserción, tales como satisfacción laboral, número de proyectos asignados y horas mensuales trabajadas.

De esta manera, el modelo predictivo se consolida como una herramienta de apoyo a la toma de decisiones estratégicas en Recursos Humanos, contribuyendo a una gestión más proactiva, basada en evidencia y orientada a la prevención de la deserción laboral.

Integración del modelo predictivo en el flujo operativo de Recursos Humanos (RRHH)

La figura 12 presenta la aplicación del modelo dentro del flujo operativo de Recursos Humanos. Los datos organizacionales son procesados y utilizados por Bosque Aleatorio para estimar la probabilidad de deserción de cada empleado. Posteriormente, estas probabilidades se segmentan en niveles de riesgo bajo, medio y alto, generando información que puede apoyar acciones preventivas, reportes analíticos, alertas tempranas y estrategias de retención.

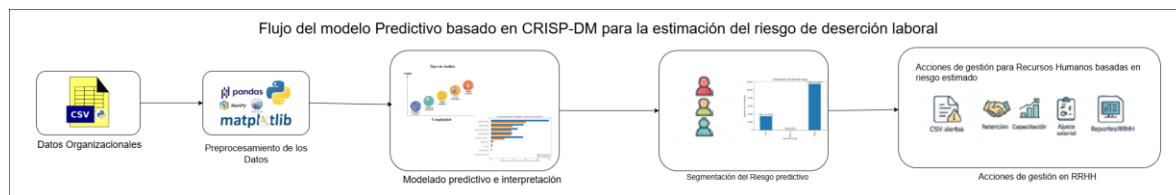


Figura 12. Aplicación del modelo predictivo en el proceso de gestión de RRHH. Fuente: Elaboración propia

6 Conclusión

El presente estudio permitió desarrollar y evaluar un modelo predictivo para estimar el riesgo de deserción laboral mediante técnicas de aprendizaje automático supervisado. De los cinco algoritmos analizados, Bosque Aleatorio obtuvo el mejor desempeño, alcanzando un *Recall* de 0,968254, lo que evidencia una alta capacidad para identificar empleados que efectivamente presentan deserción laboral. Los resultados permitieron aceptar la hipótesis de investigación, conforme a los criterios de validación establecidos y analizados en la sección 4.3, y generar información útil para identificar empleados en riesgo y apoyar la toma de decisiones en Recursos Humanos. No obstante, debido al carácter académico del conjunto de datos utilizado, la aplicación del modelo en un entorno organizacional real requiere su validación con datos propios de la organización.

6.1.1 Trabajos futuros

Como líneas de trabajo futuro, se propone validar el modelo con datos organizacionales reales para evaluar su capacidad de generalización e incorporar variables contextuales y cualitativas que permitan, ampliar el análisis de los factores asociados a la deserción laboral. Asimismo, se plantea comparar el desempeño con técnicas de *Gradient Boosting (XGBoost)*, con el propósito de evaluar si un método de ensamble basado en aprendizaje secuencial permite mejorar el desempeño predictivo obtenido en este estudio. Finalmente, se propone integrar el modelo mediante *dashboards* o sistemas de alerta que faciliten su aplicación como herramientas de apoyo a la gestión de Recursos Humanos.

Agradecimientos

Deseo expresar mi más sincero agradecimiento a mi profesor guía, José Luis Martí Lara, por su orientación, compromiso y apoyo constante durante el desarrollo de esta tesina. Asimismo, extendiendo un especial reconocimiento a **Carlos Barraza** por su dedicada gestión en el programa de MTI. Agradezco profundamente a mi familia y, en particular, a mi pareja Polette Pino Letelier, por su comprensión, motivación y apoyo incondicional en cada etapa de este proceso académico. Finalmente, agradezco a mis compañeros de estudio y colegas, cuyos valiosos comentarios y aportes enriquecieron de manera significativa el desarrollo de este trabajo.

