Machine learning en el Ausentismo de trabajadores de mina
(Minería 4.0 - Optimización de procesos en Minería con Analítica de datos)

Ludmer Arcaya Arhuata[1], Mariana Corvera[2] y Renzo Di Tolla[3]

[1] Compañía Minera Antamina, Av. El Derby 055, Lima, Perú (larcaya@antamina.com 975953500)

[2] Compañía Minera Antamina, Av. El Derby 055, Lima, Perú (mcorvera@antamina.com 981902787)

[3] Compañía Minera Antamina, Av. El Derby 055, Lima, Perú (rditolla@antamina.com 990343196)

RESUMEN

El presente trabajo técnico describe el desarrollo de una herramienta de ausentismo predictivo, diseñada para identificar a los trabajadores con mayor probabilidad de ausentarse en los próximos seis meses. Esta capacidad permitirá activar campañas preventivas orientadas a mitigar el impacto del ausentismo en la producción de material metálico, optimizando así el uso de los camiones destinados al transporte del mismo durante el proceso de extracción.

Para la implementación de la solución se empleó la metodología CRISP-DM, ampliamente reconocida en el desarrollo de modelos predictivos, en este proceso se definieron los aspectos claves de la población de estudio, el alcance de la predicción, los factores relevantes asociados al ausentismo, la preparación de los datos, el entrenamiento y despliegue de los modelos predictivo.

Los datos fueron protegidos anonimizando al trabajador mediante encriptación y ofuscando debidamente sobre las categorias.

Los modelos entrenados permiten identificar al menos al 70 % con alta probabilidad de ausentarse en los siguientes seis meses, proporcionando así una base sólida para activar campañas de mitigación y reducir el impacto del ausentismo.

1. Introducción

El ausentismo laboral representa un desafío constante en el sector minero. Como referencia en otros trabajos similares se encontró que podría estar entre 7.13% y 12.4% según (Chamana, 2015).

Esta herramienta representa una innovación, ya que proporciona información anticipada sobre el comportamiento del ausentismo en trabajadores, siendo la primera vez que se implementa una solución de este tipo en Antamina. Además, incorpora algoritmos avanzados de machine learning que permiten identificar patrones de comportamiento y predecir el ausentismo.

La información predictiva sobre el ausentismo permite activar campañas de mitigación orientadas a reducir su impacto, contribuyendo así a mejorar la productividad en la producción de material metálico. Esto se traduce en una disminución de los tiempos de paralización de camiones asociados a la ausencia de los trabajadores.

Este desarrollo aporta conocimiento técnico al seguir un proceso completo: revisión de factores, exploración de datos y experimentación con diversos algoritmos. Además, se aplicaron principios de Machine Learning Operation (MLOps) para productivizar la solución de forma robusta y el cumplimiento de estándares de seguridad, arquitectura, software y gobierno de datos.

2. Objetivos

General:

Habilitar una herramienta predictiva para el ausentismo laboral.

Específicos:

Definir los elementos clave del alcance, incluyendo los datos requeridos, el universo de análisis y los factores potencialmente relevantes asociados al ausentismo.

Entrenar modelos predictivos utilizando técnicas de machine learning para identificar a los trabajadores con alta probabilidad de ausentarse en los próximos seis meses.

Implementar y desplegar los modelos siguiendo un enfoque de Machine Learning Operations (MLOps), asegurando su integración y operación en un entorno productivo.

3. Compilación de Datos y Desarrollo del Trabajo

Compilación de Datos:

La recopilación de datos se llevó a cabo a través de entrevistas con usuarios clave vinculados al proceso de ausentismo laboral, pertenecientes a las áreas de Bienestar Social, Compensación, Salud y Operaciones.

En estas entrevistas se recopiló información sobre la problemática actual del ausentismo, las categorías utilizadas en su registro, los principales casos ocurridos en los últimos años y sus causas. Además, se logró comprender dónde se almacenan los datos relacionados con los trabajadores y sus ausencias.

Por otro lado, se identificó en detalle cada uno de los datos del trabajador y su respectivo sistema fuente, así como la disponibilidad histórica de dicha información, necesaria para el análisis del ausentismo.

Desarrollo del trabajo:

El desarrollo de este trabajo se llevó a cabo siguiendo las buenas prácticas de la metodología CRISP-DM, ampliamente utilizada para la construcción de modelos predictivos. Esta metodología se estructura en seis fases: entendimiento del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. En las secciones siguientes se describe en detalle cada una de estas etapas aplicadas en el proyecto.

Entendimiento del negocio:

En esta etapa se identificaron las preguntas clave que orientaron el desarrollo del modelo, tales como: ¿Cuáles son los principales factores que influyen en el ausentismo?, ¿Con cuánta anticipación se necesita realizar la predicción?, y ¿A quiénes están dirigidas las acciones preventivas?

Para responder a estas preguntas, se revisó el alcance de las categorías de ausencias, como se muestra en la Figura 1. Asimismo, se identificaron posibles causas del ausentismo mediante la formulación de hipótesis, las cuales se presentan en la tabla 1.

