Escuela de Administración

Escuela de Ingeniería, Ciencia y Tecnología

Maestría en Business Analytics

Sistema de predicción y alerta de deserción para empleados de una farmacéutica

Presentado por:

María Alejandra Gracia, Paola Huertas Lozano, Jaime Andrés Molina y Diana Marcela Díaz

Bogotá, D.C. 25 de noviembre de 2023

Escuela de Administración

Escuela de Ingeniería, Ciencia y Tecnología

Maestría en Business Analytics

Sistema de predicción y alerta de deserción para empleados de una farmacéutica

Presentado por:

María Alejandra Gracia, Paola Huertas Lozano, Jaime Andrés Molina y Diana Marcela Díaz

Bajo la dirección de:

Diego Güecha López

Bogotá, D.C. 25 de noviembre de 2023

Declaración de originalidad y autonomía

Declaramos bajo la gravedad del juramento, que hemos escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por nuestra propia cuenta y que, por lo tanto, su contenido es original.

Declaramos que hemos indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

María Alejandra Gracia

Paola Huertas

Jaime Andrés Molina Rodríguez

Diana Marcela Díaz

Firmado en Bogotá, D.C. el 25 de noviembre de 2023

Declaración de exoneración de responsabilidad

Declaramos que la responsabilidad intelectual del presente trabajo es exclusivamente de sus autores. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.

María Alejandra Gracia

Paola Huertas

Jaime Andrés Molina

Diana Marcela Díaz

Firmado en Bogotá, D.C. el 25 de noviembre de 2023

Tabla de contenido

Lista de Figuras

Lista de Tablas

Abreviaturas

- CRISP-DM: Cross Industry Standard Process for Data Mining
- ACP: Análisis de Componentes Principales
- ACM: Análisis de Correspondencias Múltiples
- Min.: Valor mínimo
- Max.: Valor máximo
- STD: Desviación Estándar
- SQL: Structured Query Language
- DAMA: Data Management Association
- SMOTE: Synthetic Minority Over-sampling Technique
- ADASYN: Adaptive Synthetic Sampling

Glosario

Almacenamiento de datos: Uso de medios magnéticos y de grabación para preservar y registrar información de manera digital. (IBM, s/f-a)

Arquitectura de datos: Descripción sobre la gestión de datos desde su recopilación hasta su transformación, distribución y consumo. Es el establecimiento de plan para los datos y la forma como fluyen a través de los sistemas de almacenamiento. (IBM, s/f-d)

Data Warehouse: Se refiere a un almacén de datos o un sistema de almacenamiento digital que conecta y agrupa grandes cantidades de datos de gran cantidad de fuentes diferentes. Herramienta valiosa para las organizaciones ya que les permite obtener información empresarial de gran valor para mejorar la toma de decisiones. (Oracle, s/f)

Deserción (Empresa): Reducción gradual de la fuerza laboral a medida que los empleados se van de manera voluntaria.

Esquema de estrella: Tipo de esquema de bases de datos que se compone por una tabla de hechos central asociada a distintas tablas de dimensiones.(IBM, 2021a)

Estadística multivariada: Conjunto de técnicas que permite la observación y análisis simultáneo de múltiples variables de resultado. (Díaz Monroy, 2007)

Integridad de datos: Hace referencia al almacenamiento de datos de manera precisa, completa, consistente y confiable. Garantiza que los datos sean confiables y accesibles durante su ciclo de vida.

Machine Learning: Subconjunto de la inteligencia artificial. Se enfoca en enseñar a las computadoras para que aprendan de los datos y mejoren con la experiencia, en lugar de ser explícitamente programadas para hacerlo.(IBM, s/f-c)

Minería de datos: Proceso mediante el cual se busca reconocer patrones, características, y demás información valiosa en grandes conjuntos de datos. Es de gran ayuda para las organizaciones ya que ayuda a tomar mejores decisiones. (IBM, s/f-b)

Modelo operativo centralizado: Tipo de estructura que centraliza la gestión y control de datos en una sola parte de una empresa.(Data Universe, 2023)

Modelo supervisado: Tipo de modelo en analítica que mediante variables de entrada tiene un campo de salida conocido, es decir, predicen un resultado determinado. (IBM, 2021b)

Scrum: Metodología ágil para la gestión de proyectos para desarrollarlos de manera eficaz y rápida. (Deloitte, s/f)

Variable dicotómica: En estadística corresponde a una variable indicadora que toma valores cero y uno para identificar distintas clases en una variable usualmente cualitativa. (Arias, 2021)

Resumen Ejecutivo

La alta deserción de empleados es un problema que genera grandes pérdidas y retos para las compañías hoy en día (Jain et al., 2020). Por esta razón, tener herramientas que permitan disminuir o controlar las altas tasas de abandono es una prioridad. El siguiente proyecto tiene como objetivo crear un sistema de predicción y alerta de deserción para empleados de una compañía farmacéutica que contribuya a disminuir y controlar las tasas de deserción que presenta en la actualidad.

La creación del sistema se realiza a través de la aplicación de técnicas de análisis de datos y machine learning, que acompañado de la construcción de un data warehouse permite realizar el análisis adecuado de la situación actual. En cuanto a su funcionalidad, como sistema de predicción, será utilizado para determinar la probabilidad de renuncia de cada candidato que esté postulándose a la compañía haciendo uso de una calculadora. Por otro lado, como sistema de alerta, permitirá a los tomadores de decisiones comprender qué empleados tienen mayor riesgo de renuncia, así como las áreas impactadas, esto será facilitado a los usuarios a través de un tablero organizacional. Mediante este conocimiento, los usuarios del sistema serán capaces de actuar a tiempo para lograr tomar decisiones que reduzcan el porcentaje de deserción en la compañía.

El sistema de predicción y alerta funcionará como una herramienta que permita optimizar el proceso de contratación y retención de los empleados. De esta manera, se espera que en el largo plazo los efectos negativos asociados a la deserción en la compañía sean atenuados y los procesos de contratación más efectivos.

Palabras clave: Sistema de predicción y alerta, Deserción, Aprendizaje de máquina, Farmacéutica, Almacén de datos.

Abstract

Employee attrition is a problem that generates big losses and challenges for companies nowadays (Jain et al., 2020). For this reason, it has become a priority to develop tools or systems that are able to decrease or control the high desertion rates in companies. The main purpose of this project is to create an attrition forecasting and warning system of employees from a pharmaceutical company, to help reduce the attrition rates.

The creation of the system will be developed through the application of data analysis techniques and machine learning along with a data warehouse that enables accurate analysis of the current situation. Regarding its functionality, as a forecasting system, it will be used to determine the probability of resignation of a candidate who applies for an open position in the company, by using a calculator module. As a warning system, it will allow decision makers to understand which employees are at greatest risk of resignation, as well as the impacted areas, this will be available to the users through an organizational dashboard. By using the system, users will be able to react on time and make decisions to reduce the attrition rate in the company.

The forecasting and warning system will therefore work as a tool to optimize hiring processes and control employees' retention. It is expected that in the long term, the negative effects associated with attrition in the company will be mitigated and recruitment processes will be more effective.

Keywords: Forecasting and warning system, Attrition, Machine learning, Pharmaceutical, Data warehouse**.**

1. Introducción

La industria farmacéutica juega un papel fundamental en el desarrollo científico y la investigación. Por esta razón, una compañía de este sector requiere personal especializado en el campo de la salud que cuente con la suficiente experiencia y conocimientos para llevar a cabo desarrollos innovadores y así ser competitivos frente a las demás empresas de la industria. Contar con estos perfiles le permitirá mantener los estándares de calidad de sus productos, así como generar confianza en los consumidores y alcanzar una alta reputación en el mercado. De igual manera, una baja tasa de deserción le permite a la empresa tener mejor control de sus gastos evitando sobrecostos en contratación y entrenamiento de nuevo personal.

La farmacéutica analizada en este proyecto cuenta con científicos e investigadores cuyo enfoque principal es el estudio y desarrollo de medicamentos, quienes también garantizan la seguridad y calidad de los productos. Adicionalmente, la empresa cuenta con personal ejecutivo y profesionales en mercadeo y ventas dedicados a la promoción y comercialización de los productos. Los datos de este proyecto provienen de una fuente teórica dada la confidencialidad de este tipo de información en la mayoría de las empresas.

El abandono de los empleados en esta industria genera múltiples efectos negativos. Internamente, la empresa pierde talento especializado y conocimiento difícil de encontrar en el mercado. Por un lado, esto pone en riesgo la continuidad de estudios científicos e investigaciones, y, por otro lado, la deserción en cargos de ejecutivos de ventas podría afectar el nivel de ventas de la compañía y por tanto sus indicadores financieros. Adicionalmente, la pérdida de talento genera un aumento en costos de contratación, entrenamiento y pone en riesgo la calidad del servicio brindado. Es por esto que, contratar nuevo personal resulta un desafío para el área de reclutamiento teniendo en cuenta las habilidades necesarias para este sector.

Debido a lo anterior, el proyecto busca crear un sistema de predicción y alerta de deserción que permita identificar con antelación aquellos candidatos con alta probabilidad de renuncia y que además anticipe el posible retiro de un empleado activo de manera oportuna (Jain et al., 2020). Para lograrlo, se hará uso de técnicas de minería de datos mediante machine learning, modelado y visualización de la información. Utilizando las fuentes disponibles y a través de la aplicación de técnicas de analítica de datos, se busca que el área de recursos humanos y los directivos tomen decisiones fundamentadas en la información.

Con los resultados de este proyecto, recursos humanos contará con una calculadora que le permita identificar la probabilidad de deserción de los aspirantes a la compañía. Adicionalmente, los directivos de la organización y la operación tendrán acceso a un tablero de visualización donde se expongan las variables más relevantes asociadas a la probabilidad de deserción de los empleados.

Al inicio se presentan los objetivos generales y específicos del sistema de predicción y alerta de deserción los cuales comprenden el proceso de construcción del sistema, su alcance y su aplicación esperada. En el alcance se presentan los entregables que lo componen, los procesos necesarios para su creación, los recursos actuales con los que se cuenta, sus limitaciones y por último el resultado esperado.

La siguiente sección presenta el cronograma del proyecto, el cual fue planteado haciendo uso de técnicas de gestión ágil (Scrum) para alcanzar un desarrollo óptimo y eficiente que agregue valor al negocio. El cronograma se encuentra dividido en tres fases: anteproyecto; entendimiento y preparación de datos y, finalmente desarrollo e implementación del sistema.

Además, este documento presenta un contexto de la problemática de deserción tanto a nivel global como a nivel sectorial enfocado a las farmacéuticas, con el fin de establecer el impacto de este reto empresarial y sus posibles causas.

El documento también contiene una descripción sobre la calidad de datos de las fuentes de información disponibles, tratamiento sobre inconsistencias y además incluye análisis estadístico de las mismas donde se presentan medidas de tendencia central tales como media, moda, desviación estándar, entre otras. Adicionalmente, se exponen hallazgos relevantes sobre las bases de datos, los cuales incluyen tendencias, correlaciones, patrones y creación de nuevas variables en caso de ser necesarias para el desarrollo del sistema.

De igual manera se incluye el proceso realizado para la construcción del modelo operativo que servirá como herramienta de almacenamiento de la información. Este detalle incluye una recomendación de arquitectura de tablas, de manera que la empresa farmacéutica logre centralizar sus datos y mejorar procesos de consulta, procesamiento, integridad y almacenamiento de los datos provenientes de las fuentes disponibles. Posteriormente, en la sección de análisis desde el punto de vista de los datos, se aplican métodos de estadística multivariada y el resultado sobre el análisis del primer tablero de indicadores para extraer nuevas hipótesis sobre el problema. Estas herramientas permiten además encontrar patrones sobre los datos y establecer dependencias entre variables que aporten valor al entendimiento del problema de deserción.

Seguido a esto, se presenta el detalle sobre el proceso de modelado para la creación del sistema predictivo, mediante la aplicación de modelos de machine learning supervisados para determinar riesgo de deserción en empleados y candidatos. Esta sección incluye una descripción del proceso de entrenamiento y selección de los modelos que mejor se ajustan a las necesidades,

así como el método de evaluación y finalmente una breve descripción sobre la creación de las herramientas y su conexión con el modelo supervisado.

En la sección final se presentan sugerencias de implementación del sistema para lograr el objetivo organizacional de la farmacéutica y las conclusiones del desarrollo de este proyecto.

2. Objetivos

Crear un sistema de predicción y alerta de deserción de empleados de una farmacéutica que permita estimar la probabilidad de renuncia de los candidatos a la compañía y, además, detectar dentro del personal activo aquel con mayor riesgo de abandonar la empresa. Ambos componentes ayudarán a alcanzar un mejor desempeño en la compañía mediante la toma decisiones informada y la reducción de niveles de deserción.

2.1.Objetivos específicos

1. Realizar un análisis descriptivo que permita caracterizar las variables más relevantes que inciden en el problema de deserción de la compañía.
2. Identificar relaciones entre las variables disponibles mediante el uso de técnicas de minería de datos.
3. Definir la arquitectura de datos necesaria para implementar el proyecto y crear el sistema.
4. Implementar un tablero para los ejecutivos y el área de Recursos Humanos con información general de deserción en la compañía para obtener una mejor comprensión de la situación.
5. Crear un modelo de predicción que permita identificar los empleados con mayor probabilidad de deserción para tomar acciones estratégicas.
6. Construir un tablero para ejecutivos que les permita identificar aquel personal con mayor riesgo de deserción para tomar decisiones oportunas y mitigar el riesgo de salida por parte de los empleados.

7. Desarrollar y diseñar una calculadora de deserción para el área de Recursos humanos que permita medir la probabilidad de renuncia de acuerdo con variables relacionadas al perfil de potenciales empleados.

3. Alcance

El propósito de realizar este proyecto es crear un sistema de predicción y alerta de deserción que a largo plazo permita disminuir las tasas de rotación de los empleados de una farmacéutica a través de una toma de decisiones informada y a tiempo. El proyecto está dirigido específicamente al personal de una empresa de la industria farmacéutica dedicada a la comercialización de productos medicinales, así como la investigación y desarrollo de nuevos tratamientos médicos.

El sistema permitirá a los usuarios ingresar en una calculadora ciertas características asociadas al perfil profesional de los candidatos a empleados en su proceso de selección. Ésta arrojará un porcentaje de probabilidad de renuncia para cada aspirante al cargo. El sistema, además de incorporar una herramienta de simulación, incluirá un tablero de alertas de posible deserción de los trabajadores activos en la compañía para así detectar aquellos con mayor riesgo de abandono.

El producto final del sistema será un porcentaje de probabilidad de renuncia tanto para candidatos a cargos dentro de la empresa, como para quienes se encuentran activos en la compañía. Los usuarios de éste serán los funcionarios de las áreas encargadas de reclutamiento y los directivos que tengan participación en los procesos de contratación y/o intervengan de alguna manera en el funcionamiento y operación de la farmacéutica.

Como resultado de la ejecución del proyecto se realizarán las siguientes entregas:

1. Diseño de Data Warehouse.
2. Análisis descriptivo de las variables disponibles en las diferentes fuentes de información y su presentación en un tablero dinámico

3. Modelo de machine learning que permita identificar los empleados activos con mayor probabilidad de deserción.
4. Tablero de alertas y de situación de deserción actual.
5. Modelo de machine learning que permita detectar candidatos con mayor riesgo de abandono.
6. Calculadora de probabilidad de renuncia.
7. Sugerencias sobre implementación del sistema.

Los recursos necesarios para llevar a cabo esta iniciativa son:

1. Bases de datos.
2. Documentación relacionada a métodos estadísticos.
3. Información de la industria farmacéutica.
4. Herramientas de programación y visualización.
5. Acompañamiento de experto en estadística y programación.

Los riesgos del proyecto están asociados a bases de datos inconsistentes e incompletas que dificulten la construcción de un sistema con resultados precisos.

El éxito del alcance de este proyecto se medirá a través de las pruebas de funcionalidad que se realicen en la tercera fase del proyecto. Adicionalmente, se monitorearán los tiempos de ejecución acordes al cronograma y el completo desarrollo de cada uno de los procesos necesarios para llevar a cabo la iniciativa.

4. Metodología

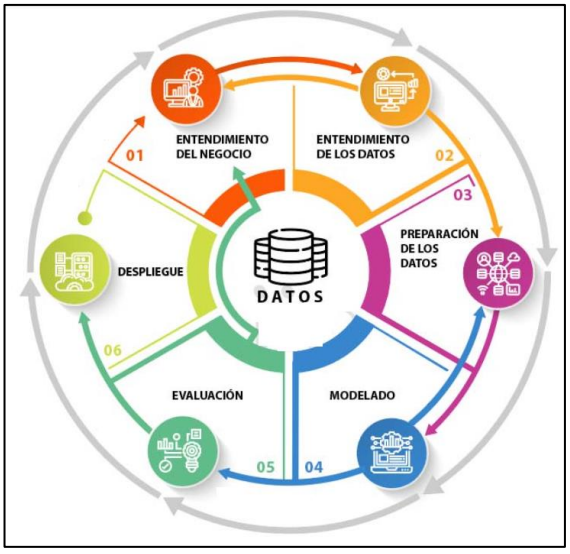
Para el desarrollo del proyecto, es necesario extraer nuevos hallazgos y conocimientos mediante herramientas de análisis de información y la aplicación de minería de datos a fuentes

relacionadas con la deserción de los empleados en la farmacéutica. Lo anterior para generar nuevos insumos que le permitan a la compañía tomar decisiones basadas en datos que le brinden beneficios y mitigue posibles riesgos. De acuerdo con lo anterior, es necesario utilizar una metodología que permita realizar un proyecto de investigación sobre un problema frecuente en las empresas, como lo es el retiro de los empleados y, a partir de esto generar tareas de minería de datos para encontrar la mejor forma de mitigarlo y así reducir los efectos negativos de éste en la empresa.

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining) fue creada y optimizada para ser neutral a cualquier industria, herramienta y/o aplicación en la que se requiera implementar un proyecto de minería de datos. Las fases del modelo de referencia de CRISP-DM se adaptan a la necesidad del proyecto, por lo que se usará para su desarrollo. Esta metodología consta de seis fases principales, que, si bien tienen un orden esquemático, no es una secuencia rígida ya que dentro de su teoría es un método que se mejora a través de la iteración entre sus fases (Chapman et al., 2000).

Figura 1

Fases CRISP-DM



Fuente: Instituto de ingeniería del conocimiento, 2021

Cada una de las fases de la metodología CRISP-DM serán ejecutadas de la siguiente manera según las recomendaciones de Chapman en CRISP-DM (Chapman et al., 2000):

4.1. Entendimiento del negocio

En esta fase se realizará una exploración del contexto de negocio de la industria farmacéutica, sin embargo, al ser la deserción de empleados un problema transversal en las empresas, se hará énfasis en indagar acerca de los factores principales que intervienen en esta problemática, el impacto que genera y las alternativas de mitigación.

4.2. Entendimiento de los datos

En esta etapa de la metodología se realiza el reconocimiento de las fuentes de información disponibles de la farmacéutica. Además, se ejecutará un análisis descriptivo de los datos y se implementarán métodos estadísticos para identificar problemas de calidad y extraer conclusiones iniciales. En esta fase se espera encontrar hallazgos para entender y atacar el problema de deserción, además de definir un plan de transformación de la información.

4.3. Preparación de los datos

Esta fase está enfocada en el proceso de tratamiento de las fuentes, iniciando por gestionar y/o reparar los problemas de calidad encontrados para tener bases de datos limpias. Posteriormente se establecerá un diseño de Data Warehouse, que funcione como centro de almacenamiento de las fuentes, y así mantener la integridad de la información de la farmacéutica, generando un estándar de formato de datos confiable, procesamiento ágil y uso posterior a la finalización del proyecto.

También se buscará crear nuevos campos o indicadores clave a partir de los datos iniciales que permitan complementar la información de la farmacéutica. La correcta ejecución de la preparación de los datos permitirá crear los tableros necesarios, para que los ejecutivos de la farmacéutica puedan realizar análisis y tomar decisiones de manera ágil e interactiva. Además de

esto, será posible generar un conjunto de datos que facilite la fase de modelado y la creación de las herramientas que componen el sistema.

4.4. Modelado

La fase de modelado utilizará los conjuntos de datos generados en la fase anterior para entrenar diferentes modelos de machine learning supervisados cuyo objetivo es determinar la probabilidad de deserción de los empleados de la farmacéutica, así como el riesgo de renuncia de candidatos. Como lo sugiere la metodología CRISP-DM, es posible que sea necesario volver a la etapa anterior para modificar la estructura de entrenamiento y tener modelos más confiables. Los modelos generados en esta etapa serán el insumo principal de la fase de evaluación.

4.5. Evaluación

Los modelos entrenados previamente se deben evaluar de forma detallada, no solo a nivel analítico, sino también de negocio e identificar si están alineados con los objetivos del proyecto.

Luego de elegir un modelo que cumpla con las expectativas según el criterio experto y estadístico, el producto final será la creación del segundo tablero. La integración de los resultados del modelo en este tablero permitirá a los ejecutivos detectar a tiempo a los empleados que estén en alto riesgo de abandonar la farmacéutica, para así tomar acciones antes de la materialización del riesgo. De manera paralela, se identificarán los factores que pueden ser evaluados a la hora de realizar una contratación y será posible también crear la calculadora de riesgo de deserción. Ésta apoyará las labores del área de recursos humanos en actividades de reclutamiento para disminuir el riesgo de contratar candidatos con alta probabilidad de renuncia.

4.6. Implementación

La fase de implementación estará enfocada en sugerir el despliegue del sistema de deserción en la compañía, el cual abarca desde la capacitación de la herramienta, además del uso

adecuado del Data Warehouse para centralización de los datos, hasta el tablero de análisis de información actual, el tablero de detección de posibles empleados desertores y la calculadora de probabilidad de deserción de candidatos.