7 Referencias

- [1] S. N. d. C. y. E. (SENCE), «Servicio Nacional de Capacitación y Empleo (SENCE) — Informe anual “Sistema de Análisis de Bolsas de Empleo (SABE)”», [En línea]. Available: https://sence.gob.cl/sites/default/files/informe_anual_2024_sabe_sence.pdf.
- [2] D. I. Gutiérrez, «BlogUdLap», [En línea]. Available: <https://blog.udlap.mx/blog/2015/01/ladesercionlaboralsusrepercusiones/>.
- [3] Q. Pro, «Questionpro», [En línea]. Available: <https://www.questionpro.com/es/analisis-de-datos.html>.
- [4] G. P.-s. P. S. Usama Fayyad, From Data Mining to Knowledge Discovery in Databases, AI Magazine, 1996.
- [5] S. E. Miner, «<https://support.sas.com/>», [En línea]. Available: <https://support.sas.com/documentation/cdl/en/emcs/66392/HTML/default/viewer.htm#n0pejm83csbja4n1xueveo2uoujy.htm>.
- [6] IBM, «Machine Learning», [En línea]. Available: <https://www.ibm.com/mx-es/think/topics/machine-learning>.
- [7] I. A. d. datos, «<https://www.ibm.com/>», [En línea]. Available: <https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview>.
- [8] E. (. I. t. M. L. (. e. M. P. Alpaydin, Aprendizaje supervisado e no supervisado, 2020.
- [9] A. (. Géron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow (2nd ed.). O'Reilly Media., 2019.
- [10] L. Breiman, Machine Learnig, Robert E.Schapiro.
- [11] 7Puentes, «7Puentes», 4 Agosto 2022. [En línea]. Available: <https://www.7puentes.com/blog/2022/08/04/ciencia-de-datos-mercado-laboral/>.
- [12] python.org, «www.python.org/», [En línea]. Available: <https://www.python.org/>.
- [13] maximaformacion.es, «<https://www.maximaformacion.es/blog-dat/que-es-r-software/>», [En línea]. Available: <https://www.maximaformacion.es/blog-dat/que-es-r-software/>.
- [14] pontia.tech, «www.pontia.tech/», [En línea]. Available: <https://www.pontia.tech/rapidminer-clave-en-data-science/>.
- [15] orangedatamining, «<https://orangedatamining.com/>», [En línea]. Available: <https://orangedatamining.com/>.
- [16] tableau, «<https://www.tableau.com/es-es/>», [En línea]. Available: <https://www.tableau.com/es-es>.
- [17] flourish.studio, «flourish.studio/», [En línea]. Available: <https://flourish.studio/>.