5. Cronograma

El proyecto cuenta con un plan de ejecución dividido en tres fases. A continuación, se presentan cada una de las fases junto con sus resultados esperados para dar cumplimiento a los objetivos. Se hará uso de metodologías ágiles para alcanzar un desarrollo óptimo y eficiente en busca de generar valor al negocio.

La primera fase corresponde al anteproyecto, en ésta se realiza el entendimiento del negocio para comprender la necesidad específica del proyecto. Además, se determina la manera de abordarlo, su alcance y se realiza el inventario de las fuentes disponibles. Con base en esto, se crean prototipos, diagramas de los entregables y los procesos necesarios.

Figura 2

Fase 1- Anteproyecto

Fase 1 Anteproyecto - Etapas	Noviembre 2022	Diciembre 2022	Enero 2023
1. Entendimiento del negocio			
2. Definición y alcance del proyecto			
3. Recolección de fuentes disponibles			
4. Creación de prototipos (Tablero y Calculadora)			
5. Sesión de revisión con director			
6. Sesión de retrospectiva con el equipo			

Fuente: elaboración propia

La segunda fase corresponde al entendimiento y preparación de los datos. Esta etapa incluye la limpieza de los datos, medición de calidad de los datos, creación de nuevos indicadores para realizar posteriormente la bodega de datos y así, finalmente el entendimiento de las

principales fuentes disponibles para la creación del sistema mediante análisis descriptivo y multivariado.

Figura 3

Fase 2 – Entendimiento y preparación de datos

Fase 2 Entendimiento y preparación de datos - Etapas	Febrero 2023	Marzo 2022	Abril 2023	Mayo 2023
1. Entendimiento de datos en fuentes disponibles				
2. Exploración inicial de datos				
2.1. Cáculo de calidad de datos				
2.2. Limpieza de datos				
2.3. Creación de nuevas variables				
2.4. Cálculo de estadísticos principales				
2.5. Hallazgos principales				
3. Creación de Data warehouse				
4. Sesión de revisión con el director				
5. Carga de datos en Data warehouse				
6. Versión inicial tablero - analítica descriptiva				
7. Sesión revisión con director y principales interesados				
8. Sesión de retrospectiva con el equipo				

Fuente: elaboración propia

La última fase hace referencia a la creación del sistema mediante la aplicación de modelos de machine learning. Se desarrollan las etapas de modelado, evaluación y sugerencias de implementación según la metodología CRISP-DM. Una vez se seleccione el modelo se crean las herramientas propuestas; el tablero de alerta de deserción y calculadora de probabilidad de deserción para candidatos y para finalizar se realizan sugerencias de implementación y conclusiones del proyecto.

Figura 4

Fase 3 – Desarrollo e implementación del sistema

Fase 3 Desarrollo e implementación del sistema - Etapas	Agosto 2023	Septiembre 2022	Octubre 2023	Noviembre 2023
1. Investigación sobre modelos supervisados - árboles de decisión				
2. Aplicación de modelos supervisados - árboles decisión				
3. Sesión de revisión con director				
4. Sesión de retrospectiva con el equipo				
5. Selección del mejor modelo para deserción activos				
6. Selección del mejor modelo para deserción de candidatos				
7. Integración modelo deserción activos al tablero				
8. Creación y desarrollo de calculadora				
9. Integración del modelo				
10. Sugerencias de implementación y conclusiones				

Fuente: elaboración propia

6. Entendimiento del negocio: Situación Organizacional

La rotación de personal es una problemática transversal a todas las compañías y afecta a todos los sectores. Varios autores entre ellos Gallup y Deloitte afirman el costo de reemplazar un empleado puede llegar a ser hasta el 150% de su salario anual o que puede llegar a ser 10.000 dólares al doble de su salario anual. De igual manera, según el Centro para el Progreso Estadounidense, el costo de una compañía que tenga un gran movimiento de colaboradores es del 213% del costo anual del salario medio de esa Organización. (Pourshasb, 2021).

Son múltiples los impactos que genera para una compañía la deserción de sus empleados:

- Las funciones desempeñadas por un colaborador que abandona la compañía deben ser suspendidas o reasignadas a las demás personas del equipo. Por un lado, detenerlas puede generar incumplimiento de objetivos, suspensión de proyectos, pérdida de inversión o de oportunidades de negocio, entre otros. Por otro lado, asignarlas a otras personas en algunos casos puede ocasionar una sobrecarga laboral que a su vez impacta motivación, equilibrio entre la vida laboral y personal, así como disminución en su productividad.

- La pérdida de talento para una empresa puede significar una reducción de eficiencia e innovación.
- La deserción de un empleado en la compañía lleva consigo la contratación de nuevas personas en el equipo, este proceso genera una inversión de recursos en áreas de talento humano, reclutamiento y en los líderes encargados que puede tardar varios meses.
- El ingreso de nuevas personas a la compañía o el cambio de cargos dentro de la misma trae consigo realización de entrenamientos que implican una inversión en tiempo y a su vez incrementan el riesgo de errores en los procesos por carencia de experiencia.
- El conocimiento de negocio es un aprendizaje continuo que se adquiere en gran medida por el tiempo de experiencia laboral de un colaborador en una compañía o sector. Por esta razón, el entrenamiento relacionado a la comprensión y familiarización de una industria, su dinámica operacional y estrategias, resulta costoso en términos de tiempo y esfuerzo.

Las causas de una alta deserción están asociadas a diversos factores tanto internos como externos a la compañía. Los más comunes son las condiciones laborales (salario, horario, tipo de contrato, falta de reconocimiento) exceso de carga laboral, mejor oportunidad laboral, condiciones de salud, familiares o personales, entre otros.

Teniendo en cuenta la trascendencia de la problemática descrita anteriormente, el contexto bajo el cual se analizarán los datos de la compañía farmacéutica será desde el punto de vista de rotación del personal. Los roles considerados para este análisis se encuentran dentro de las áreas de recursos humanos, ventas y desarrollo e investigación, ya que en general representan más del 50% de la estructura organizacional de una compañía del sector farmacéutico (Fitzgerald & Wilson, 2023):

Tabla 1

Distribución áreas en farmacéuticas

Departamento	Porcentaje empleados	Características
Manufactura	50%	Actividades que ocurren al principio y al final de la cadena de suministro y producción, manufactura, ciencia y tecnología y empaque
Calidad	30%	Control de calidad y cumplimiento de reglas
Operaciones	10%	Servicio técnico, TI y automatización
Investigación y desarrollo	0% - 5%	Según los productos
Otros	10%	Cadena de suministro, mercadeo, recursos humanos y finanzas

Fuente: (Fitzgerald & Wilson, 2023)

La compañía analizada para esta investigación contiene un gran porcentaje de biólogos y médicos que se desempeñan en su mayoría dentro de las áreas de investigación y desarrollo. Considerando que parte del alcance de estas compañías está asociado con ensayos clínicos los cuales “toman entre 10 y 15 años para ser completados hasta las etapas de licenciamiento”

(Bioclever, 2021), el conocimiento especializado de los empleados y su experiencia se convierten en el activo más valioso para la empresa.

Adicionalmente, otro de los departamentos que contiene un gran número de empleados es el área de ventas. En el sector farmacéutico, la actividad comercial cumple un papel fundamental a la hora de culminar con la cadena de valor, ya que es en esta etapa en la que se aseguran los ingresos y se incrementa el patrimonio de la empresa.

El vendedor, es quien tiene contacto directo y logra mantener las relaciones comerciales con los clientes, por esta razón es muy importante contar con personal capacitado que conozca a profundidad los productos y sus características, beneficios y contraindicaciones. Este rol actúa como promotor y difusor de información relacionada a nuevos productos, investigaciones clínicas y posibles tratamientos a enfermedades. Sus principales clientes hacen parte del personal de la salud por lo cual su conocimiento y experiencia en este campo es muy relevante para la reputación y reconocimiento de la compañía.

Teniendo en cuenta los impactos de una alta rotación mencionados previamente y el contexto para una compañía farmacéutica, una alta deserción de empleados generaría grandes impactos en continuidad de proyectos de investigación, ensayos clínicos y ventas, así como reputación empresarial que además de los costos financieros, podría generar una potencial pérdida de ingresos.

7. Entendimiento y exploración de las fuentes de información disponibles

Para abordar los objetivos mencionados en este proyecto, se cuenta con cuatro bases de datos como fuentes de información. Estas bases se caracterizan principalmente por almacenar información demográfica, indicadores históricos individuales sobre desempeño de los empleados

y su percepción sobre su trabajo y la compañía. Sin embargo, se evaluará la posibilidad de incluir fuentes adicionales de acuerdo con el avance del proyecto.

7.1. Base de Datos General Empleados

Almacena la información general de los empleados de tres áreas de la farmacéutica. Incluye identificadores, datos demográficos, jerárquicos y cambios de cargo en la empresa.

7.1.1. Características Generales y Descripción de Campos

La base contiene 4.410 registros, donde cada uno de ellos corresponde a un empleado y además contiene 24 variables adicionales que serán detalladas en esta sección.

Esta base contiene el registro de trabajadores activos en la compañía, pero también preserva datos de empleados inactivos. Parte de los datos registrados en la base están asociados a características demográficas como edad, género, nivel de educación, entre otros. Así mismo, incluye datos laborales de los empleados como años desde último ascenso, años de experiencia, salario mensual, entre otros.

Los siguientes son los campos que conforman la base:

Tabla 2

Campos - General Empleados

Campo	Descripción	Tipo de dato	Ejemplo
Edad	Edad del empleado	Numérico	25
Desercion	Indica si el empleado abandonó la empresa	Texto	Si
Frecuencia_viaje	Frecuencia de viaje por trabajo	Texto	Esporádico

Area	Área en la empresa	Numérico	1
Distancia_Casa	Distancia en kilómetros hogar - oficina	Numérico	15
Nivel_Educativo	Nivel de educación actual de 1 a 5	Numérico	3
Campo_Educativo	Área de educación	Texto	Medicina
Recuento_Empleado	Indicador empleado valido	Numérico	1
ID empleado	Identificador único por empleado	Numérico	20
Genero	Género	Numérico	2
Nivel_Rol	Nivel del cargo en la compañía de 1 a 5	Numérico	5
Rol	Cargo del empleado	Texto	Ejecutivo de ventas
Estado_Civil	Estado civil	Texto	Soltero
Salario_Mensual	Ingresos mensuales del trabajador en rupias	Numérico	110,31
Companias	Número de empresas en las que el empleado ha trabajado	Numérico	3

Mayor18	Indica si el trabajador tiene más de 18 años	Texto	Y
Porcentaje_Aumento_Salarial	Ultimo porcentaje de aumento salarial	Numérico	13
Horario	Horas de trabajo habituales	Numérico	8
Nivel_Opcion_Acciones	Nivel de derecho de compra de acciones de la compañía	Numérico	3
Anos_Trabajados	Número de años trabajados	Numérico	10
Entrenamientos	Número de capacitaciones recibidas en el último año	Numérico	5
Antiguedad	Número de años que el empleado ha estado en la empresa	Numérico	6
Anos_Ultima_Promocion	Años desde el último ascenso	Numérico	0
Anos_Lider_Actual	Número de años con líder actual	Numérico	4

Fuente: Elaboración propia

7.1.2. Descripción Estadística

La siguiente tabla indica las principales características de las variables cuantitativas de la base General Empleados.

Tabla 3

Características campos numéricos – General Empleados

Campo	Edad	Distancia_Casa	Nivel_Educativo	Recuento_Empleado
Cantidad	4.410	4.410	4.410	4.410
%no nulos	100,0%	100,0%	100,0%	100,0%
media	36,92	9,19	2,91	1
std	9,13	8,10	1,02	0
min	18	1	1	1
25%	30	2	2	1
50%	36	7	3	1
75%	43	14	4	1
max	60	29	5	1

Campo	Nivel_Rol	Salario_Mensual	Companias	Porcentaje_Aumento_Salarial	Horario
Cantidad	4.410	4.410	4.391	4.410	4.410
%no nulos	100,0%	100,0%	99,6%	100,0%	100,0%
media	2,84	65.029,31	2,69	15,20	8
std	1,25	47.068,88	2,49	3,65	0
min	1	10.090	0	11	8
25%	2	29.110	1	12	8

50%	3	49.190	2	14	8
75%	3	83.800	4	18	8
max	5	199.990	9	25	8

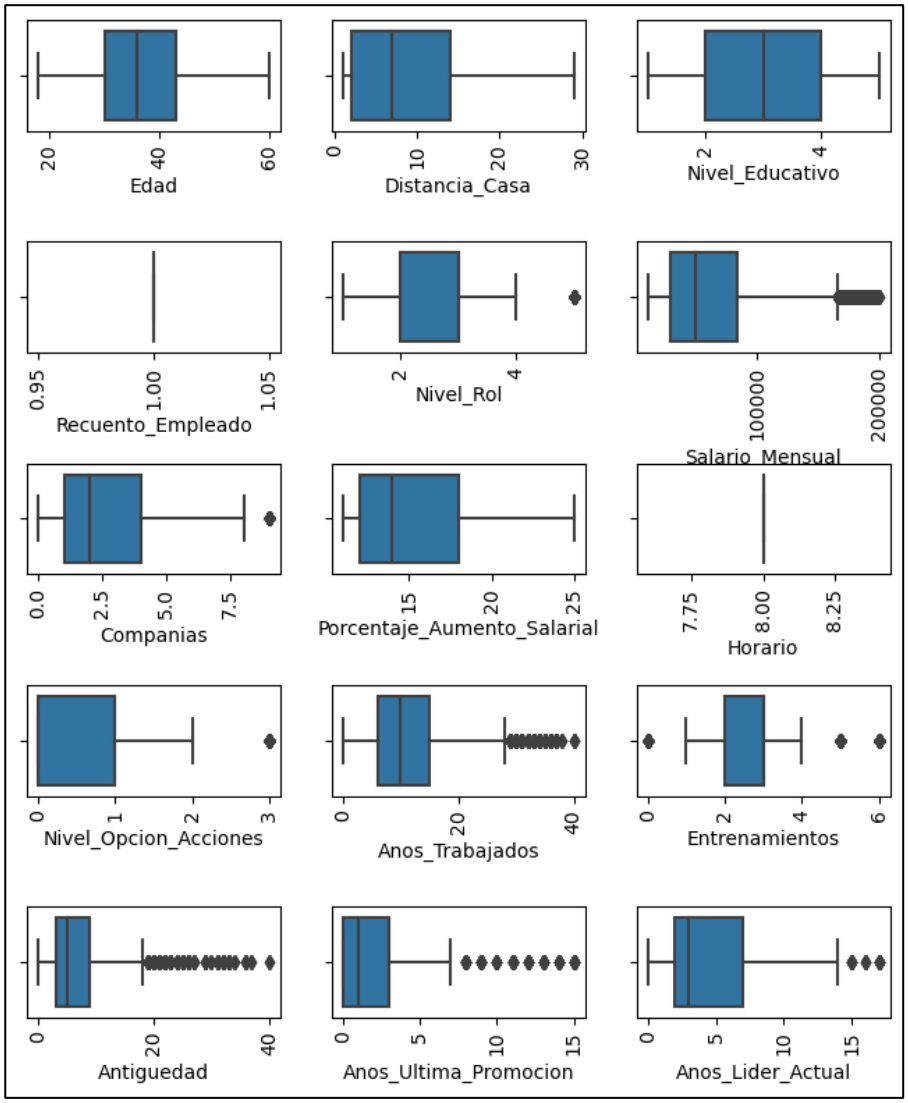
Campo	**Nivel_Opcion Acciones**	**Anos_ Trabajados**	**Entrenamientos**	**Antiguedad**	**Anos_Ultima Promocion**	**Anos_Lider_ Actual**
Cantidad	4.410	4.401	4.410	4.410	4.410	4.410
%no nulos	100,0%	99,8%	100,0%	100,0%	100,0%	100,0%
media	0,79	11,27	2,79	7,00	2,18	4,12
std	0,85	7,782	1,28	6,12	3,22	3,56
min	0	0	0	0	0	0
25%	0	6	2	3	0	2
50%	1	10	3	5	1	3
75%	1	15	3	9	3	7
max	3	40	6	40	15	17

Fuente: Elaboración propia

Para complementar el entendimiento de la base se crean diagramas de caja para encontrar la simetría de las variables, la distribución de datos y la existencia de valores atípicos en estos campos. De igual manera, este permite observar el nivel de dispersión de la variable. Los hallazgos serán resumidos a continuación.

Figura 5

Diagrama de caja – General Empleados

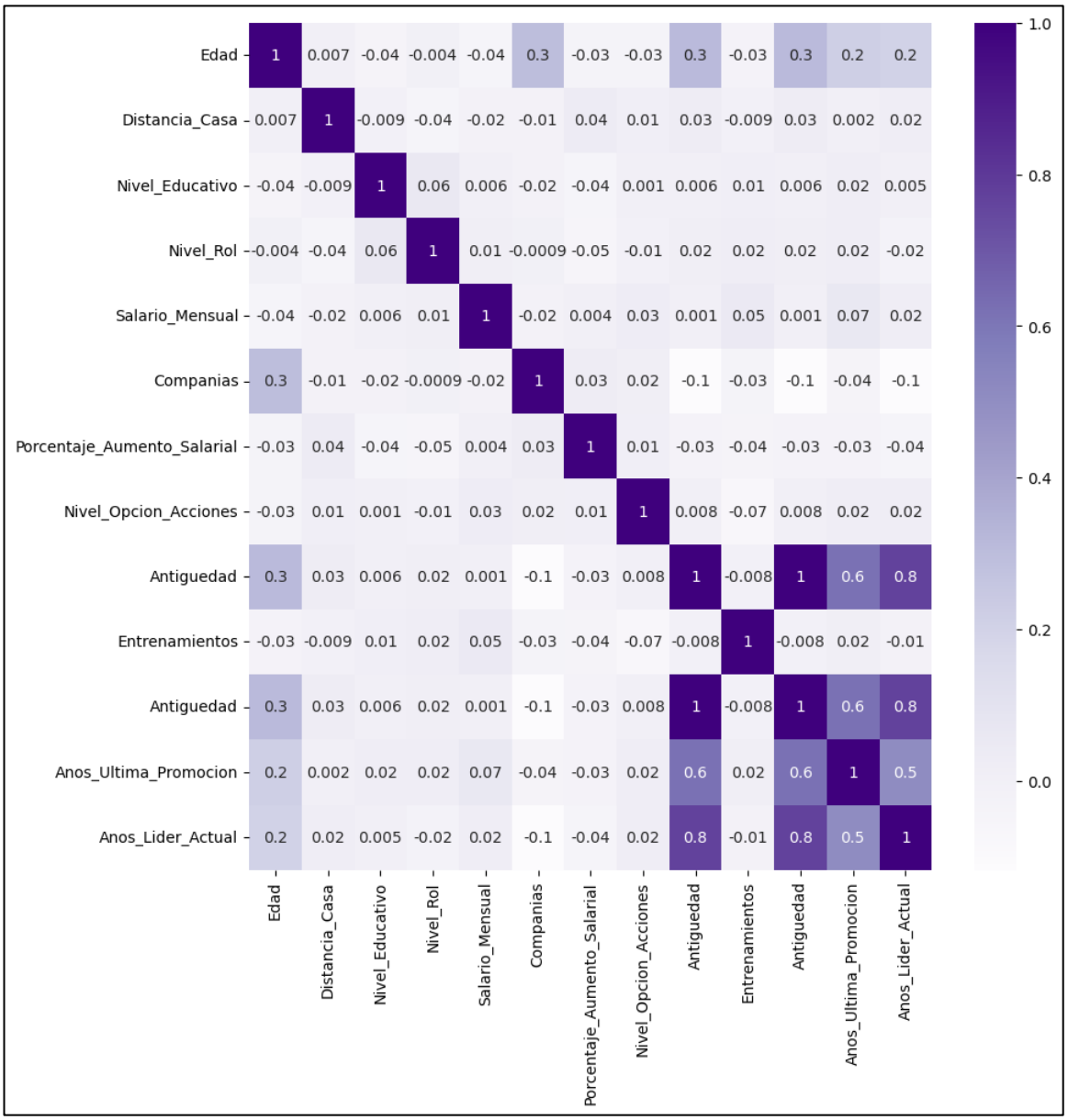


Fuente: elaboración propia

La siguiente matriz de correlaciones expone las relaciones existentes entre las distintas variables de la base General Empleados.

Figura 6

Matriz de correlaciones - General Empleados



Fuente: elaboración propia

La siguiente tabla muestra las características de las variables categóricas de la base General Empleados, indicando el porcentaje de campos no nulos, valores únicos y moda.

Tabla 4

Características campos categóricos – General Empleados

Campo	**Desercion**	**Frecuencia_viaje**	**Area**	**Campo_Educativo**
Cantidad	4.410	4.410	4.410	4.410
Únicos	2	3	3	6
Moda	No	Esporádico	Investigación & Desarrollo	Biología
Frecuencia	3.699	3.129	2.535	1.818

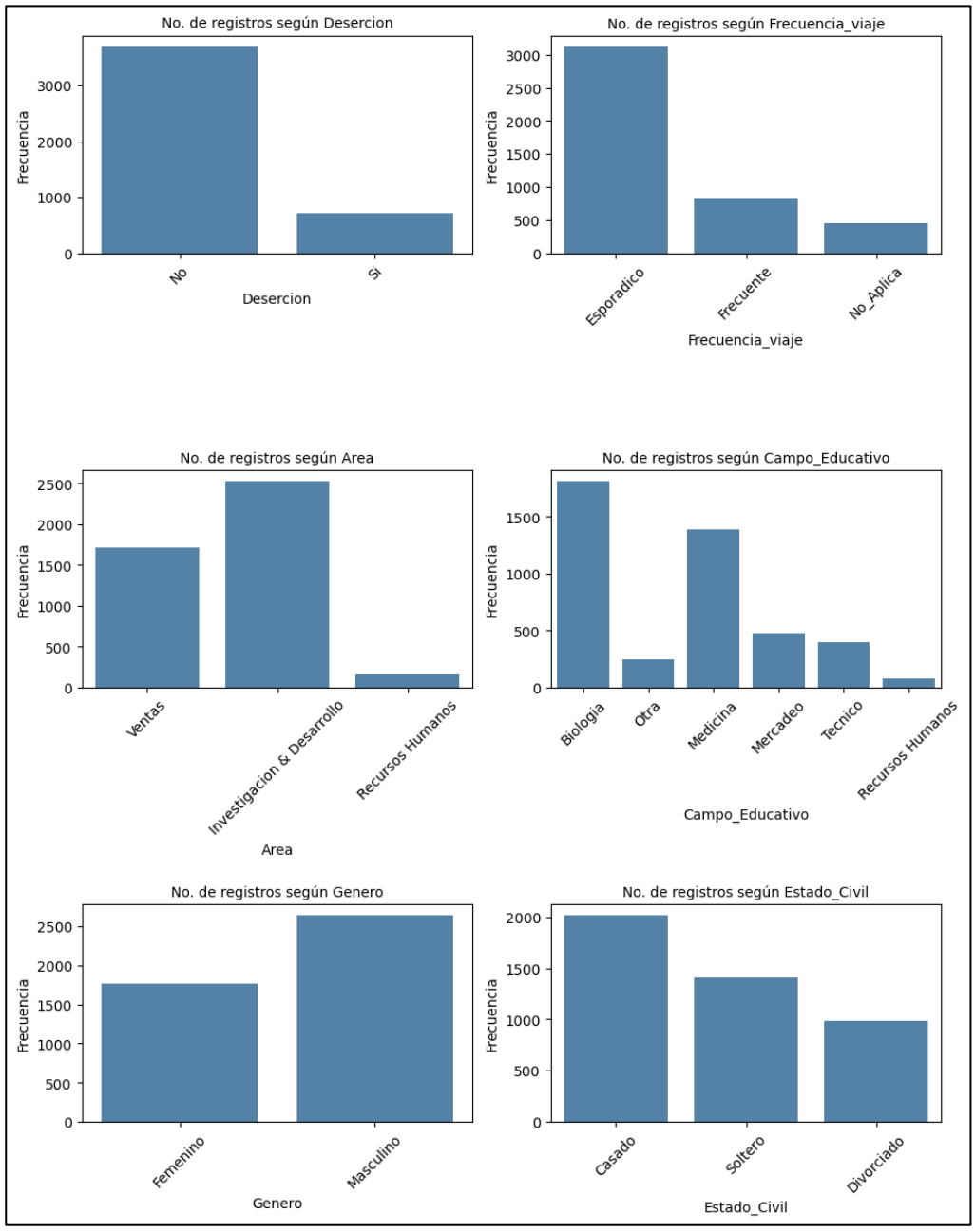
Campo	**Genero**	**Rol**	**Estado_Civil**	**Mayor18**
Cantidad	4.410	4.410	4.410	4.410
Únicos	2	9	3	1
Moda	Masculino	Ejecutivo de ventas	Casado	Y
Frecuencia	2.646	978	2.019	4.410

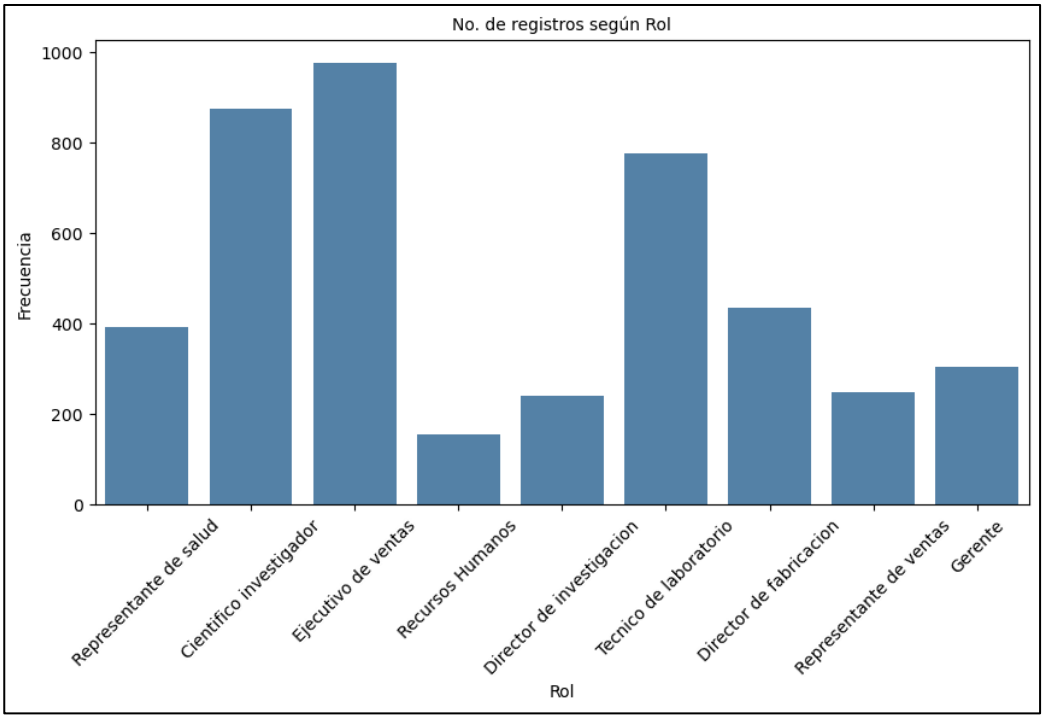
Fuente: Elaboración propia

Para tener un mejor entendimiento de estos datos se realizan gráficos de barras donde es posible observar la distribución de los registros de los campos categóricos.

Figura 7

Gráficos de barras – General Empleados





Fuente: elaboración propia

7.1.3. Hallazgos Iniciales

Los campos numéricos de la base General Empleados en su mayoría no tienen nulos y aparentemente no tienen datos inconsistentes o fuera de los rangos normales para la naturaleza de los campos.

La información permite observar que la edad promedio de los empleados es 37 años y su rango de edad está entre 30 y 43 años; el nivel educativo promedio de los trabajadores es pregrado junto con un promedio de 11 años de experiencia laboral y una media de 7 años trabajando directamente en la compañía.

Respecto a compensación, el salario mensual promedio se encuentra en 65.029 rupias, su valor mínimo 10.090 rupias y el máximo de 199.990 lo cual parece indicar la existencia de una alta dispersión en los ingresos. La base muestra personas con gran antigüedad en la compañía, su valor promedio es de 7 años y presenta valores atípicos con personas entre 30 y 40 años de antigüedad. Existe una brecha significativa en los años que tienen las personas después de ser

promovidas, con un promedio de 2 años, sin embargo, hay algunos valores atípicos de personas que tienen entre 7 y 15 años desde que lograron su última promoción.

El análisis de las variables categóricas permite profundizar la caracterización de los perfiles de los empleados. Se identifica que en su mayoría trabajan en el área de Investigación y Desarrollo, su campo de educación más frecuente es Medicina y Biología y cuentan con una frecuencia de viaje esporádica. Por otro lado, los roles más representativos son: ejecutivo de ventas, científico investigador y técnico de laboratorio. En cuanto a género se encuentra mayor frecuencia de hombres, sin embargo, la diferencia de empleados por género no es significativa.

Teniendo en cuenta los perfiles anteriores, los datos muestran que la compañía realiza capacitaciones frecuentemente a sus empleados. En promedio un empleado recibe 3 capacitaciones al año. Sin embargo, pueden verse valores atípicos de empleados que recibieron hasta 6 entrenamientos durante un año, así como también empleados que no recibieron ninguno.

El cambio de líder no parece ser una práctica frecuente en la compañía, ya que en promedio las personas llevan 4 años con su líder actual y en su mayoría han estado entre 3 y 6 años con el mismo líder.

La matriz de correlación de la figura 6 evidencia algunas relaciones importantes entre variables de la fuente General Empleados. La edad, está altamente correlacionada de manera positiva con los años de experiencia laboral (con un valor de 0.7). De igual manera, el total de años trabajados tiene una alta correlación con los años que la persona ha trabajado en la Farmacéutica (0.6), así como con los años que la persona lleva con el líder actual (0.8). Los años desde la última promoción, están correlacionados positivamente con la experiencia de las personas (0.4) y su antigüedad en la compañía (0.6).

7.2. Base de Datos Encuesta de Empleados

Esta base incluye los resultados de las encuestas realizadas a los trabajadores de la farmacéutica con el objetivo de medir su nivel de satisfacción con el ambiente laboral, su cargo desempeñado, y también su sensación frente al equilibrio entre vida personal y laboral.

7.2.1. Características Generales y Descripción de Campos

La base consta de 4.410 registros, y los demás campos corresponden a las ultimas calificaciones dadas por el empleado a las siguientes preguntas:

- ¿Cuál es su nivel de satisfacción con el entorno de trabajo? (empresa, instalaciones, relación con compañeros, cultura empresarial).
- ¿Qué tan satisfecho está con su cargo actual y las funciones desempeñadas?
- ¿Cómo considera su nivel de equilibrio de vida personal con vida laboral?

A continuación, se describen los campos disponibles en la base:

Tabla 5

Campos – Encuesta de empleados

Campo	Descripción	Tipo de dato	Ejemplo
ID_Empleado	Identificador único del empleado	Numérico	50
Satisfaccion_Entorno	Calificación sobre satisfacción con el ambiente laboral entre 1 y 4.	Numérico	3

	Donde 1 es bajo, 2 es medio, 3 es alto y 4 muy alto. Si no tiene respuesta tiene el valor de “NA”.		
Satisfaccion_Laboral	Calificación sobre satisfacción con el cargo desempeñado entre 1 y 4. Donde 1 es bajo, 2 es medio, 3 es alto y 4 muy alto. Si no tiene respuesta tiene el valor de “NA”.	Numérico	4
Balance_VidaTrabajo	Calificación sobre el nivel de equilibrio personal y laboral entre 1 y 4. Donde 1 es malo, 2 es bueno, 3 es muy bueno y 4 el mejor. Si no tiene respuesta tiene el valor de “NA”.	Numérico	1

Fuente: Elaboración propia

7.2.2. Descripción Estadística

Todos los campos de la base son de tipo numérico, cuyas medidas de tendencia central más importantes se describen en la siguiente tabla.

Tabla 6

Características campos numéricos – Encuestas de empleados

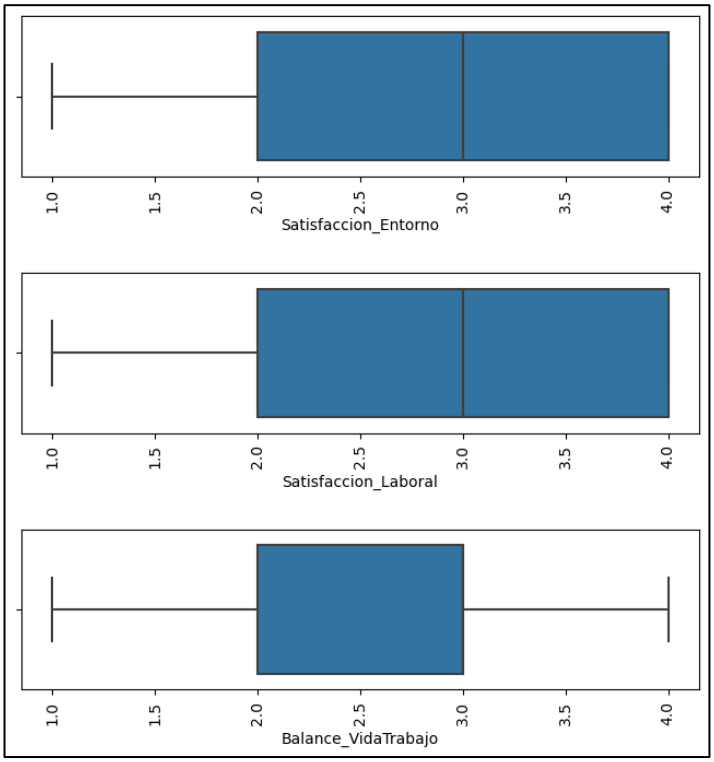
Campo	Satisfaccion_Entorno	Satisfaccion_Laboral	Balance_VidaTrabajo
Cantidad	4.385	4.390	4.372
%no nulos	99,4%	99,5%	99,1%
media	2,72	2,73	2,76
std	1,09	1,10	0,71
min	1	1	1
25%	2	2	2
50%	3	3	3
75%	4	4	3
max	4	4	4

Fuente: elaboración propia

Por medio de diagramas de caja es posible ver la distribución y tendencia central de las variables que conforman la base de encuestas.

Figura 8

Diagramas de caja - Encuestas empleados

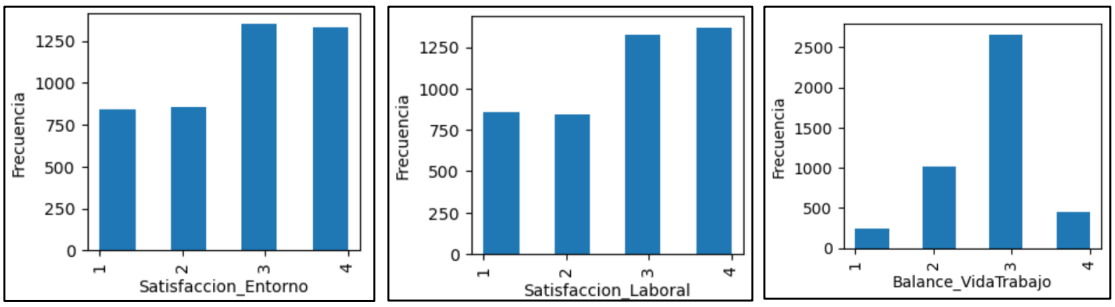


Fuente: elaboración propia

Para tener otra perspectiva de los datos, se construyen histogramas para cada uno de los campos de interés:

Figura 9

Histogramas - Encuestas empleados

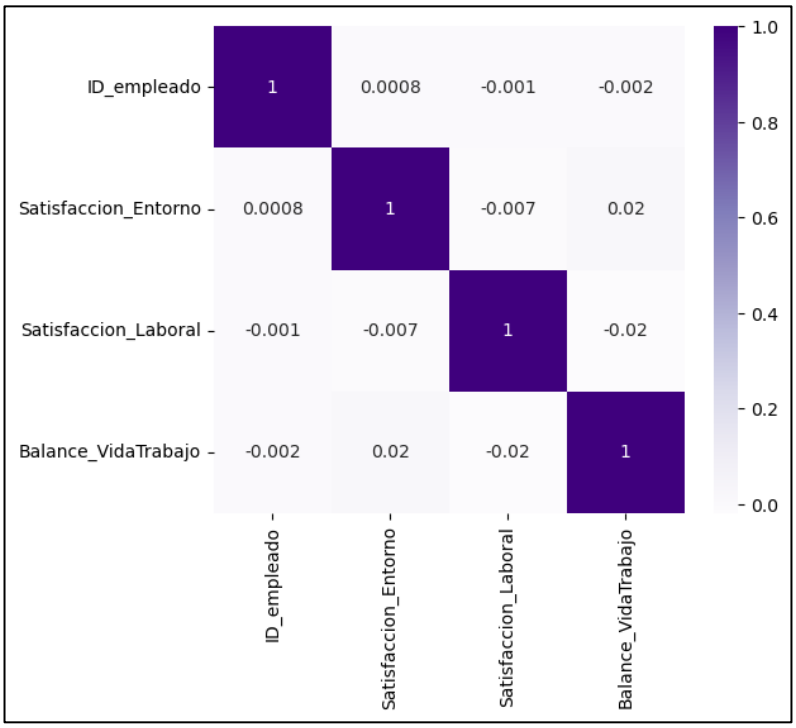


Fuente: elaboración propia

Es posible ver la correlación entre las variables numéricas de esta base en la siguiente figura.

Figura 10

Matriz de correlaciones - Encuesta empleados



Fuente: elaboración propia

7.2.3. Hallazgos Iniciales

La base de encuestas de empleados cuenta con pocas variables, sin embargo, sus registros pueden ser clave a la hora de determinar las causas de renuncia de un empleado. Esta fuente cuenta únicamente con campos numéricos, teniendo en cuenta que uno de ellos es un campo de identificación.

La cantidad de valores nulos en la base de datos es baja, de los 4.410 registros, se tienen solo 25 con datos ausentes de la variable Satisfaccion_Entorno, 20 nulos para la variable Satisfaccion_Laboral y 38 registros sin información para el campo Balance_VidaTrabajo. Estos

datos vacíos, aunque representan menos del 1% de la base, se deben tener en cuenta para procesos de imputación.

Por otra parte, las distribuciones de los campos Satisfaccion_Entorno y Satisfaccion_Laboral tienen un comportamiento casi idéntico, sesgadas a la derecha con valores típicos de 3 y 4, con solo algunas diferencias que se pueden apreciar en sus promedios y desviaciones estándar, así como en la cantidad de nulos ya descrita. Finalmente, el campo Balance_VidaTrabajo tiene una distribución más cercana a la normal, con un leve sesgo a la derecha, presentando valores típicos cercanos a 3. Los campos de la base no presentan valores atípicos.

La matriz de correlación muestra que entre los campos no hay relaciones fuertes, aunque sus distribuciones sean similares, la correlación entre ellas es cercana a cero.

Los histogramas realizados sugieren que la base cuenta con empleados que en su mayoría tienen un nivel de satisfacción bueno frente a su ambiente laboral y el cargo desempeñado ya que los resultados más frecuentes se encuentran entre 3 y 4. La sensación de equilibrio de vida personal y laboral de los empleados de la farmacéutica tiende a ser también positiva al obtener 3 (Muy buena) como calificación más frecuente.

Si bien la base cuenta en su mayoría con datos de personas satisfechas, no todos los empleados están ubicados hacia resultados positivos en términos de satisfacción, por el contrario, cuenta con una proporción de un 25% que presenta cierto nivel de insatisfacción en cada pregunta, lo que hace que la base de encuestas de empleados cause interés dentro del análisis del problema.

7.3. Base de datos Encuesta Líder

Esta fuente de información contiene los resultados de una encuesta que realizan los líderes para evaluar el desempeño y nivel de compromiso de los empleados con su posición. Esta

percepción podría ser un factor importante para determinar que una persona en la farmacéutica está próxima a abandonar la compañía.

7.3.1. Características Generales y Descripción de Campos

Cuenta con 4.410 registros que detallan el resultado por empleado de la calificación dada por sus líderes. Esta encuesta consta de las siguientes preguntas:

- ¿Qué tan identificado está el empleado con su posición? O ¿Cuál es el nivel de participación del empleado en su posición?
- Calificación de desempeño del empleado en el año anterior.

A continuación, se detallan los campos que componen la base:

Tabla 7

Campos - Encuestas Líder

Campo	Descripción	Tipo de dato	Ejemplo
ID_Empleado	Identificador único del empleado	Numérico	1564
CompromisoTrabajo	Nivel de participación del empleado en su posición entre 1 y 4. Donde 1 es baja, 2 es media, 3 es alta y 4 muy alta. Si no tiene respuesta tiene el valor de “NA”.	Numérico	1

Rendimiento	Calificación de desempeñó del empleado en el año anterior entre 1 y 4. Donde 1 es bajo, 2 es bueno, 3 es excelente y 4 excepcional. Si no tiene respuesta tiene el valor de "NA".	Numérico	4

Fuente: Elaboración propia

7.3.2. Descripción Estadística

A continuación, se exponen las características principales de los campos numéricos de la base de encuestas de líder.

Tabla 8

Características de campos numéricos - Encuestas Líder

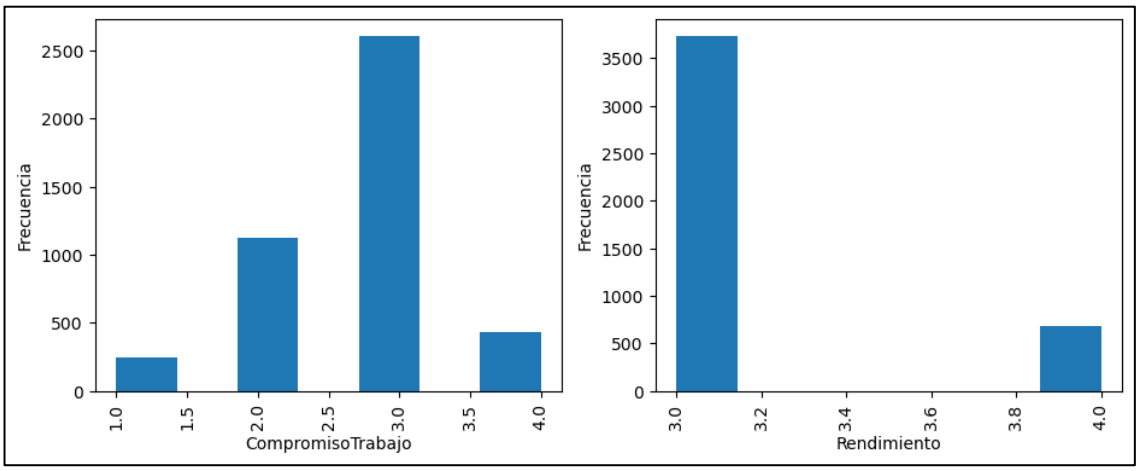
Campo	**CompromisoTrabajo**	**Rendimiento**
Cantidad	4.410	4.410
%no nulos	100%	100%
media	2,73	3,15
std	0,71	0,36
min	1	3
25%	2	3
50%	3	3
75%	3	3
max	4	4

Fuente: Elaboración propia

Los siguientes histogramas permiten observar la distribución de los valores de los campos numéricos de las dos preguntas que conforman la encuesta.

Figura 11

Histogramas - Encuesta Líder



Fuente: elaboración propia

La correlación para las dos variables numéricas es de 0.0106, un valor casi nulo.