- [18] www.microsoft.com, «<https://www.microsoft.com>,» [En línea]. Available: <https://www.microsoft.com/es-es/power-platform/products/power-bi>.
- [19] interactivechaos, «interactivechaos,» [En línea]. Available: <https://interactivechaos.com/es/manual/tutorial-de-python/librerias-en-data-science>.
- [20] Jupyter.org, «Jupyter.org,» [En línea]. Available: <https://docs.jupyter.org/en/latest/>.
- [21] V. V. & P. G. K. (Saradhi, «ScienceDirect,» 2011. [En línea]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0957417410007621?via%3Dihub>.
- [22] E. M. S. Buriticá, «Repository.eafit.edu.co,» Facultad de Ciencias, Universidad Eafit, Medellín, Colombia, [En línea]. Available: <https://repository.eafit.edu.co/server/api/core/bitstreams/40e1422e-7f64-4ece-8944-12945b5501db/content..>
- [23] A. Š. D. & F. J. (P. o. t. e. t. i. u. d. t. Živković, «<https://www.scitepress.org>,» 2024. [En línea]. Available: <https://www.scitepress.org/Papers/2024/125384/125384.pdf>.
- [24] A. M. S. & C. M. (E. a. p. u. m. l. t. J. o. K. f. M. a. C. Rana, «<https://journal.uokufa.edu.iq>,» 2024. [En línea]. Available: <https://journal.uokufa.edu.iq/index.php/jkmc/article/view/15735/13401>.
- [25] K. (H. A. E. T. Dataset, «<https://www.kaggle.com/datasets>,» [En línea]. Available: <https://www.kaggle.com/datasets>.
- [26] Amazon, «aws.amazon,» [En línea]. Available: <https://aws.amazon.com/es/what-is/python/#:~:text=Python%20es%20un%20lenguaje%20de,ejecutar%20en%20muchas%20plataformas%20diferentes..>
- [27] G. d. Chile, «G. d. Chile, «Gobierno de Chile,» 12 03 2023. [En línea]. Available: <https://www.gob.cl/noticias/40-horas-es-ley-conozca-los-detalles-del-proyecto-impulsado-por-el-gobierno/..>
- [28] E. Digitala, «Enpresa Digitala,» [En línea]. Available: <https://enpresadigitala.spri.eus/es/introduccion-a-machine-learning-con-orange-3-onli/#:~:text=Orange%203%20es%20un%20software,o%20el%20an%C3%A1lisis%20de%20datos..>
- [29] A. d. D. Google, «Coursera,» [En línea]. Available: <https://www.coursera.org/programs/google-learning-program-len1k/professional-certificates/analisis-de-datos-de-google?collectionId=KhW1R>.
- [30] healthdataminer, «healthdataminer.com,» [En línea]. Available: <https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/>.
- [31] J. R. Molina, «Repository.uniandes.edu.co,» [En línea]. Available: <https://repository.uniandes.edu.co/server/api/core/bitstreams/5e096195-3e80-4d97-b379-9e87ab2d4494/content..>
- [32] Oracle, «Oracle Chile,» [En línea]. Available: [https://www.oracle.com/cl/database/what-is-database/#:~:text=Una%20base%20de%20datos%20es,bases%20de%20datos%20\(DBMS\)..](https://www.oracle.com/cl/database/what-is-database/#:~:text=Una%20base%20de%20datos%20es,bases%20de%20datos%20(DBMS)..)
- [33] V. J. Reyes, «LinkedIn,» 9 Noviembre 2021. [En línea]. Available: https://www.linkedin.com/pulse/qu%C3%A9-es-machine-learning-y-para-sirve-vidal-j-reyes/?trk=public_profile_article_view.
- [34] A. d. talento, «Asesor del talento,» [En línea]. Available: <https://www.asesordeltalento.cl/blog/delega-de-manera-facil-4/como-reducir-la-rotacion-de-personal-31>.
- [35] R. Zambrano, «OpenWebinars,» 15 Octubre 2019. [En línea]. Available: <https://openwebinars.net/blog/modelos-de-machine-learning/>.
- [36] K. D. Base, «kaggle.com,» [En línea]. Available: <https://www.kaggle.com/datasets/giripujar/hr-analytics>.
- [37] E. M. S. Buriticá, «Repository.eafit.edu.co,» [En línea]. Available: <https://repository.eafit.edu.co/server/api/core/bitstreams/40e1422e-7f64-4ece-8944-12945b5501db/content>.
- [38] O. %. (d. 2. Instituto Nacional de Estadísticas (INE) — “Tasa de Rotación Laboral de junio de 2024 se situó en 30, «www.inec.gob.cl/,» [En línea]. Available: <https://www.inec.gob.cl/sala-de-prensa/prensa/general/noticia/2024/12/18/tasa-de-rotaci%C3%B3n-laboral-de-junio-de-2024-se-situ%C3%B3-en-30-0>.
- [39] F. C. M. G. R. & W. D. L. E. (P. e. a. u. m. l. t. C. allucchi, «Mdpi,» 2020. [En línea]. Available: <https://www.mdpi.com/2073-431X/9/4/86>.
- [40] Jupyter, «Jupyter,» [En línea]. Available: <https://docs.jupyter.org/en/latest/>.
- [41] <https://scikit-learn.org>, «<https://scikit-learn.org>,» [En línea]. Available: <https://scikit-learn.org/stable/modules/svm.html#classification>.



UNIVERSIDAD TÉCNICA
FEDERICO SANTA MARÍA

DEPARTAMENTO
DE INFORMÁTICA



MAGÍSTER EN
TECNOLOGÍAS DE
LA INFORMACIÓN