7.3.3. Hallazgos Iniciales

La encuesta posee 3 campos numéricos, uno de ellos es un identificador de empleado, por lo que contamos con solo 2 variables de resultados. Sin embargo, estos campos ayudan a entender el desempeño de los empleados junto con su nivel de participación en su posición actual.

La base no cuenta con datos nulos ni con valores atípicos, lo cual resulta beneficioso para el análisis.

En cuanto a los histogramas de las variables, el campo CompromisoTrabajo tiene una distribución levemente sesgada a la izquierda, teniendo la nota de 3 (asociada a la característica “alta” en el nivel de participación e identificación con el rol), además, es la calificación más frecuente. Esto podría indicar que los trabajadores en general tienen una buena relación con su rol

y por tanto podrían mantener un nivel de compromiso alto y estable con la compañía. Sin embargo, aquellos empleados con bajo nivel de involucramiento pueden ser de interés para el estudio.

Por otro lado, el campo Rendimiento muestra una distribución cargada a la derecha (con ausencia de valores entre 1 y 2) y con valor más frecuente de 3 (que va ligado a la característica "Excelente" en el desempeño del empleado en el año anterior).

Las dos variables presentan una correlación casi nula, sugiriendo que no hay relación entre ellas, algo importante para evitar redundancia de información en el análisis del problema. Sin embargo, es interesante ya que los empleados obtuvieron una calificación de desempeño alta (entre 3 y 4) aun teniendo bajos niveles de participación para aquellos con niveles entre 1 y 2.

7.4. Bases de Datos de Entrada y de Salida de los Empleados

Estas fuentes almacenan el registro de hora de entrada y de hora de salida para cada uno de los días del año anterior. Contiene el momento exacto en el que cada trabajador ingresó y salió de la farmacéutica para cumplir con sus labores, así como también registra su ausencia.

Las bases están almacenadas en dos archivos diferentes, una para horas de entrada y otra para horas de salida.

7.4.1. Características Generales y Descripción de Campos

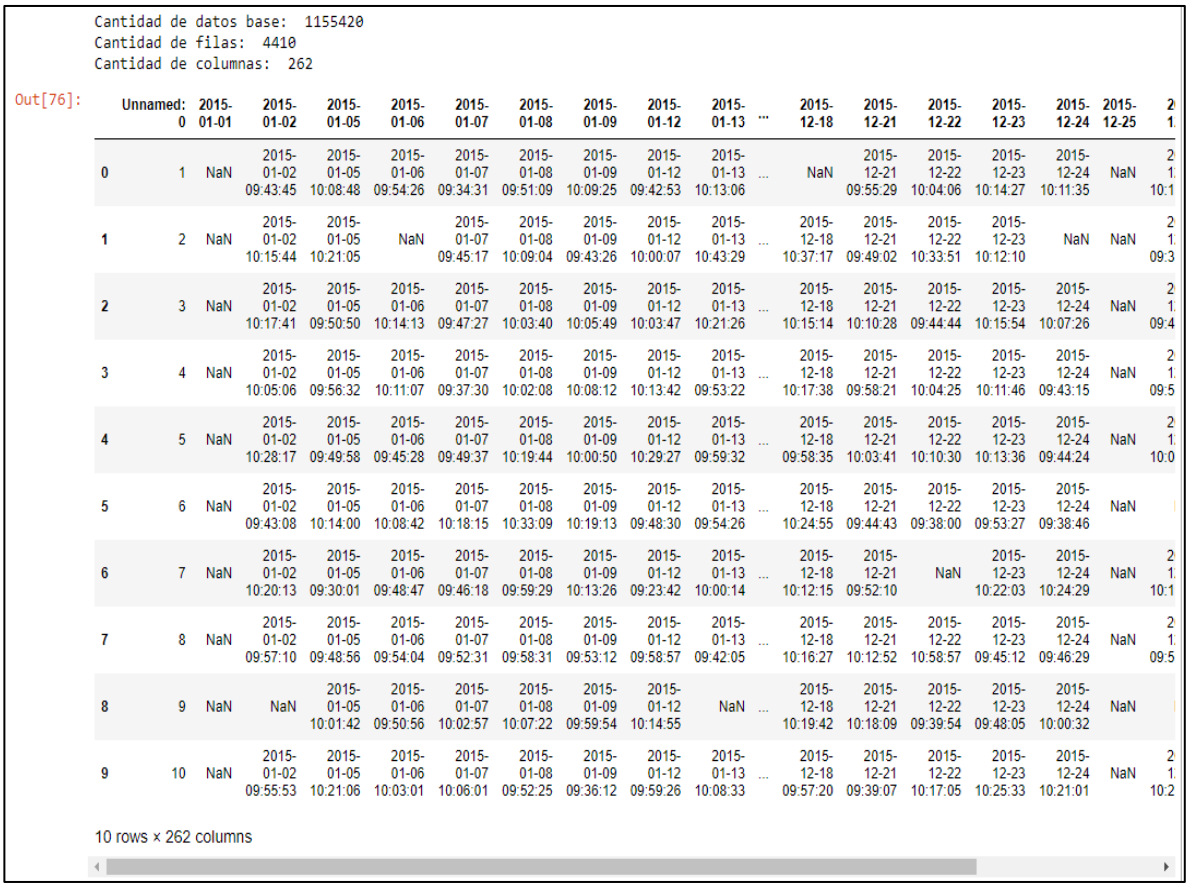
Ambas fuentes de información cuentan con 4.410 registros, asociados cada uno a un empleado de la farmacéutica y con la misma estructura. Las bases están conformadas por 262 columnas, la primera de ellas almacena el número de identificación del empleado (ID_Empleado), mientras que las otras 261 columnas corresponden a los días laborales del año. Cabe resaltar que no se tienen 365 campos para los días del año debido a que la farmacéutica solamente tiene en cuenta los días laborales entre lunes y viernes. Estos campos contienen información de tipo fecha

con formato AAAA-MM-DD HH:MM:SS y para ausencias se tienen registros marcados como “NA”.

A continuación, es posible evidenciar una muestra de los datos de las bases originales.

Figura 12

Muestra de datos - Entrada de empleados



```
Cantidad de datos base:  1155420
Cantidad de filas:  4410
Cantidad de columnas:  262
```

Out[76]:

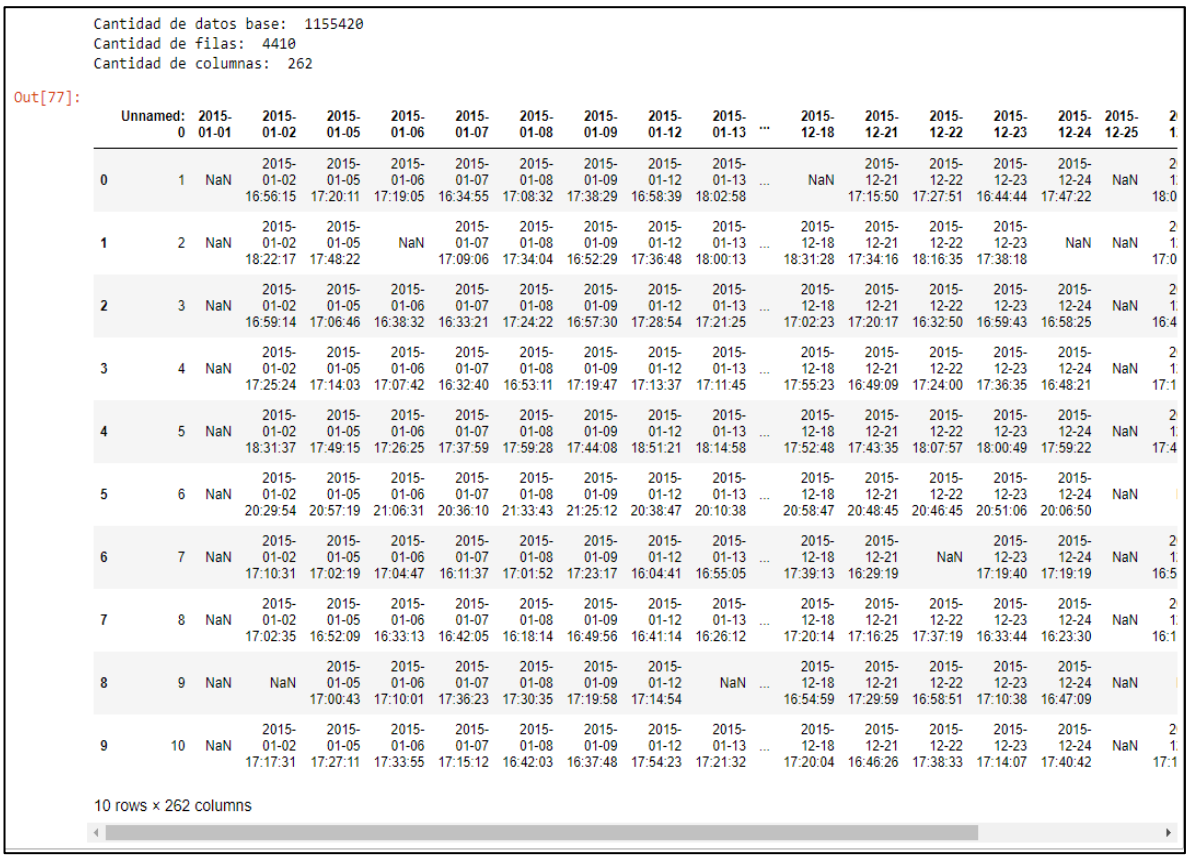
	Unnamed: 0	2015-01-01	2015-01-02	2015-01-05	2015-01-06	2015-01-07	2015-01-08	2015-01-09	2015-01-12	2015-01-13	...	2015-12-18	2015-12-21	2015-12-22	2015-12-23	2015-12-24	2015-12-25	2 1
0	1	NaN	2015-01-02 09:43:45	2015-01-05 10:08:48	2015-01-06 09:54:26	2015-01-07 09:34:31	2015-01-08 09:51:09	2015-01-09 10:09:25	2015-01-12 09:42:53	2015-01-13 10:13:06	...	NaN	2015-12-21 09:55:29	2015-12-22 10:04:06	2015-12-23 10:14:27	2015-12-24 10:11:35	NaN	2 1 10:1
1	2	NaN	2015-01-02 10:15:44	2015-01-05 10:21:05	NaN	2015-01-07 09:45:17	2015-01-08 10:09:04	2015-01-09 09:43:26	2015-01-12 10:00:07	2015-01-13 10:43:29	...	2015-12-18 10:37:17	2015-12-21 09:49:02	2015-12-22 10:33:51	2015-12-23 10:12:10	NaN	NaN	2 1 09:3
2	3	NaN	2015-01-02 10:17:41	2015-01-05 09:50:50	2015-01-06 10:14:13	2015-01-07 09:47:27	2015-01-08 10:03:40	2015-01-09 10:05:49	2015-01-12 10:03:47	2015-01-13 10:21:26	...	2015-12-18 10:15:14	2015-12-21 10:10:28	2015-12-22 09:44:44	2015-12-23 10:15:54	2015-12-24 10:07:26	NaN	2 1 09:4
3	4	NaN	2015-01-02 10:05:06	2015-01-05 09:56:32	2015-01-06 10:11:07	2015-01-07 09:37:30	2015-01-08 10:02:08	2015-01-09 10:08:12	2015-01-12 10:13:42	2015-01-13 09:53:22	...	2015-12-18 10:17:38	2015-12-21 09:58:21	2015-12-22 10:04:25	2015-12-23 10:11:46	2015-12-24 09:43:15	NaN	2 1 09:5
4	5	NaN	2015-01-02 10:28:17	2015-01-05 09:49:58	2015-01-06 09:45:28	2015-01-07 09:49:37	2015-01-08 10:19:44	2015-01-09 10:00:50	2015-01-12 10:29:27	2015-01-13 09:59:32	...	2015-12-18 09:58:35	2015-12-21 10:03:41	2015-12-22 10:10:30	2015-12-23 10:13:36	2015-12-24 09:44:24	NaN	2 1 10:0
5	6	NaN	2015-01-02 09:43:08	2015-01-05 10:14:00	2015-01-06 10:08:42	2015-01-07 10:18:15	2015-01-08 10:33:09	2015-01-09 10:19:13	2015-01-12 09:48:30	2015-01-13 09:54:26	...	2015-12-18 10:24:55	2015-12-21 09:44:43	2015-12-22 09:38:00	2015-12-23 09:53:27	2015-12-24 09:38:46	NaN	
6	7	NaN	2015-01-02 10:20:13	2015-01-05 09:30:01	2015-01-06 09:48:47	2015-01-07 09:46:18	2015-01-08 09:59:29	2015-01-09 10:13:26	2015-01-12 09:23:42	2015-01-13 10:00:14	...	2015-12-18 10:12:15	2015-12-21 09:52:10	NaN	2015-12-23 10:22:03	2015-12-24 10:24:29	NaN	2 1 10:1
7	8	NaN	2015-01-02 09:57:10	2015-01-05 09:48:56	2015-01-06 09:54:04	2015-01-07 09:52:31	2015-01-08 09:58:31	2015-01-09 09:53:12	2015-01-12 09:58:57	2015-01-13 09:42:05	...	2015-12-18 10:16:27	2015-12-21 10:12:52	2015-12-22 10:58:57	2015-12-23 09:45:12	2015-12-24 09:46:29	NaN	2 1 09:5
8	9	NaN	NaN	2015-01-05 10:01:42	2015-01-06 09:50:56	2015-01-07 10:02:57	2015-01-08 10:07:22	2015-01-09 09:59:54	2015-01-12 10:14:55	NaN	...	2015-12-18 10:19:42	2015-12-21 10:18:09	2015-12-22 09:39:54	2015-12-23 09:48:05	2015-12-24 10:00:32	NaN	
9	10	NaN	2015-01-02 09:55:53	2015-01-05 10:21:06	2015-01-06 10:03:01	2015-01-07 10:06:01	2015-01-08 09:52:25	2015-01-09 09:36:12	2015-01-12 09:59:26	2015-01-13 10:08:33	...	2015-12-18 09:57:20	2015-12-21 09:39:07	2015-12-22 10:17:05	2015-12-23 10:25:33	2015-12-24 10:21:01	NaN	2 1 10:2

10 rows × 262 columns

Fuente: elaboración propia

Figura 13

Muestra de datos - Salida de empleados



```
Cantidad de datos base:  1155420
Cantidad de filas:  4410
Cantidad de columnas:  262
```

Out[77]:

	Unnamed: 0	2015-01-01	2015-01-02	2015-01-05	2015-01-06	2015-01-07	2015-01-08	2015-01-09	2015-01-12	2015-01-13	...	2015-12-18	2015-12-21	2015-12-22	2015-12-23	2015-12-24	2015-12-25	2 1
0	1	NaN	2015-01-02 16:56:15	2015-01-05 17:20:11	2015-01-06 17:19:05	2015-01-07 16:34:55	2015-01-08 17:08:32	2015-01-09 17:38:29	2015-01-12 16:58:39	2015-01-13 18:02:58	...	NaN	2015-12-21 17:15:50	2015-12-22 17:27:51	2015-12-23 16:44:44	2015-12-24 17:47:22	NaN	2 1 18:0
1	2	NaN	2015-01-02 18:22:17	2015-01-05 17:48:22	NaN	2015-01-07 17:09:06	2015-01-08 17:34:04	2015-01-09 16:52:29	2015-01-12 17:36:48	2015-01-13 18:00:13	...	2015-12-18 18:31:28	2015-12-21 17:34:16	2015-12-22 18:16:35	2015-12-23 17:38:18	NaN	NaN	2 1 17:0
2	3	NaN	2015-01-02 16:59:14	2015-01-05 17:06:46	2015-01-06 16:38:32	2015-01-07 16:33:21	2015-01-08 17:24:22	2015-01-09 16:57:30	2015-01-12 17:28:54	2015-01-13 17:21:25	...	2015-12-18 17:02:23	2015-12-21 17:20:17	2015-12-22 16:32:50	2015-12-23 16:59:43	2015-12-24 16:58:25	NaN	2 1 16:4
3	4	NaN	2015-01-02 17:25:24	2015-01-05 17:14:03	2015-01-06 17:07:42	2015-01-07 16:32:40	2015-01-08 16:53:11	2015-01-09 17:19:47	2015-01-12 17:13:37	2015-01-13 17:11:45	...	2015-12-18 17:55:23	2015-12-21 16:49:09	2015-12-22 17:24:00	2015-12-23 17:36:35	2015-12-24 16:48:21	NaN	2 1 17:1
4	5	NaN	2015-01-02 18:31:37	2015-01-05 17:49:15	2015-01-06 17:26:25	2015-01-07 17:37:59	2015-01-08 17:59:28	2015-01-09 17:44:08	2015-01-12 18:51:21	2015-01-13 18:14:58	...	2015-12-18 17:52:48	2015-12-21 17:43:35	2015-12-22 18:07:57	2015-12-23 18:00:49	2015-12-24 17:59:22	NaN	2 1 17:4
5	6	NaN	2015-01-02 20:29:54	2015-01-05 20:57:19	2015-01-06 21:06:31	2015-01-07 20:36:10	2015-01-08 21:33:43	2015-01-09 21:25:12	2015-01-12 20:38:47	2015-01-13 20:10:38	...	2015-12-18 20:58:47	2015-12-21 20:48:45	2015-12-22 20:46:45	2015-12-23 20:51:06	2015-12-24 20:06:50	NaN	
6	7	NaN	2015-01-02 17:10:31	2015-01-05 17:02:19	2015-01-06 17:04:47	2015-01-07 16:11:37	2015-01-08 17:01:52	2015-01-09 17:23:17	2015-01-12 16:04:41	2015-01-13 16:55:05	...	2015-12-18 17:39:13	2015-12-21 16:29:19	NaN	2015-12-23 17:19:40	2015-12-24 17:19:19	NaN	2 1 16:5
7	8	NaN	2015-01-02 17:02:35	2015-01-05 16:52:09	2015-01-06 16:33:13	2015-01-07 16:42:05	2015-01-08 16:18:14	2015-01-09 16:49:56	2015-01-12 16:41:14	2015-01-13 16:26:12	...	2015-12-18 17:20:14	2015-12-21 17:16:25	2015-12-22 17:37:19	2015-12-23 16:33:44	2015-12-24 16:23:30	NaN	2 1 16:1
8	9	NaN	NaN	2015-01-05 17:00:43	2015-01-06 17:10:01	2015-01-07 17:36:23	2015-01-08 17:30:35	2015-01-09 17:19:58	2015-01-12 17:14:54	NaN	...	2015-12-18 16:54:59	2015-12-21 17:29:59	2015-12-22 16:58:51	2015-12-23 17:10:38	2015-12-24 16:47:09	NaN	
9	10	NaN	2015-01-02 17:17:31	2015-01-05 17:27:11	2015-01-06 17:33:55	2015-01-07 17:15:12	2015-01-08 16:42:03	2015-01-09 16:37:48	2015-01-12 17:54:23	2015-01-13 17:21:32	...	2015-12-18 17:20:04	2015-12-21 16:46:26	2015-12-22 17:38:33	2015-12-23 17:14:07	2015-12-24 17:40:42	NaN	2 1 17:1

10 rows × 262 columns

Fuente: elaboración propia

Las bases anteriores contienen datos de horarios de entrada y salida de los empleados de la Farmacéutica que podrá ser utilizada para identificar ausentismos, así como retrasos y sobrecarga laboral. Para obtener esta información será necesario realizar transformaciones que permitan extraer resultados relacionados al comportamiento de los empleados y tendencias de deserción. Del mismo modo, será necesario crear indicadores adicionales calculados a partir de los datos anteriores para extraer resultados de estadística descriptiva y analizar patrones de asistencia de los empleados. Este tratamiento de datos será detallado en la sección posterior.

8. Preparación de la información

Con el contexto inicial de las fuentes de información, se identifica la necesidad de realizar un tratamiento adicional sobre los datos, enfocado en gestionar en primera instancia los problemas de calidad encontrados de tal forma que no impacte los procesos posteriores de minería de datos. Adicionalmente, es necesario crear un modelo operativo centralizado para su almacenamiento y escalabilidad, que facilite la gestión del equipo responsable de gobierno de datos dentro de la farmacéutica al elevar la homogenización y estandarización de los campos y acoplarse al despliegue de recursos especializados en el futuro.

8.1. Calidad de los datos

A partir del entendimiento de las fuentes, se conoce en primer lugar que cada una de las bases cuenta con el registro de 4.410 empleados. Para esta totalidad de registros y siguiendo algunas dimensiones de calidad descritas en (DAMA International, 2020) es posible establecer una evaluación de los siguientes indicadores de calidad de datos:

- Completitud: Grado en el que los atributos requeridos están presentes en la base. Se mide como el porcentaje de datos cuyos valores están informados sobre el total de registros de la base.
- Validez: Grado de conformidad de los datos de acuerdo con el formato deseado, tipo de dato y rango de valores. Se calcula como el porcentaje de datos que cumple con los requerimientos válidos sobre el total de registros de la base.
- Unicidad: Es el porcentaje de valores únicos que existen. Se puede calcular mediante la división entre el número de valores únicos sobre el total de registros. Esta dimensión se aplica solamente cuando no se requieren duplicados en un campo.

- Precisión: Grado en que los datos representan la realidad según contexto del negocio y significado de los campos. Se puede calcular como porcentaje de valores precisos sobre el total de registros de la base.

De acuerdo con las dimensiones de calidad mencionadas, estos son los resultados para cada una de las bases disponibles.

Tabla 9

Resultados de calidad sobre base General Empleados

Campo	Completitud	Validez	Unicidad	Precisión
ID empleado	100,0%	100,0%	100,0%	100,0%
Edad	100,0%	100,0%	1,0%	100,0%
Desercion	100,0%	100,0%	0,0%	100,0%
Frecuencia_viaje	100,0%	100,0%	0,1%	100,0%
Area	100,0%	100,0%	0,1%	100,0%
Distancia_Casa	100,0%	100,0%	0,7%	100,0%
Nivel_Educativo	100,0%	100,0%	0,1%	100,0%
Campo_Educativo	100,0%	100,0%	0,1%	100,0%
Recuento_Empleado	100,0%	100,0%	0,0%	100,0%
Genero	100,0%	100,0%	0,0%	100,0%
Nivel_Rol	100,0%	100,0%	0,1%	100,0%
Rol	100,0%	100,0%	0,2%	100,0%
Estado_Civil	100,0%	100,0%	0,1%	100,0%
Salario_Mensual	100,0%	100,0%	30,6%	100,0%

Companias	99,6%	100,0%	0,2%	100,0%
Mayor18	100,0%	100,0%	0,0%	100,0%
Porcentaje_Aumento_Salarial	100,0%	100,0%	0,3%	100,0%
Horario	100,0%	100,0%	0,0%	100,0%
Nivel_Opcion_Acciones	100,0%	100,0%	0,1%	100,0%
Anos_Trabajados	99,8%	100,0%	0,9%	100,0%
Entrenamientos	100,0%	100,0%	0,2%	100,0%
Antiguedad	100,0%	100,0%	0,8%	100,0%
Anos_Ultima_Promocion	100,0%	100,0%	0,4%	100,0%
Anos_Lider_Actual	100,0%	100,0%	0,4%	100,0%

Fuente: elaboración propia

Tabla 10

Resultados de calidad sobre base Encuestas empleados

Campo	Completitud	Validez	Unicidad	Precision
ID_Empleado	100%	100,0%	100,0%	100,0%
Satisfaccion_Entorno	99,4%	100,0%	0,1%	100,0%
Satisfaccion_Laboral	99,5%	100,0%	0,1%	100,0%
Balance_VidaTrabajo	99,1%	100,0%	0,1%	100,0%

Fuente: elaboración propia

Tabla 11

Resultados de calidad sobre base Encuestas Líder

Campo	Completitud	Validez	Unicidad	Precision
ID_Empleado	100%	100,0%	100,0%	100,0%
CompromisoTrabajo	100%	100,0%	0,1%	100,0%
Rendimiento	100%	100,0%	0,0%	100,0%

Fuente: elaboración propia

Por otra parte, como ya se especificó en la exploración de fuentes, las bases de entrada y salida de empleados constan de 262 campos (1 de ID de empleado y 261 para días de asistencia en el año). Todos los valores de las columnas de fecha de asistencia tienen un formato consistente (fecha-hora o NA para registros de inasistencia) y los identificadores tienen calidad de 100% teniendo en cuenta las dimensiones medidas.

8.2. Tratamiento de Inconsistencias en las Fuentes

Las inconsistencias encontradas están enfocadas en la completitud de los datos en los campos de Companias y Anos_Trabajados para la base de General Empleados, y para la base de encuestas de empleados en los campos de Satisfaccion_Entorno, Satisfaccion_Laboral, Balance_VidaTrabajo. Sin embargo, la completitud de los campos mencionados supera el 99% por lo que no representa un impacto mayor en la robustez de la información.

Para efectos de almacenamiento y análisis del estado actual de la farmacéutica los campos con valores nulos se dejarán sin modificación, teniendo en cuenta que la ausencia de empleados puede ser un factor crítico en los tableros de análisis que usarán los ejecutivos de la farmacéutica y el área de recursos humanos. Sin embargo, para las fases que involucran el uso de modelos analíticos para estimación de probabilidad de deserción, los empleados con valores ausentes no se

tendrán en cuenta ya que representan aproximadamente el 1% de los datos y podrían influir negativamente en el entrenamiento de los modelos. En este orden de ideas, para bases enfocadas en procesos analíticos se usará información de solamente 4.300 empleados con todos los datos diligenciados.

8.3. Transformaciones Iniciales y Generación de Nuevos Indicadores Clave

Dentro de las bases analizadas se identificó que las bases de entradas y salidas de los empleados contaban con un formato inadecuado para realizar un apropiado análisis que contribuyera al proyecto. Dado lo anterior, se decidió realizar un cambio en las dimensiones de las bases de hora entrada y hora salida para hallar nuevos indicadores que permitan entender una dimensión adicional de los empleados de la farmacéutica.

Se transformaron los registros de cada día para obtener las variables fecha, hora_entrada y hora_salida. Estas variables, junto con el ID de los empleados hacen posible calcular el promedio de horas trabajadas por empleado, días de ausentismo y porcentaje de ausentismo durante el año. Para el cálculo de estos nuevos campos se realizó el siguiente proceso.

- Promedio de horas trabajadas: se toma la duración entre hora entrada y salida por empleado, se excluyen los días en los que no asiste a trabajar y se calcula el promedio de horas por empleado.
- Días y porcentaje de ausentismo: en la base original el ausentismo se identificaba como un registro cuyo valor era NA. Para poder contabilizarlo se creó una variable dummy con valor 1 cuando es NA y 0 cuando es diferente a este valor. Luego de hallar el valor total en días donde el empleado estuvo ausente se calculó el porcentaje de estos sobre los días laborales del año.

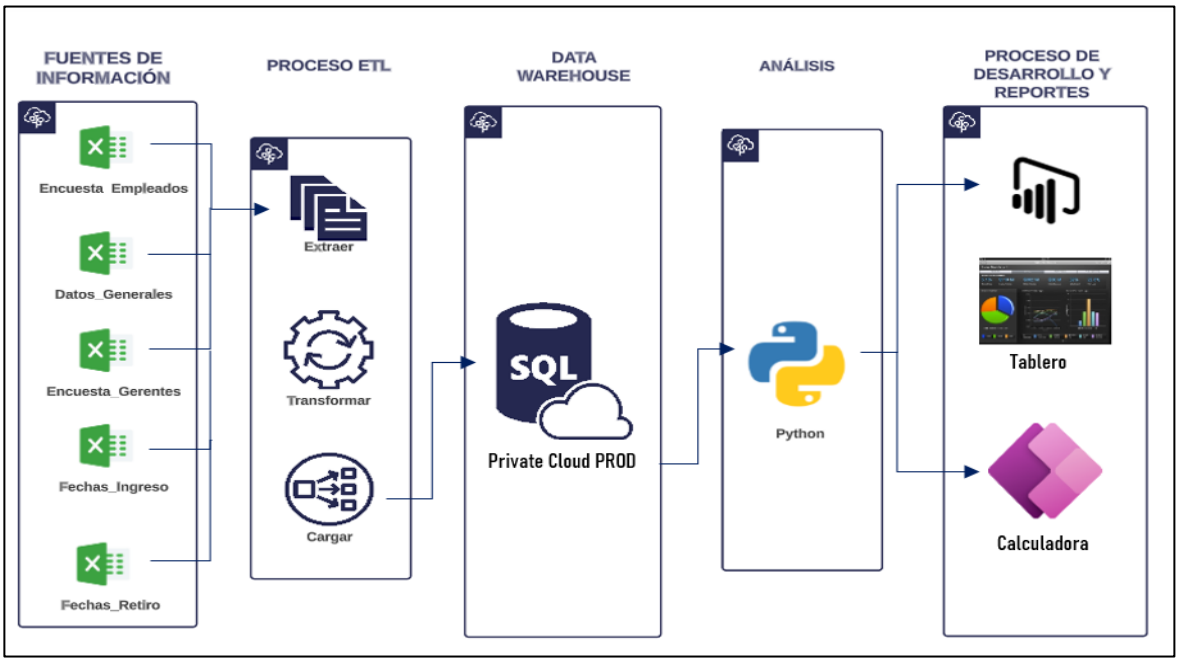
Estas variables servirán de insumo para analizar la sobrecarga laboral, así como también, determinar los empleados con un alto número de ausentismo. Se espera que estas variables sean de gran importancia a la hora de desarrollar el sistema de alerta por empleado para identificar su probabilidad de deserción.

8.4. Modelo Operativo Centralizado

Se propone la siguiente arquitectura como punto de partida para la organización del manejo de la información desde su extracción, pasando por su minería de datos correspondiente y finalizando en entregables orientados a la toma de decisiones por parte de los ejecutivos de la farmacéutica, sustentadas con datos.

Figura 14

Diagrama de Arquitectura Data Warehouse



Fuente: elaboración propia

Como se observa, las fuentes principales del proceso de deserción vienen desde diferentes sistemas, por lo que es necesaria la inclusión dentro del flujo de datos de un proceso de extracción

de estas fuentes, transformación, limpieza y generación de nuevos indicadores y cargue de los datos en un almacén central (proceso ETL). Se propone que el tratamiento de datos sea estandarizado y desarrollado por un equipo de ingenieros de datos específico, su principal objetivo será unificar estructuras e información en un almacén de datos.

Teniendo este enfoque, la metodología Kimball se adapta a la arquitectura esperada. Esta metodología se centra en un diseño progresivo de abajo hacia arriba enmarcando un almacén de datos a nivel dimensional que divide la información en una tabla de hechos (datos de tipo transaccional) y sus respectivas dimensiones (que se apoyan o sustentan en la tabla hechos) (Naeem, 2020). Partiendo del uso de esta metodología, el almacén de datos será diseñado con una arquitectura tipo estrella, la cual permitirá una manipulación de los datos más rápida. Este modelo facilita el uso de la información debido a que aplica tanto para datos normalizados como datos sin normalizar, además de permitir una recuperación rápida de información.

Después de realizar el análisis de las variables disponibles, se encuentran 13 dimensiones para el modelo estrella. Cada una de ellas con llaves foráneas nuevas para datos descriptivos. Estas dimensiones tienen las siguientes características.

Tabla 12

Dimensiones Esquema Estrella

Nombre Dimensión	Descripción	Esquema
DIM_TIEMPO	Permite asociar y llevar un control de los campos tipo fecha de los datos	DIA (SMALLINT) FECHA (DATE) AÑO (SMALLINT) TRIMESTRE (SMALLINT) MES (SMALLINT)

	que componen el Data Warehouse.	
DIM_SATISFACCION_ENTORNO	Facilita la relación o conexión de cada uno de los nombres niveles de satisfacción con el entorno de trabajo de cada empleado que se encuentre asociada a la tabla de hechos.	CODIGO_EN_SAT (INT) NOMBRE_EN_SAT (NVARCHAR (30))
DIM_NIVEL_EDUCATIVO	Conecta los nombres de los niveles educativos de cada uno de los datos de los empleados con los datos alojados en la tabla hechos.	COD_NIV_EDUCATIVO (INT) NOMBRE_NIV_EDUCATIVO NVARCHAR (20)
DIM_BALANCE_ VIDATRABAJO	Brinda la posibilidad de identificar las categorías que están	COD_TR_BAL (INT) NOMBRE_TR_BAL NVARCHAR (30)

	asociadas al campo Balance_VidaTrabajo con los datos alojados en la tabla hechos.	
DIM_RENDIMIENTO	Asocia los nombres de los niveles de rendimiento para el año estudiado con los datos que se encuentran en la tabla hechos.	COD_RENDIMIENTO(INT) NOMBRE_RENDIMIENTO NVARCHAR (30)
DIM_COMPROMISO_TRABAJO	Relaciona directamente los nombres de los niveles de compromiso de trabajo con los datos existentes en la tabla hechos.	ID_COMP_TR (INT) NOMBRE_COMP_TR NVARCHAR (30)
DIM_SATISFACCION_LABORAL	Relaciona los nombres de los niveles de	COD_SAT_LABORAL (INT) NOMBRE_SAT_LABORAL NVARCHAR (30)

	satisfacción de trabajo con los datos almacenados en la tabla hechos.	
DIM_ROL	Vincula los nombres de los roles de cargo con los datos existentes en la tabla de hechos.	COD_ROL (INT) NOMBRE_ROL NVARCHAR (40)
DIM_AREA	Conecta los nombres de las áreas existentes en la compañía con los datos alojados en la tabla hechos.	COD_AREA (INT) NOMBRE_AREA NVARCHAR (60)
DIM_GENERO	Enlaza el género de los empleados con la tabla hechos.	COD_GEN (INT) NOMBRE_GEN NVARCHAR (40)
DIM_ESTADO_CIVIL	Relaciona el campo estado civil existente para cada registro que se encuentran en la tabla de hecho.	COD_ES_C (INT) NOMBRE_ES_C NVARCHAR (40)

DIM_FRECUENCIA_VIAJE	Realiza la vinculación entre el campo frecuencia de viaje con los datos existentes en la tabla de hechos.	COD_FV (INT) NOMBRE_FV NVARCHAR (40)
DIM_NIVEL_ROL	Vincula el campo nivel de cargo con los datos almacenados en la tabla hechos.	COD_N_ROL (INT) NOMBRE_N_ROL NVARCHAR (60)

Fuente: elaboración propia

Posteriormente, se identifican los campos que harán parte de la tabla hechos. Esta entidad será construida mediante la consolidación de los campos de las bases de General Empleados, Encuestas de Empleados, Encuestas de Líderes, Entradas y Salidas. A continuación, se muestran las características de dicha tabla.

Tabla 13

Tabla de hechos Esquema Estrella

FACT_DATA_GENERAL_DESERC	**Tipo Dato**
ID_empleado	int
Edad	int
Desercion	nvarchar(20)
Frecuencia_viaje	int

Area	int
Distancia_Casa	int
Nivel_Educativo	int
Campo_Educativo	nvarchar(20)
Recuento_Empleado	int
Genero	int
Nivel_Rol	int
Rol	int
Estado_Civil	int
Salario_Mensual	money
Companias	int
Mayor18	nvarchar(10)
Porcentaje_Aumento_Salarial	decimal(18,0)
Horario	int
Nivel_Opcion_Acciones	int
Anos_Trabajados	int
Entrenamientos	int
Antiguedad	int
Anos_Ultima_Promocion	int
Anos_Lider_Actual	int
Satisfaccion_Entorno	int
Satisfaccion_Laboral	int
Balance_VidaTrabajo	int

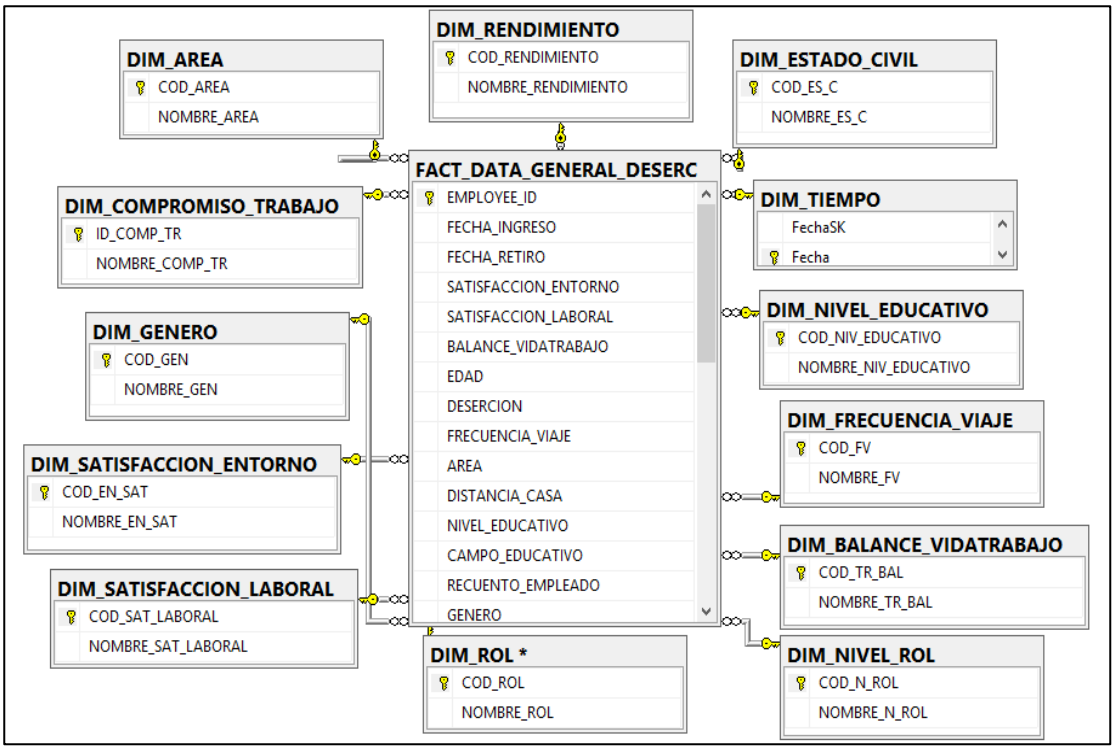
CompromisoTrabajo	int
Rendimiento	int
Prom_Horas	int
Porcentaje_Ausentismo	decimal(18,0)
Ausentismo	int

Fuente: elaboración propia

Al finalizar el ejercicio de modelado, el diseño final del almacén de datos con esquema estrella es el siguiente.

Figura 15

Esquema relacional de Data Warehouse



Fuente: elaboración propia

En referencia a la infraestructura que se utilizará para almacenar los datos, se propone un sistema alojado en la nube, con un motor de base de datos SQL. Esto para adquirir las ventajas de

gestión de acceso, administración de privacidad y escalabilidad. Como la información disponible actualmente no es de gran volumen se tiene en cuenta un margen de crecimiento del 30% sobre los datos como punto de referencia para capacidad de almacenamiento. Al estar en la nube, también ofrece las bondades de una política de respaldo de datos mínima que incluye creación de puntos de restauración diarios y mensuales.

Con la información centralizada, es posible hacer uso de datos para analítica por medio de herramientas especializadas. Se sugiere el establecimiento de un equipo conformado por analistas y científicos de datos para realizar estos procesos de minería de información. Se recomienda el uso de herramientas específicas de analítica basadas en Python para generar modelos y hallazgos sobre los datos. A su vez, se plantea en la arquitectura llegar a usuarios finales por medio de tableros de indicadores desarrollados en Power BI y alojados de tal manera que los usuarios puedan acceder mediante cualquier explorador, conectados en la red de la compañía (implementación de Power BI Server).

9. Análisis de la Problemática desde el Punto de Vista de la Información

Mediante métodos estadísticos multivariados y análisis descriptivo sobre la información preparada, es posible extraer las primeras hipótesis sobre los factores que pueden estar influenciando en la deserción de los empleados de la farmacéutica. A su vez, estos procesos de análisis permitirán elegir variables fundamentales para el entrenamiento de modelos en etapas posteriores y descartar aquellos componentes que puedan generar redundancia y sesgo en el entrenamiento.

9.1. Análisis de Variables Mediante Métodos Estadísticos Multivariados

Una manera de extraer las primeras relaciones e identificar patrones entre las variables disponibles en las fuentes de información es mediante el uso de métodos estadísticos multivariados

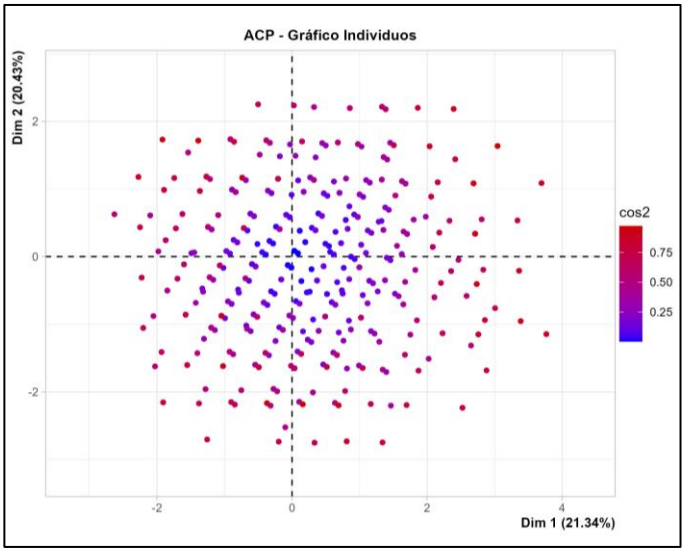
como lo son el Análisis de Componentes Principales (ACP) y el Análisis de Correspondencias Múltiples (ACM).

Mediante el ACP es posible analizar las variables cuantitativas a través de la reducción de su dimensionalidad. El primer análisis de variables cuantitativas está enfocado en los resultados de las encuestas de empleado y adicionalmente la información de las encuestas de los líderes. Esta segmentación se hizo con el fin de encontrar relaciones de satisfacción del empleado con su rendimiento, de esta manera empezar a detectar campos que causen redundancia y patrones.

Los resultados de este primer análisis se muestran a continuación.

Figura 16

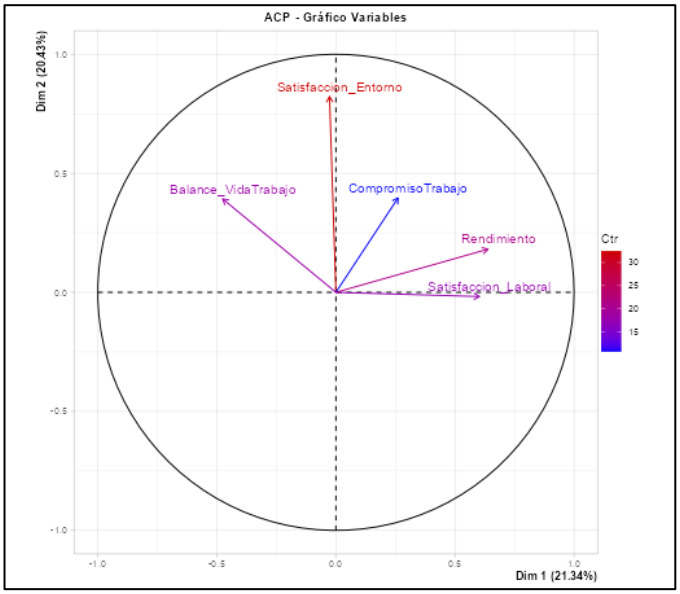
Gráfico individuos análisis de ACP encuestas



Fuente: elaboración propia

Figura 17

Gráfico de variables análisis de ACP encuestas



Fuente: elaboración propia

El primer plano factorial conformado por las dos componentes principales almacena una inercia de 41,76%. A partir de la nube de individuos, se puede destacar que en cuanto a resultados de encuestas no hay una tendencia, se distribuyen dentro de todo el plano.

El gráfico de variables indica que la Satisfacción con el cargo (Satisfaccion_Laboral) y el desempeño según el líder (Rendimiento) son los que más contribuyen en la primera componente, mientras que el resultado de Satisfacción con ambiente laboral (Satisfaccion_Entorno) ejerce la mayor contribución sobre la segunda componente.

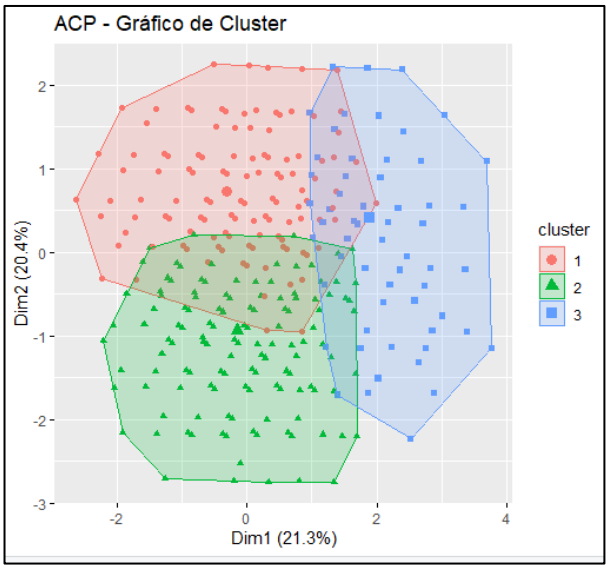
Mediante los ángulos entre las variables, se evidencia que la satisfacción con el ambiente laboral y la satisfacción con el cargo son componentes totalmente independientes, sin embargo, estas variables guardan cierta relación con el compromiso del empleado. Además, la satisfacción con el cargo tiene fuerte relación con el rendimiento del empleado según su líder. Finalmente es

posible ver que el balance de vida laboral y personal no tiene relación con el compromiso con el cargo, pero si esta positivamente relacionada con el ambiente de trabajo.

Aplicando método de agrupación de datos, se obtienen los siguientes resultados:

Figura 18

Gráfico de clúster sobre análisis de ACP encuestas



Fuente: elaboración propia

Tabla 14

Características de cada clúster de ACP encuestas

Campo	Cluster 1	Cluster 2	Cluster 3
Balance_VidaTrabajo	20.4	-13.6	-11.2
CompromisoTrabajo	11.4	-15.4	6.08
Rendimiento	-11	-17.8	45.7
Satisfaccion_Entorno	42.3	-45.5	4.19
Satisfaccion_Laboral	-8.15	-0.965	14.6

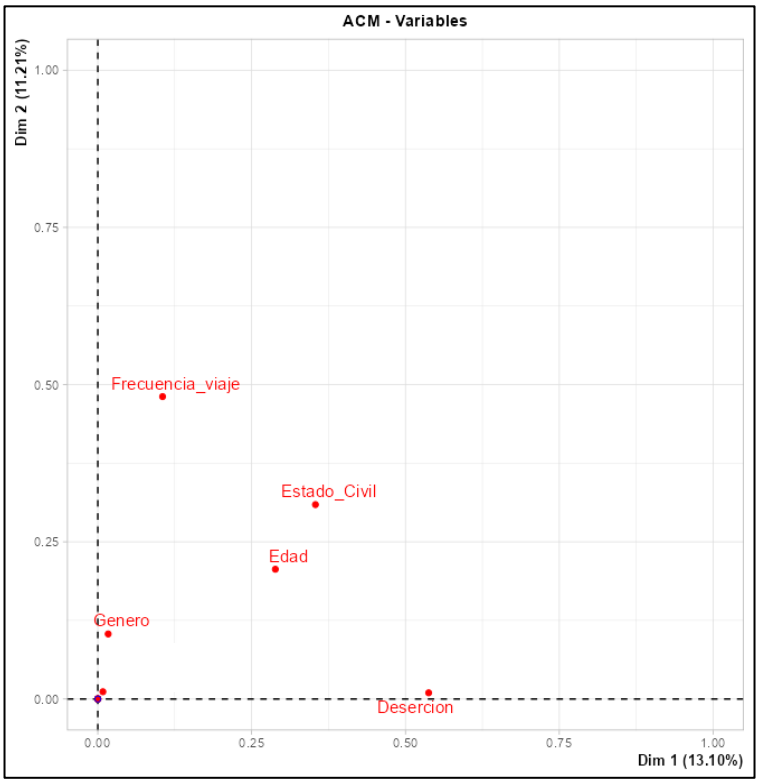
Fuente: elaboración propia

En esta agrupación de individuos se puede identificar que hay más empleados cuya satisfacción con el ambiente laboral es menor a la media, igual que su desempeño (Clúster 2). Adicionalmente, se puede establecer que el grupo más pequeño de empleados tiene una evaluación de desempeño por encima de la media, al igual que su satisfacción con el cargo (Clúster 3).

Por otra parte, el ACM es un método que permite analizar variables cualitativas para identificar perfiles de individuos. Se incluyeron variables que permitan agrupar a los empleados entre desertores y no desertores por rango de edad, frecuencia de viaje, área, género y estado civil.

Figura 19

Gráfico variables ACM



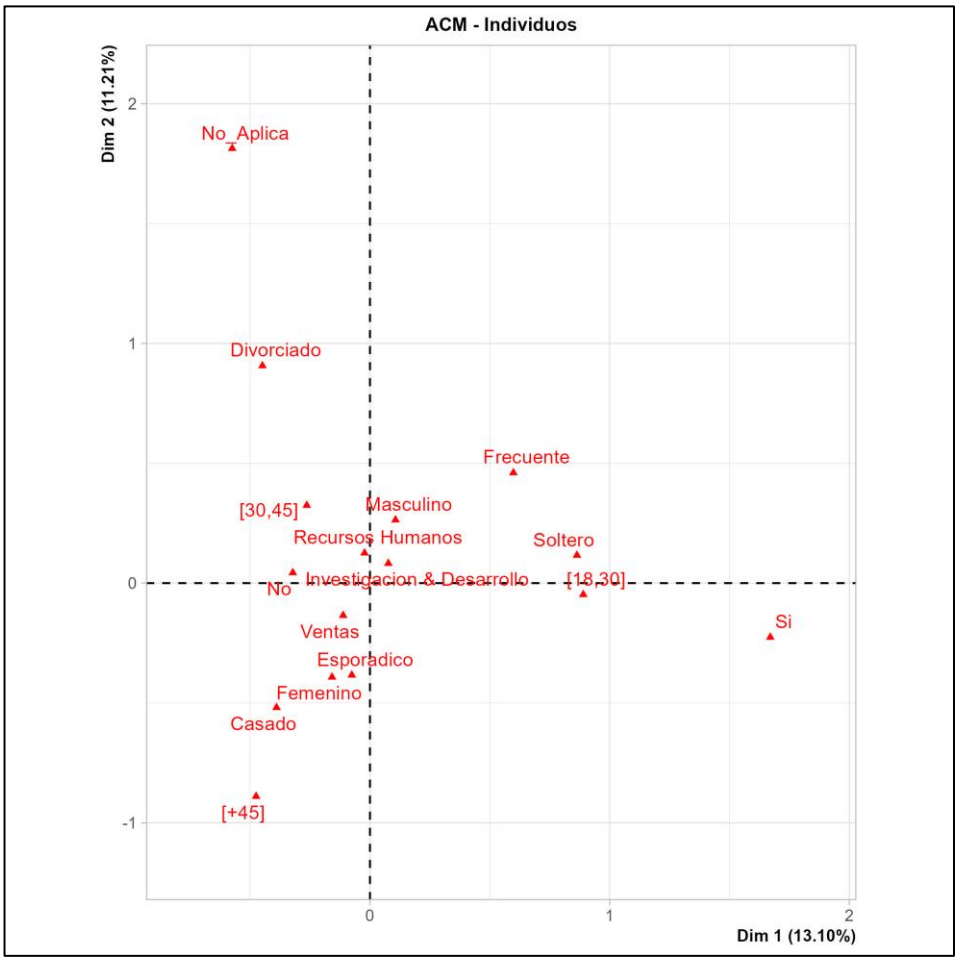
Fuente: elaboración propia

El primer plano factorial conserva el 24,31% de la información o inercia. Se puede observar cómo la variable de deserción es la que más contribuye al primer componente al mostrar una mayor cercanía a este eje en el gráfico de variables. Las demás variables contribuyen en mayor proporción

al segundo componente; especialmente género, área y frecuencia de viaje, éstas se pueden resumir como características demográficas.

Figura 20

Gráfico individuos ACM



Fuente: elaboración propia

Ahora bien, el ACM indica que aquellos valores más cercanos al centro son los más frecuentes. Para este caso, analizando el plano de individuos, es más frecuente pertenecer al área de Investigación y desarrollo o Ventas, ser de género masculino y además estar casado. De igual manera, es más frecuente no ser desertor de la compañía. En la edad se puede evidenciar un efecto Guttman en el plano factorial que va desde el rango de 18 a 30, 31 a 45 y 45+ años. Esto puede

deberse a un efecto de ordenamiento en las edades. Sin embargo, el rango de edad más frecuente en los empleados analizados es 30 a 45 años.

Por otro lado, al tratar de reconocer las características demográficas para los desertores se puede observar que es más frecuente que sean solteros, viajen frecuentemente y que estén en un rango de edad entre 18 a 30 años. En cuanto al área y el género no se identifica una relación clara ya que hay gran cantidad de no desertores y esto no permite realizar una caracterización exacta para estas variables.

9.2.Tablero de Análisis de Indicadores y Estado Actual de la Farmacéutica

Una vez se han preparado y entendido los datos mediante técnicas como ACP y ACM se procede a realizar un análisis descriptivo inicial sobre la situación actual de deserción en la farmacéutica. Para realizar el análisis se utilizó Power BI como herramienta para la creación de un tablero que permita a los ejecutivos de la empresa y otros actores importantes entender el comportamiento de los retiros en la empresa y sus características principales.

Figura 21

Tablero Deserción Farmacéutica – Reporte

Fuente: elaboración propia

La primera parte del tablero cuenta con 3 secciones que incluyen información de los empleados de la farmacéutica relacionada con sus retiros, desempeño y datos demográficos. La primera sección está compuesta por datos demográficos. Durante el año analizado se retiraron 711 empleados de la compañía, lo que representó una tasa de deserción del 16%. De este total, el 62% son hombres y el 38% mujeres. Por otro lado, la edad promedio de los desertores es de 34 años y cuentan con una antigüedad promedio de 5 años en la empresa.

La siguiente sección permite identificar el porcentaje de deserción en las áreas analizadas de la compañía, así como también los resultados más relevantes de las encuestas realizadas a los empleados. El área de Investigación y Desarrollo es la que cuenta con mayor número de desertores con 414 empleados, lo que corresponde a más del 50% del total.

Es importante resaltar que un análisis de los resultados de las encuestas a los empleados muestra que la satisfacción con el ambiente laboral y satisfacción con el nivel de cargo no difiere

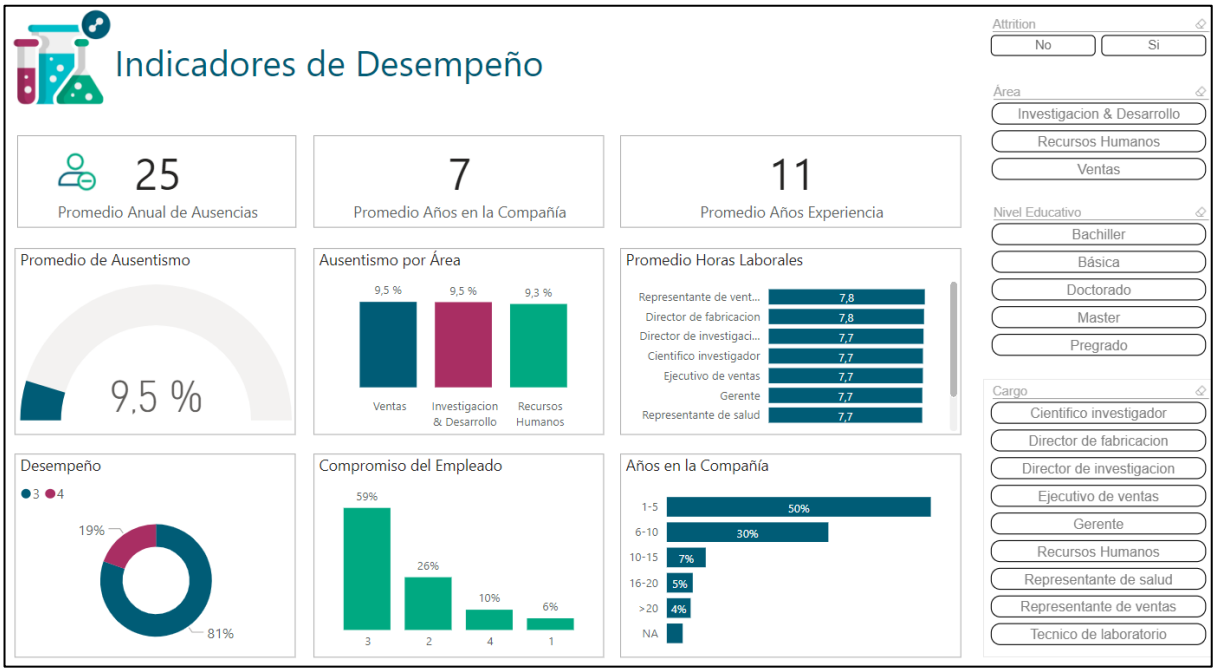
entre los empleados que han abandonado la compañía y los empleados activos, ya que estas métricas obtuvieron un valor promedio de 2,77 y 2,47 respectivamente.

Para concluir el análisis demográfico y laboral de los empleados se identifican los roles con mayor frecuencia de deserción. Estos fueron: Ejecutivo de ventas (165), Científico investigador (159) y Técnico de laboratorio (126). Los anteriores con un número de desertores mayor a 100. Por último, el nivel de educación más frecuente para los empleados que desertan es Pregrado y Maestría.

Este análisis descriptivo permite empezar a entender e identificar características representativas de los empleados que tienen propensión a retirarse de la compañía. Se espera que este tablero ayude a los actores correspondientes a entender la situación y desarrollar planes de acción para evitar efectos negativos en la farmacéutica.

Figura 22

Tablero Deserción Farmacéutica – Desempeño



Fuente: elaboración propia

La segunda pestaña del tablero contiene información del desempeño de los empleados, así como datos de ausentismo y variables asociadas a su rendimiento. Con la información disponible en el tablero se evidencia un promedio de ausencias de 9.5% que representan 25 días al año, siendo recursos humanos el área con el menor porcentaje (9.3%).

El desempeño de todos los empleados está entre nivel 3 y 4, lo cual es un indicio de una cultura de alto desempeño a lo largo de la organización. Para las personas que salieron de la compañía el porcentaje de desempeño es más alto comparado con las que continúan en ella, con una diferencia de 3 puntos porcentuales.

En términos de experiencia, los empleados tienen en promedio 11 años de experiencia laboral, siendo más baja la media para quienes salieron de la empresa (8 años). El 50% de las personas llevan entre 1 y 5 años trabajando en la empresa, sin embargo, este porcentaje es más alto para quienes no trabajan más en la empresa (62%).

En términos generales, los datos presentados en esta sección permitirán a todos los niveles de la compañía identificar comportamientos de falta de motivación, así como generar alertas de pérdida de compromiso de los empleados. Adicionalmente, proporciona información para el seguimiento de desempeño individual y con ello optimización de gestión del tiempo, que incentive una cultura de alto desempeño para desarrollar planes de acción específicos y personalizados en los equipos. Las cifras de desempeño contribuyen a su vez a analizar posibles riesgos de renuncia por bajas en los resultados.

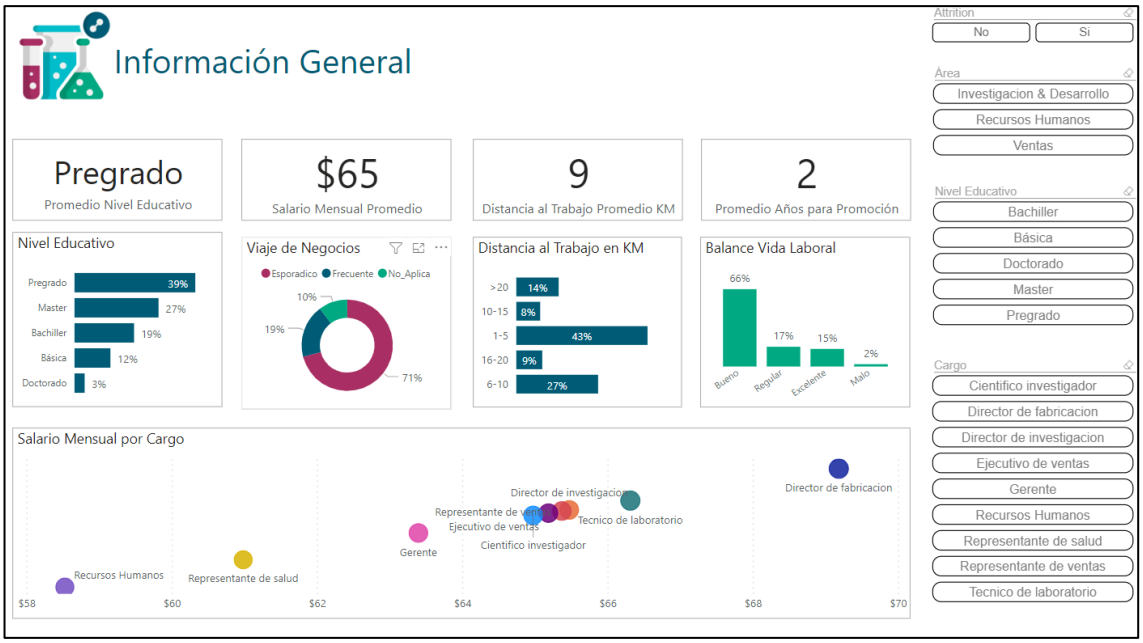
En línea con lo anterior, la sección de desempeño expone datos de horas laboradas por los empleados, lo cual podría indicar posibles excesos de horas adicionales que a futuro podrían impedir un equilibrio entre la vida laboral y personal de los colaboradores. De igual manera, expone datos de experiencia, permanencia en la compañía, compromiso y desempeño que

proporcionan herramientas para tomar decisiones informadas como posibles promociones de los empleados.

La última pestaña del tablero contiene información de nivel educativo, distancia al trabajo, balance entre la vida laboral y personal, así como salarios promedio por rol. Ésta permite a los usuarios analizar información personal de los empleados, y da señales sobre el contexto de bienestar de estos.

Figura 23

Tablero Deserción Farmacéutica – Información General



Fuente: elaboración propia

Los datos muestran que en promedio los empleados tienen nivel de estudios de pregrado, viven a 9 kilómetros de distancia del trabajo y su salario mensual es de $65.000 rupias. En este

sentido, el 66% califica el balance de su vida laboral como “Bueno”, por lo cual se pueden intuir que en general las condiciones laborales son favorables. Una variable que puede influir en lo anterior es la distancia al trabajo, ya que el 70% de los trabajadores vive a menos de 10 kilómetros de distancia del trabajo. En cuanto al salario, los directores de fabricación son los que más devengan en la compañía, seguido de los técnicos de laboratorio y directores de investigación.

Adicionalmente, es importante resaltar que el salario promedio de las personas que se fueron de la compañía es en promedio más bajo ($62.000 rupias). El promedio de número de años desde la última promoción son 2 y el 71% de los empleados viaja por negocios de manera esporádica.

Para concluir, la pestaña de información general permite a todos los niveles de la organización tomar decisiones de potencial crecimiento de los empleados y oportunidades de desarrollo que mitiguen posibles deserciones a futuro. De igual forma, permite analizar el bienestar actual de los empleados e identificar áreas de mejora y oportunidades para incrementar la satisfacción laboral que en el largo plazo contribuye a una mayor retención de talento.

En conjunto, todas las secciones del tablero otorgan una visión objetiva de los equipos y los empleados de la farmacéutica que contribuyen a una correcta gestión de rotación del personal e identificación de áreas de mejora mediante los hallazgos anteriormente expuestos.

10. Modelo de Probabilidad de Deserción para Empleados Activos

Una vez se ha realizado el análisis descriptivo y multivariado de los datos disponibles se procede a realizar la fase de modelamiento del proyecto. En esa sección se describe el proceso de entrenamiento, evaluación y selección del mejor modelo para medir la probabilidad de retiro de empleados activos y finalmente se expone el proceso de creación del tablero de alerta que integra el modelo seleccionado.

10.1. Entrenamiento

Con el propósito de llevar a cabo el entrenamiento de modelos para obtener las predicciones deseadas, se toman los datos integrados de las fuentes mencionadas previamente (General Empleados, Encuestas de empleados, Encuestas de líderes y Entradas y salida de empleados). Lo anterior permite consolidar toda la información en una única base, facilitando el manejo de los datos. Adicionalmente, a este consolidado se le eliminaron las columnas que no aportaban al entrenamiento del modelo como el ID del empleado, así como aquellas que fueran redundantes al estar altamente correlacionadas (**Figura 6**), como antigüedad y número de años con el líder actual. Finalmente, se transformaron las variables categóricas en variables binarias para tener únicamente entradas numéricas a los modelos que se entrenarán.

Una vez realizados los procesos de entendimiento y preparación de los datos se cuenta con las herramientas necesarias para entrenarlos y así obtener predicciones. De acuerdo con los resultados descriptivos previos, inicialmente se observó un desbalanceo en la distribución de los valores del campo “deserción”. La base de datos muestra un 83% de empleados clasificados en la categoría “no desertores”, lo cual puede afectar la capacidad del modelo de predecir de manera correcta ya que ésta es la variable objetivo. Tener una proporción significativamente mayor de “no desertores” lleva a que los resultados puedan sesgarse hacia la clase mayoritaria, disminuyendo así su nivel de precisión.

Con el propósito de corroborar las consecuencias del desbalanceo de los datos se ejecutó un modelo de regresión logística cuyo nivel de predicción para los “no desertores” fue de 98%, en comparación con un 18% para la categoría de “desertores”. Esto confirma la necesidad de utilizar estrategias que mitiguen los efectos del desbalanceo en el modelo predictivo, antes de proceder con el entrenamiento a distintos modelos.

Para balancear la muestra se hizo uso de dos estrategias: SMOTE y ADASYN. Éstas comparten el objetivo común de reducir el desbalanceo de la muestra a través de la generación de registros sintéticos de la clase minoritaria. Sin embargo, tienen un enfoque distinto en la manera en la que adaptan el proceso a la muestra.

En el caso de la estrategia SMOTE, la creación de registros sintéticos se realiza interpolando características entre registros minoritarios, en este caso de los "desertores". La estrategia ADASYN por su parte, se ajusta a la distribución de densidad de los datos para generar los registros sintéticos de la categoría minoritaria. Realizando estos métodos se obtuvo un cambio en el número de registros por categoría de la siguiente manera:

Tabla 15

Registros con métodos de balanceo

Categoría	**# Registros Modelo Original**	**# Registros SMOTE**	**# Registros ADASYN**
0 – No desertores	3.605	3.605	3.605
1 - Desertores	695	3.605	3.491

Fuente: elaboración propia

Una vez que la fuente ha sido balanceada, se procede a la fase de entrenamiento, para lo cual se hará uso de modelos de machine learning que cumplan con el objetivo de predecir la deserción de empleados.

Para realizar la selección del mejor modelo se ejecutó inicialmente un modelo de regresión logística, el cual tiene ventajas en términos de interpretación y es menos sensible a problemas de sobreajuste que otros modelos similares. Adicionalmente, se realizaron tres modelos supervisados basados en árboles de decisión (Random Forest, Gradient Boosting y Árbol de Decisión) teniendo

en cuenta que no asumen relaciones lineales entre las entradas y salidas, lo cual permite modelar patrones de datos de mayor complejidad y tener una mayor precisión en las predicciones.

Los modelos anteriores se entrenan después de cada una de las estrategias de balanceo SMOTE y ADASYN y posteriormente para cada uno de los modelos se calcula la importancia relativa de todas las variables para así identificar aquellas que contribuyen de manera significativa a la predicción y seleccionar solo las características relevantes para simplificar el modelo. Para este último paso, se seleccionaron las 10 variables con mayor contribución a cada modelo, para luego entrenar nuevamente y comparar los resultados.

Posteriormente, teniendo en cuenta los resultados de las ocho alternativas mencionadas se realizó evaluación de rendimiento haciendo uso de métricas importantes como: sensibilidad por clase, Error Tipo I y II, AUC, sobreajuste y precisión para así identificar el mejor modelo.

Se obtuvieron los siguientes resultados para modelos según estrategia SMOTE con todas sus variables:

Tabla 16

Evaluación de modelos – SMOTE (empleados activos)

SMOTE	**Sensibilidad clase 0**	**Sensibilidad clase 1**	**Error tipo 1**	**Error tipo 2**	**AUC Train**	**AUC Test**	**Sobreajuste: ABS (AUC train – AUC test)**	**Precisión**
Regresión Logística	0,89	0,77	100	203	0,8323	0,8317	0,060%	0,8319
Árbol de Decisión	0,85	0,75	133	221	0,8036	0,8035	0,010%	0,8036

Random Forest	0,91	0,83	77	151	0,8752	0,8734	0,180%	0,8735
Gradient Boosting	0,98	0,98	15	17	0,9972	0,9822	1,500%	0,9822

Fuente: elaboración propia

Posteriormente se ejecutaron los cuatro modelos excluyendo aquellas variables con menor importancia y se obtuvieron los siguientes resultados.

Tabla 17

Evaluación de modelos – SMOTE variables significativas (empleados activos)

SMOTE – Variables Significativas	**Sensibilidad clase 0**	**Sensibilidad clase 1**	**Error tipo 1**	**Error tipo 2**	**AUC Train**	**AUC Test**	**Sobreajuste: ABS (AUC train – AUC test)**	**Precisión**
Regresión Logística	0,89	0,77	96	211	0,826	0,8295	0,350%	0,8297
Árbol de Decisión	0,85	0,75	134	221	0,8032	0,8029	0,030%	0,8031
Random Forest	0,91	0,81	78	173	0,8581	0,8606	0,250%	0,8607
Gradient Boosting	0,98	0,98	20	22	0,995	0,9767	1,830%	0,9767

Fuente: elaboración propia

Una vez finalizada la evaluación de modelos por estrategia SMOTE se realizó el mismo ejercicio aplicando la estrategia ADASYN y se obtuvieron los siguientes resultados en la primera fase con todas sus variables:

Tabla 18

Evaluación de modelos – ADASYN (empleados activos)

ADASYN	Sensibilidad clase 0	Sensibilidad clase 1	Error tipo 1	Error tipo 2	AUC Train	AUC Test	Sobreajuste: ABS (AUC train - AUC test)	Precisión
Regresion Logística	0,9	0,73	85	240	0,8267	0,817	0,970%	0,8167
Árbol de Decisión	0,88	0,71	106	255	0,8051	0,7967	0,840%	0,7965
Random Forest	0,92	0,84	70	140	0,8884	0,8817	0,670%	0,8816
Gradient Boosting	0,99	0,98	10	18	0,9988	0,9842	1,460%	0,9842

Fuente: elaboración propia

Luego de identificar las variables con mayor importancia relativa se ejecutan nuevamente los cuatro modelos bajo estrategia ADASYN y se obtienen los siguientes resultados:

Tabla 19

Evaluación de modelos – ADASYN variables significativas (empleados activos)

ADASYN – Variables Significativas	**Sensibilidad clase 0**	**Sensibilidad clase 1**	**Error tipo 1**	**Error tipo 2**	**AUC Train**	**AUC Test**	**Sobreajuste: ABS (AUC train - AUC test)**	**Precisión**
Regresión Logística	0,89	0,73	93	237	0,8253	0,8142	1,110%	0,8139
Árbol de Decisión	0,88	0,72	106	253	0,8041	0,7979	0,620%	0,7976
Random Forest	0,92	0,84	71	143	0,8795	0,8793	0,020%	0,8793
Gradient Boosting	0,99	0,97	13	31	0,9963	0,9752	2,110%	0,9751

Fuente: elaboración propia

10.2. Evaluación

Con los resultados anteriores, se evalúa la calidad y rendimiento de los modelos construidos para elegir el más destacado.

Teniendo en cuenta el objetivo de negocio, el primer criterio para seleccionar el mejor modelo fue la implicación que pueden tener los Errores Tipo I y II en la farmacéutica. Por un lado, el Error Tipo II representa el número de empleados que efectivamente desertaron, pero que el modelo predijo que no lo harían, mientras que el Error Tipo I se refiere a las personas que no desertaron y el modelo predijo que sí se retirarían. La implicación de estos errores está relacionada a la toma de decisiones y acciones anticipadas frente a empleados que tienen indicios de abandonar la compañía, por lo que resulta más costoso para la empresa no identificar personas con alta probabilidad de desertar (Error Tipo II), que identificar de manera errónea personas que realmente no tenían intención de irse de la compañía (Error Tipo I). Por lo anterior el primer criterio para seleccionar el modelo fue tomar aquellos con menor número de Errores Tipo II, en este caso el Random Forest y el Gradient Boosting descartando así la regresión logística y el árbol de decisión.

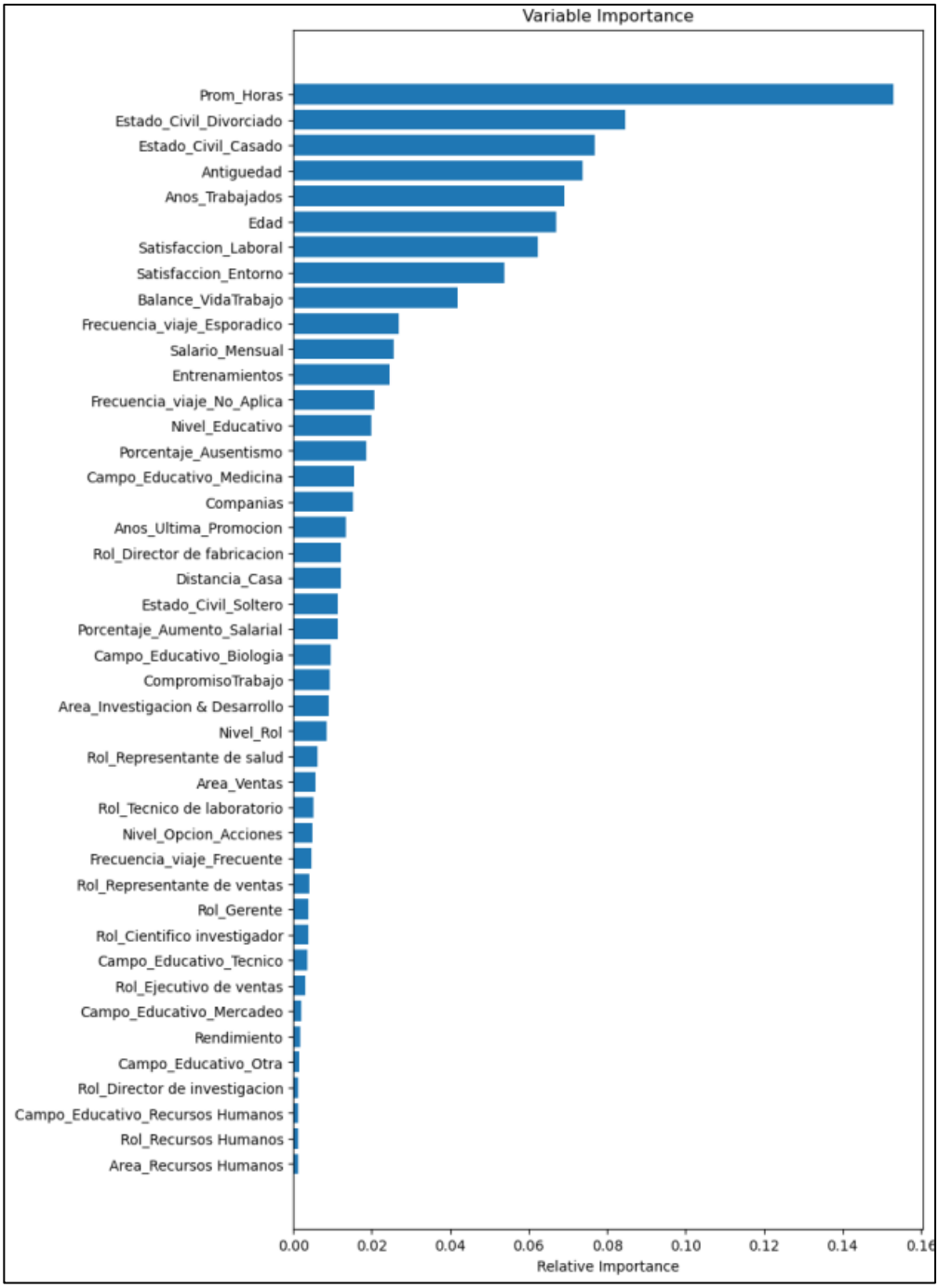
El segundo criterio para elegir el mejor modelo es el sobreajuste. Para este caso se busca encontrar el modelo que tenga los menores niveles de sobreajuste de manera que su desempeño sea exitoso con datos distintos a los usados para su entrenamiento, es decir, se busca seleccionar el que tenga mayor capacidad para generalizar los conceptos. Para ello, se calculó la diferencia entre el AUC del set de entrenamiento y el de prueba, encontrando un menor nivel de sobreajuste para el Random Forest. Para estos dos modelos las diferencias entre criterios no tienen gran magnitud, sin embargo se tuvo en cuenta también seleccionar aquel con menor complejidad en términos de procesamiento de datos masivos.

Para elegir la metodología de balanceo de datos, SMOTE (**Tabla 16** y **Tabla 17**) o ADASYN (**Tabla 18** y **Tabla 19**), se consideraron los resultados del Random Forest en cada caso, encontrando mejores resultados del modelo bajo el método ADASYN. Éste, además de tener menores Errores Tipo I y II en comparación con el SMOTE, presenta un mayor nivel de precisión ya que se ajusta a la distribución de densidad del conjunto de datos.

Finalmente, con el fin de entender los factores que tienen un mayor impacto sobre la predicción del modelo, se verificó la importancia relativa de cada una de las variables, obteniendo los siguientes resultados:

Figura 24

Importancia relativa de variables – Modelo empleados activos



Fuente: elaboración propia

De acuerdo con la gráfica anterior, se evidencia que las variables relacionadas con los roles, áreas, educación, compromiso y rendimiento no son características que tengan un impacto significativo en la capacidad de predicción del modelo, por lo cual se eliminan y se entrena

nuevamente. Lo anterior contribuye a simplificar el modelo mejorando su interpretabilidad y disminuyendo los Errores Tipo I y II.

En conclusión, el modelo elegido para la predicción del modelo de probabilidad de renuncia para los empleados de la farmacéutica es Random Forest, con estrategia de balanceo ADASYN y únicamente incluyendo las variables significativas para el entrenamiento.

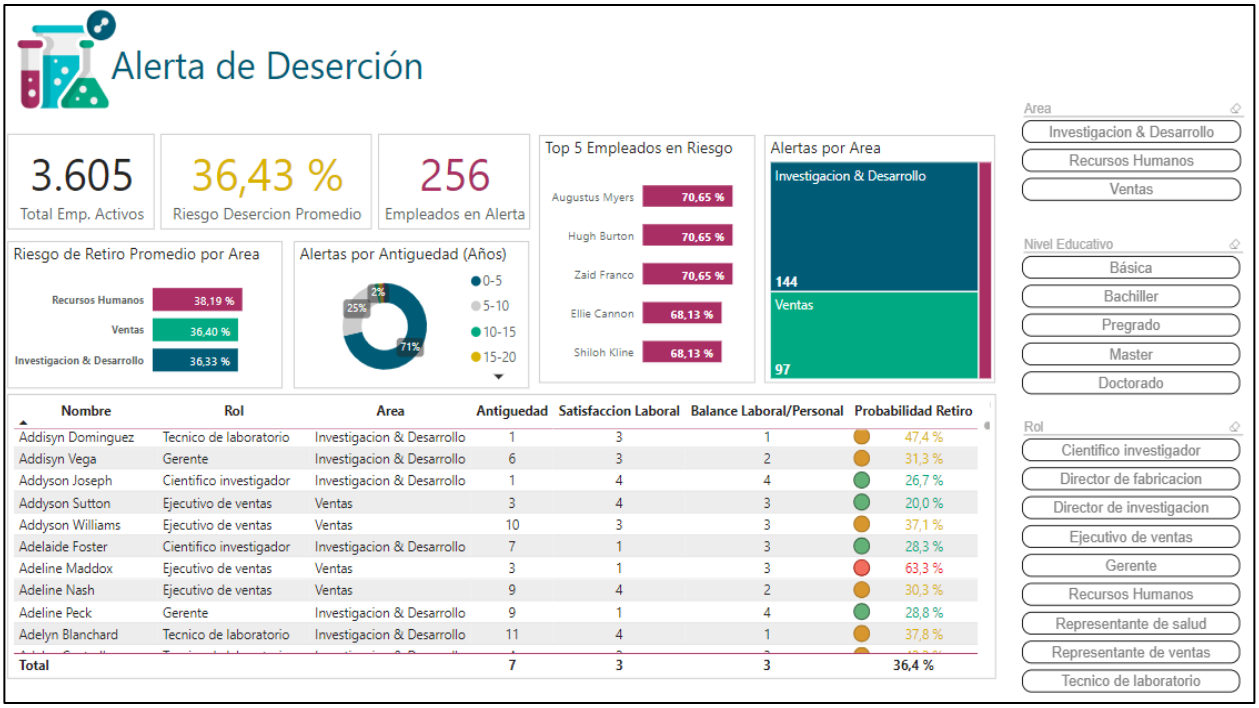
10.3. Generación de Tablero de alerta de deserción

Una vez elegido el modelo, se construyó una nueva sección dentro del tablero organizacional, cuyo objetivo principal es la detección temprana de empleados activos con riesgo de abandonar la empresa.

La sección está compuesta inicialmente por indicadores relacionados con la probabilidad de deserción tanto a nivel empresa, como de área y antigüedad, esto teniendo en cuenta la necesidad del negocio de retener talentos con experiencia y detectar en qué áreas es necesario enfocar acciones. Adicionalmente brinda un top de empleados con mayor probabilidad de deserción que permite tomar decisiones críticas a nivel empleado. Finalmente, esta sección muestra un detalle de todos los empleados, con características importantes que pueden influir en una decisión de abandonar la empresa y un indicador de probabilidad de deserción clasificado por colores para facilitar la detección de individuos con alerta o probabilidad de retiro mayor. Cabe resaltar que en esta sección se puede filtrar por Área, Nivel Educativo y Rol según la necesidad del usuario.

Figura 25

Tabla de deserción farmacéutica – Alerta de deserción



Nombre	Rol	Area	Antiguedad	Satisfaccion Laboral	Balance Laboral/Personal	Probabilidad Retiro
Addisyn Dominguez	Tecnico de laboratorio	Investigacion & Desarrollo	1	3	1	47,4 %
Addisyn Vega	Gerente	Investigacion & Desarrollo	6	3	2	31,3 %
Addyson Joseph	Cientifico investigador	Investigacion & Desarrollo	1	4	4	26,7 %
Addyson Sutton	Ejecutivo de ventas	Ventas	3	4	3	20,0 %
Addyson Williams	Ejecutivo de ventas	Ventas	10	3	3	37,1 %
Adelaide Foster	Cientifico investigador	Investigacion & Desarrollo	7	1	3	28,3 %
Adeline Maddox	Ejecutivo de ventas	Ventas	3	1	3	63,3 %
Adeline Nash	Ejecutivo de ventas	Ventas	9	4	2	30,3 %
Adeline Peck	Gerente	Investigacion & Desarrollo	9	1	4	28,8 %
Adelyn Blanchard	Tecnico de laboratorio	Investigacion & Desarrollo	11	4	1	37,8 %
Total			7	3	3	36,4 %

Fuente: elaboración propia

11. Modelo de Probabilidad de Deserción de Candidatos

Dando continuidad a la fase de modelamiento del proyecto y habiendo seleccionado el mejor modelo para medir la probabilidad de deserción de empleados activos se procede a realizar la selección de un nuevo modelo para la predicción de deserción de candidatos de la farmacéutica. Esta sección incluye de igual manera una descripción sobre el proceso de creación de la calculadora de deserción para candidatos.

11.1. Entrenamiento

Teniendo en cuenta que para el modelo de probabilidad de deserción de empleados activos se utilizaron todas las variables significativas, se hace necesaria la creación de una nueva base de datos que contenga únicamente los campos a diligenciar por candidatos que tengan la intención de ingresar a la compañía. Es decir, se eliminan todas las variables relacionadas a satisfacción laboral, compromiso, ausentismo y todas aquellas que tengan relación con el contexto laboral actual de los

empleados, ya que no pueden ser utilizadas como variables de entrada para la calculadora de candidatos.

Para realizar el entrenamiento, se siguió el mismo proceso anterior de balanceo con la nueva base de datos utilizando las estrategias SMOTE y ADASYN, para posteriormente entrenar los cuatro modelos supervisados (Regresión Logística, Árbol de Decisión, Random Forest y Gradient Boosting).

Realizando balanceo bajo el método SMOTE se obtuvieron los siguientes resultados:

Tabla 20

Evaluación de modelos – SMOTE (candidatos)

SMOTE	Sensibilidad clase 0	Sensibilidad clase 1	Error tipo 1	Error tipo 2	AUC Train	AUC Test	Sobreajuste: ABS (AUC train - AUC test)	Precisión
Regresión Logística	0,88	0,67	112	294	0,781307	0,774538	0,677%	0,77
Árbol de Decisión	0,89	0,66	101	310	0,762636	0,771723	0,909%	0,77
Random Forest	0,86	0,79	124	191	0,8173	0,825186	0,789%	0,83
Gradient Boosting	0,96	0,95	35	43	0,976333	0,956726	1,961%	0,96

Fuente: elaboración propia

Al identificar solo las 10 variables con mayor importancia se obtuvieron los siguientes resultados:

Tabla 21

Evaluación de modelos – SMOTE variables significativas (candidatos)

SMOTE Variables Significativas	Sensibilidad clase 0	Sensibilidad clase 1	Error tipo 1	Error tipo 2	AUC Train	AUC Test	Sobreajuste: ABS (AUC train - AUC test)	Precisión
Regresión Logistica	0,88	0,7	105	269	0,787589	0,792314	0,472%	0,79
Árbol de Decisión	0,89	0,66	101	310	0,762636	0,771723	0,909%	0,77
Random Forest	0,87	0,75	121	227	0,807154	0,806823	0,033%	0,81
Gradient Boosting	0,97	0,95	30	45	0,982988	0,958379	2,461%	0,96

Fuente: elaboración propia

Por otro lado, se obtuvieron los siguientes resultados realizando balanceo bajo la estrategia ADASYN:

Tabla 22

Evaluación de modelos – ADASYN (candidatos)

ADASYN	Sensibilidad clase 0	Sensibilidad clase 1	Error tipo 1	Error tipo 2	AUC Train	AUC Test	Sobreajuste: ABS (AUC train - AUC test)	Precisión
Regresión Logística	0,9	0,62	90	322	0,775472	0,762411	1,306%	0,77
Árbol de Decisión	0,86	0,68	130	272	0,767144	0,769483	0,234%	0,77
Random Forest	0,84	0,78	148	185	0,825519	0,810297	1,522%	0,81
Gradient Boosting	0,96	0,96	33	38	0,984327	0,959597	2,473%	0,96

Fuente: elaboración propia

Al tomar solo las variables significativas se obtuvieron los siguientes resultados:

Tabla 23

Evaluación de modelos – ADASYN variables significativas (candidatos)

ADASYN Variables Significativas	Sensibilidad clase 0	Sensibilidad clase 1	Error tipo 1	Error tipo 2	AUC Train	AUC Test	Sobreajuste: ABS (AUC train - AUC test)	Precisión
Regresión Logística	0,9	0,56	89	380	0,75785	0,729125	2,873%	0,73
Árbol de Decisión	0,7	0,73	276	230	0,73108	0,713324	1,776%	0,71
Random Forest	0,83	0,75	154	213	0,82551	0,790646	3,486%	0,79
Gradient Boosting	0,96	0,95	36	41	0,98224	0,956189	2,605%	0,96

Fuente: elaboración propia

11.2. Evaluación

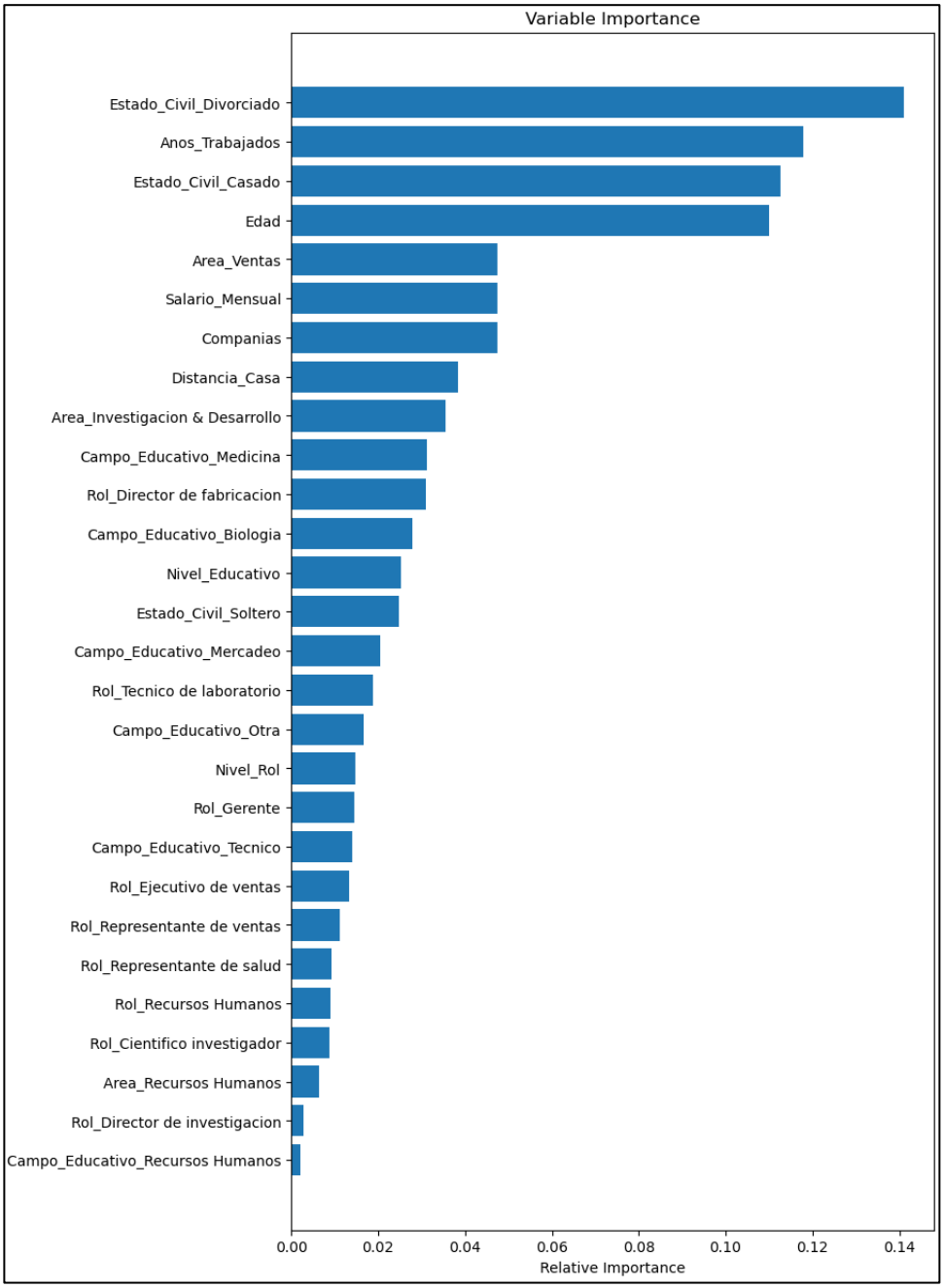
Los criterios de evaluación para elegir el modelo de probabilidad de deserción de candidatos fueron muy similares al modelo anterior. Sin embargo, los resultados difieren ya que hay pérdida de información ocasionada por la eliminación de variables, que cambia la generalización realizada por el modelo en términos de predicción.

El primer criterio de selección es nuevamente el modelo con menor número de Errores Tipo II, ya que estos implican costos elevados que la compañía asume al contratar y entrenar personas que deciden desertar en el corto o mediano plazo. Por lo anterior, nuevamente se descartan la regresión logística y el árbol de decisión, al ser los modelos con mayor número de Errores tipo II.

Posteriormente, se consideran tanto el nivel de sobreajuste como la precisión para elegir el mejor modelo, siendo el Random Forest el que mejor cumple las características mencionadas. Tanto para Random Forest como para Gradient Boosting las diferencias entre criterios no presentan una gran magnitud, sin embargo, se tuvo en cuenta seleccionar aquel modelo con menor complejidad en términos de procesamiento de datos masivos. Finalmente, se evalúa tanto la metodología de balanceo como la inclusión de variables, encontrando para este caso que, dada la distribución de los datos, el ADASYN es el que mejor realiza el proceso de balanceo. Nuevamente se verifica la importancia relativa de las variables, encontrando los siguientes resultados:

Figura 26

Importancia relativa de variables – Modelo candidatos



Fuente: elaboración propia

Después de entrenar el modelo, se evidencia que para este caso incluir únicamente las variables significativas reduce precisión e incrementa el nivel de sobreajuste (**Tabla 23**). Por lo

cual, incluir todas las variables garantiza un modelo más robusto con mayor capacidad de generalizar de manera correcta los datos.

Para concluir, el modelo que mejor se ajusta al cálculo de probabilidad de deserción de los candidatos es el Random Forest bajo la metodología de balanceo ADASYN, incluyendo todas las variables de la base de datos.

11.3. Construcción de Calculadora de Deserción

Figura 27

Calculadora de Probabilidad de Deserción para Candidatos

Para realizar la interfaz gráfica de la calculadora se optó por utilizar el lenguaje Python integrado con la librería tkinter. Inicialmente se definieron las variables que ingresarán mediante la calculadora y con las cuales se realizará la evaluación sobre el modelo. Las variables de entrada fueron definidas de la siguiente manera:

- N° Candidato: Campo para ingresar el número del candidato a evaluar.
- Área: Lista desplegable para elegir entre área de Investigación y Desarrollo, Ventas o Recursos Humanos.
- Distancia –Km: Este campo es de valor numérico y se debe diligenciar la distancia que existe entre la compañía y el lugar de residencia del candidato.
- Nivel Rol: Campo numérico con lista desplegable del 1 al 5, según el nivel de la posición activa.
- Rol: Lista desplegable que contiene las opciones de Científico Investigador, Representante de la Salud, Ejecutivo de Ventas, Recursos Humanos, Técnico de Laboratorio, Director de Investigación, Representante de ventas y Gerente.
- Campo Educativo: Lista desplegable con las opciones de Medicina, Biología, Mercadeo, Recursos Humanos, Técnico y Otro.
- Nivel Educativo: Lista desplegable del 1 al 5.
- Experiencia - Años: Campo numérico libre para diligenciar.
- Compañías: Campo con entrada libre con el número de las compañías en las cuales el candidato ha laborado.
- Salario Mensual: Campo de tipo numérico para libre diligenciamiento.
- Edad: Entrada de tipo numérico.
- Estado Civil: Campo de lista desplegable con las opciones: Casado, Soltero, Divorciado.

Estas variables de entrada son capturadas al activar el botón “Evaluar” o “Cargar”. Al momento del cargue ingresan a un arreglo definido como entrenamiento, posteriormente estos valores del arreglo pasan a un dataframe que es el que finalmente ingresará al modelo para su

posterior evaluación. Después de realizar esta evaluación la calculadora muestra el porcentaje de probabilidad de deserción del candidato a evaluar según el resultado que arroje el modelo entrenado previamente.

La calculadora consta de dos tipos de resultados: El primero es individual que contempla en el escenario cuando los valores del candidato son ingresados 1 a 1 en cada uno de los campos del formulario de la calculadora, en este caso cuando el modelo retorne el resultado se verá reflejado en la parte superior derecha del formulario de la calculadora representado como un porcentaje. El segundo escenario es grupal, la calculadora cuenta con una opción de cargue de archivo de varios registros de candidatos, esta opción es realizada mediante el botón “Cargar”. En este caso los resultados se reflejarán en la parte inferior de la calculadora en una tabla la cual detalla la probabilidad de deserción de cada uno de los candidatos evaluados.

La interfaz gráfica tiene implementado dentro de los resultados un estilo de semaforización para mayor entendimiento e interpretación de los resultados finales. De 0% a 30 % la probabilidad de deserción es considerada como baja y es representada en color verde, de 31% a 55% la probabilidad es media y es representada por el color amarillo y mayor a 55% la probabilidad es alta y es representada por el color rojo.

12. Sugerencias de implementación

Luego de seleccionar los modelos más adecuados para el sistema de deserción y haber creado las herramientas que lo componen, se realizan las siguientes recomendaciones para una implementación efectiva.

Iniciar con la creación de un equipo integral que se encargue de la implementación, comunicación, seguimiento y alcance del sistema. Se recomienda que esté constituido por especialistas en datos, ejecutivos e integrantes de distintas áreas como finanzas, recursos humanos,

mercadeo y demás áreas que se encarguen de realizar seguimiento de nuevos proyectos en la farmacéutica. Este equipo se reunirá con los especialistas en datos para conocer a fondo el objetivo y el alcance del sistema para posteriormente crear estrategias de retención y contratación efectiva que ayuden a alcanzar el nivel de deserción esperado en la farmacéutica.

Una vez creado el equipo, se puede iniciar proceso de creación y aprobación de estrategias. En cuanto a las iniciativas de retención el equipo debe establecer umbrales sobre la probabilidad de deserción y crear medidas de acción de acuerdo con el nivel de riesgo.

Según la probabilidad de deserción se crearían tres categorías:

1. Empleados sin riesgo de deserción: con probabilidad de retiro entre 0% y 30%
2. Empleados en riesgo: con probabilidad entre 31% y 55%
3. Empleados de alto riesgo: con probabilidad mayor a 55%

Luego de establecer los umbrales se debe hacer un reconocimiento extensivo de las variables o características de mayor importancia para aquellos empleados en riesgo. Dentro de las variables más importantes para determinar deserción identificadas por el modelo están: promedio de horas trabajadas, antigüedad, satisfacción laboral, balance vida trabajo y salario mensual. Con estas variables el equipo podría crear planes de acción para revisar el balance de carga laboral hablando con los jefes directos de los empleados. Proponer un estudio sobre el nivel salarial de la farmacéutica para asegurar que la compensación de los empleados sea competitiva en el mercado e internamente equitativa. En cuanto a antigüedad se podrían desarrollar estrategias que mantengan la fidelidad de los empleados con la farmacéutica. La creación de un equipo multidisciplinar que pueda mantener comunicación entre las áreas implicadas para su realización y con integrantes que tengan poder de decisión se convierte en un factor clave para el éxito del proyecto.

Otra de las tareas del equipo consiste en realizar un plan de comunicación y adopción de las herramientas para los usuarios principales como lo son el equipo de reclutamiento, recursos humanos y los ejecutivos a cargo de las contrataciones. Este plan debe incluir una sesión donde se exponga el sistema de predicción de deserción junto con las estrategias una vez hayan sido aprobadas. De manera adicional para brindar soporte se pueden crear instructivos de uso, sesiones de capacitación para el uso adecuado de las herramientas, sesiones de seguimiento, crear un canal efectivo de comunicación con preguntas y respuestas frecuentes y finalmente garantizar que los usuarios del sistema puedan dar sugerencias de mejora.

Por último, vale la pena resaltar que hay responsabilidades asociadas al mantenimiento del modelo que sirven como fuente principal del sistema que deben ser llevadas a cabo por el equipo de datos. Para realizar este mantenimiento se recomienda actualizar la información base del modelo y calibrarlo una vez por trimestre para identificar con tiempo posibles cambios o nuevos hallazgos relevantes.

De igual manera, se recomienda realizar un seguimiento trimestral a la frecuencia de uso de las herramientas y a los resultados de los planes de acción creados para retención y contratación efectiva.

Finalmente, cada semestre se debe evaluar la evolución de la tasa de deserción en la farmacéutica y cómo la herramienta ha impactado estos resultados. Todo lo anterior con fin de incentivar el uso del sistema para conseguir el objetivo organizacional y poder controlar las consecuencias negativas que conlleva una alta tasa de retiro.

13. Conclusiones

Los datos disponibles de manera cruda en la farmacéutica no brindan conocimiento per se para la empresa. Partiendo únicamente de estadística descriptiva y organización de la información

se pueden obtener hallazgos básicos y empezar a generar valor a través del entendimiento del estado actual de la compañía. Al incluir técnicas más avanzadas de análisis se pueden extraer patrones y asociaciones entre las características tanto demográficas como de desempeño de los empleados, algo que puede ser clave a la hora de crear planes de acción enfocados en mejorar indicadores.

Así como la disponibilidad de información combinada con la minería de datos es un punto de partida para extraer valor de las fuentes, el almacenamiento de los datos y el modelo operativo juegan un rol clave en la gestión de este recurso. En este proyecto se establece un modelo centralizado Data Warehouse, capaz de facilitar la estandarización de los metadatos de la información guardada y además permitir la extracción de datos de manera eficiente y organizada, destacando la facilidad de este modelo operativo a nivel de administración. Este modelo es un buen punto de partida para una empresa que está iniciando con la gestión de su información, sin embargo, si la cantidad de fuentes y volúmenes de datos crece rápidamente, así como la cantidad de personas que requieren aprovechar la información, se recomienda a la farmacéutica revisar de manera recurrente el diseño del modelo e incluso evaluar la alternativa de descentralizar de manera progresiva para facilitar el acceso.

La construcción del tablero de análisis actual de la farmacéutica permite a la empresa condensar la gran cantidad de datos disponibles en gráficos claros y concisos sobre preguntas principales del negocio. Las visualizaciones del tablero se convierten en una herramienta fundamental para entender el panorama de los empleados y detectar características iniciales del comportamiento de deserción en la compañía.

El análisis descriptivo se complementó a través de técnicas estadísticas como ACP y ACM. Estos aportaron al entendimiento de la situación hallando relaciones entre variables significativas

y reconociendo perfiles de empleados desertores y no desertores. En cuanto a relaciones entre variables se comprobó que la satisfacción con el ambiente laboral y el cargo son componentes independientes. Sin embargo, el ambiente en el trabajo parece influir en el compromiso de los trabajadores.

Las altas tasas de deserción tienen implicaciones negativas para el rendimiento a nivel financiero, comercial, investigativo y laboral en las empresas. Para este caso, una tasa de deserción del 16% representa altos riesgos de pérdida de talento clave para la farmacéutica y aumento de costos de contratación, por esta razón el proyecto adoptó técnicas de analítica guiadas bajo la metodología CRISP-DM para generar productos analíticos capaces de informar sobre posibles riesgos de abandono de los empleados y facilitar la toma de decisiones enfocadas en reducir las tasas de deserción y por consiguiente sus efectos negativos.

De esta manera, se generaron dos modelos de machine learning orientados a la detección del riesgo de deserción en empleados activos, así como de nuevos candidatos a ingresar a la farmacéutica, respectivamente. El primer modelo, enriquece el tablero del sistema ya que el resultado de probabilidad de abandono por cada empleado activo es la fuente principal de la sección de alertas por riesgo de deserción. Así, el tablero del sistema se convierte finalmente en un elemento que logra combinar la analítica descriptiva con la analítica predictiva. Por otra parte, el modelo de riesgo de abandono para nuevos candidatos brinda un apoyo al equipo de recursos humanos para la toma de decisiones basadas en información de perfiles, que, si bien no da una verdad absoluta sobre el futuro de la persona en la compañía, puede facilitar al analista de contratación la creación de preguntas de mayor valor en el proceso de entrevistas y enriquecer su perspectiva. La probabilidad generada por este modelo tiene un porcentaje de precisión más bajo que el modelo de riesgo de empleados activos, debido a que el modelo para la calculadora toma

en cuenta únicamente variables que se pueden consultar en una entrevista de contratación, dejando de lado factores relacionados al desempeño y la satisfacción, por lo cual es primordial que la decisión de recursos humanos también esté enriquecida por el criterio experto.

Aunque el uso individual de los dos modelos mencionados genera información de valor a directivos y áreas de recursos humanos, el uso combinado de estos refuerza aún más el aprovechamiento de los datos. Una vez un candidato pasa por la evaluación de la calculadora de deserción y se convierte en un empleado activo, empieza a ser parte del tablero de alertas de abandono, la evolución de su riesgo de deserción mientras está en la compañía ofrece hallazgos importantes al área de contratación acerca del uso de la calculadora para nuevos candidatos.

La creación de este sistema de análisis de deserción requirió capacidades técnicas en distintas herramientas y bases sólidas en manejo de la información, sin embargo, una vez se implemente en la farmacéutica, ésta debería contar con personal técnico capaz de realizar actualizaciones periódicas sobre los conjuntos de datos utilizados, además de administrar correctamente el sistema centralizado de almacenamiento y establecer un proceso de gobernanza de datos firme. Lo anterior, acompañado de una buena infraestructura tecnológica es fundamental para el soporte y mantenimiento del sistema, y aseguraría su uso a largo plazo.

Referencias Bibliográficas

Arias, E. R. (2021, febrero 1). *Variable dicotómica*. Economipedia. https://economipedia.com/definiciones/variable-dicotomica.html

Bioclever. (2021, mayo 8). *Duración de los ensayos clínicos y los estudios observacionales*. Bioclever. https://www.bioclever.com/es-ES/duracion-de-los-ensayos-clinicos-los-estudios-observacionales-n-46-es

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72

Data Universe. (2023, marzo 1). *Modelo Operativo Centralizado*. https://data-universe.org/modelo-operativo-centralizado/

Deloitte. (s/f). *¿Qué es Scrum?* Deloitte Spain. Recuperado el 23 de noviembre de 2023, de https://www2.deloitte.com/es/es/pages/technology/articles/que-es-scrum.html

Díaz Monroy, L. G. (2007). *Estadística multivariada: Inferencia y métodos* (2. ed). Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia.

Fitzgerald, D., & Wilson, C. (2023, junio 26). *The Structure and Departments in a Pharmaceutical Manufacturing Company*. GetReskilled. https://www.getreskilled.com/pharmaceutical-companies/structure/

IBM. (s/f-a). *¿Qué es el almacenamiento de datos? | IBM*. ¿Qué es el almacenamiento de datos? Recuperado el 23 de noviembre de 2023, de https://www.ibm.com/mx-es/topics/data-storage

IBM. (s/f-b). *¿Qué es la minería de datos? | IBM*. Recuperado el 23 de noviembre de 2023, de https://www.ibm.com/mx-es/topics/data-mining

IBM. (s/f-c). *¿Qué es machine learning? | IBM*. ¿Qué es machine learning? Recuperado el 23 de noviembre de 2023, de https://www.ibm.com/mx-es/topics/machine-learning

IBM. (s/f-d). *¿Qué es una arquitectura de datos? | IBM*. ¿Qué es una arquitectura de datos? Recuperado el 23 de noviembre de 2023, de https://www.ibm.com/es-es/topics/data-architecture

IBM. (2021a, marzo 8). *Esquemas de estrella*. https://www.ibm.com/docs/es/ida/9.1.2?topic=schemas-star

IBM. (2021b, agosto 17). *Tipos de modelos*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=mining-types-models

Jain, P. K., Jain, M., & Pamula, R. (2020). Explaining and predicting employees' attrition: A machine learning approach. *SN Applied Sciences*, *2*(4), 757. https://doi.org/10.1007/s42452-020-2519-4

Naeem, T. (2020, febrero 3). Conceptos de Data Warehouse: Enfoque de Kimball vs. Inmon. *Astera*. https://www.astera.com/es/tipo/blog/conceptos-de-almac%C3%A9n-de-datos/

Oracle. (s/f). *¿Qué es un almacén de datos?* Recuperado el 23 de noviembre de 2023, de https://www.oracle.com/co/database/what-is-a-data-warehouse/

Pourshasb, N. (2021, febrero 8). *¿Cuánto pierdes por el estrés financiero de tus trabajadores?* El Economista. https://www.eleconomista.com.mx/revistaimef/Cuanto-pierdes-por-el-estres-financiero-de-tus-trabajadores--20210208-0031.html

Anexos Técnicos

- Tablero Deserción: Archivo de Power BI para el tablero de deserción que contiene el panorama actual de la compañía y la gestión de alertas de deserción.
- Modelamiento-Probabilidad Deserción Candidatos: Archivo formato ipynb con el código construido en Python de los modelos entrenados para obtener la probabilidad de deserción de los candidatos.
- Modelamiento-Probabilidad Deserción Empleados: Archivo formato ipynb con el código construido en Python de los modelos entrenados para obtener la probabilidad de deserción de los empleados activos.
- Calculadora Deserción Candidatos: Archivo formato ipynb con el código construido en Python que contiene la integración del modelo elegido de probabilidad de deserción de candidatos, con la herramienta de la calculadora final creada dentro del entorno de Python.