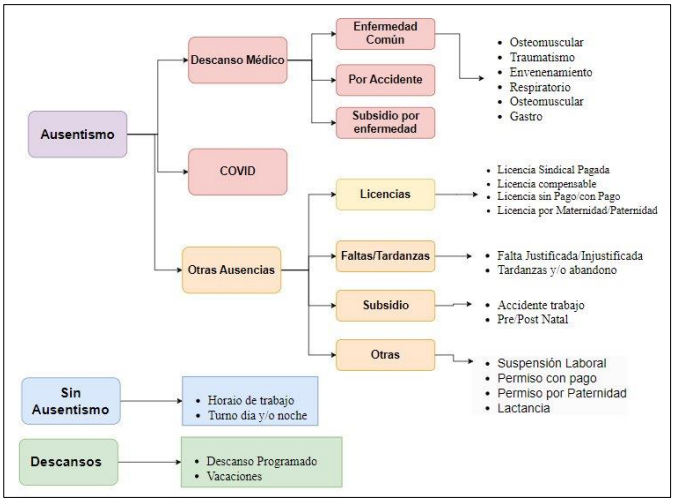

Figura 1. Categorización del Ausentismo Laboral
Fuente. Elaboración propia

Tabla 1. Factores potenciales relacionado con el ausentismo laboral

Dominios	Factores
Salud	Respiratorias, Osteomusculares, Digestivas, Mental, Traumatismos.
Socio-demográfico	Edad, Sexo, Antiguedad, Procedencia, Departamento, Hijos, Educación, y Estado Civil.
Personal	Estrés, Salud de hijo, Fallecimiento, Motivación, Claridad en Objetivos, Deudas, Retención Judicial, Cambios de Vivienda, Cambios de Celular.
Empresa	Gerencia, Procesos, Turno, Tipo Supervisión, Contrato, Flexibilidad horaria.
Contexto	Fiestas patronales, Feriados, Transporte, Clima, político, Económico, Estacionalidad, día, mes, feriados.
Rendimiento	Nivel de Satisfacción Laboral, Rendimiento de los objetivos

Fuente: Elaboración propia

El ausentismo se registra a nivel individual por trabajador, y la información relevante para el análisis abarca un horizonte de seis meses. Las acciones preventivas están dirigidas a los

trabajadores con mayor probabilidad de ausentarse durante ese periodo. Por ello, es fundamental contar con datos actualizados que permitan anticipar escenarios dentro de los próximos seis meses. En la siguiente tabla se detalla el grupo objetivo al cual están dirigidas dichas acciones, alineando así el enfoque del análisis predictivo.

Tabla 2. Especificaciones del análisis

Definicion	Detalle
Universo de Análisis	*Las asignaciones de trabajo durante la jornada Dia o Noche de los trabajadores realizadas durante 2022 y 2023, acotada según el Perfil potencial (Top 6 superintendencias de la Jornada JT-1. Debidamente anonimizada al trabajador usando encriptación y ofuscamiento.*
Unidad de estudio	*Asignación de trabajo del trabajador en un turno especifico (Dia O Noche)*
Target de Análisis	*Ausencia de trabajo en la asignación planificada por algún motivo según lo definido para esta fase.*
Periodo de anticipación	Se requiere predecir para los 6 meses siguientes

Fuente: Elaboración propia

Comprensión de los datos:

En esta etapa luego de acceder a los datos del trabajador debidamente anonimizada, encriptado, enmascarado y ofuscado, se procedió con la revisión en detalle la información relevante asociada al ausentismo, comenzando por la marca de ausentismo, la cual se registra diariamente según el tipo de ausencia, como se muestra en la Figura 1.

Los trabajadores que no presentan ausencias registran el concepto ejecutado de la actividad, así como los valores de planificación correspondientes al turno asignado. Cabe señalar que los días de descanso programado y las vacaciones no fueron considerados en el análisis.

En la siguiente figura se muestra el ranking de categorías de ausentismo junto con su acumulado. Por criterios de representatividad y conveniencia analítica, se excluyeron del análisis aquellas categorías que registraron menos de 100 ocurrencias.

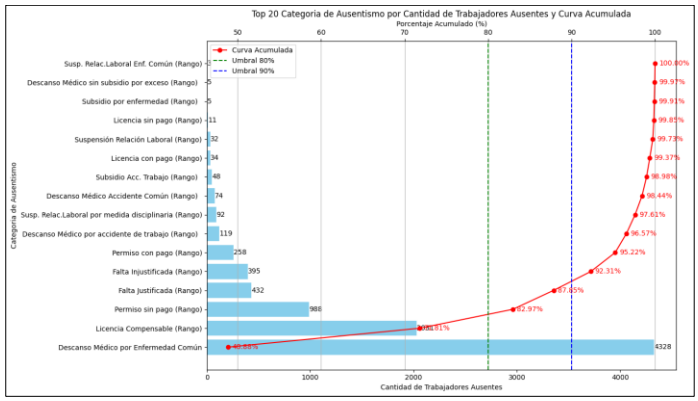

Figura 2. Distribución acumulada de las categorías del ausentismo
Fuente. Elaboración propia

Las superintendencias agrupan las áreas según su actividad específica. En este contexto, los trabajadores pertenecen a distintas superintendencias. En la siguiente tabla se presenta el ranking de superintendencias según la cantidad de registros de ausentismo. Por motivos prácticos y de representatividad, se definió trabajar con el top 6 superintendencias, las cuales concentran el 82.6 % del total de casos de ausentismo.

Tabla 3. Distribución del ausentismo por el top de Superintendencias, datos del año 2023

Top	Superintendencia	Ausentismo	Porcentaje	Acumulado
1	A	3038	46%	46.10%
2	B	696	11%	56.60%
3	C	627	10%	66.10%
4	D	418	6%	72.50%
5	E	417	6%	78.80%
6	F	253	4%	82.60%
7	G	177	3%	85.30%
8	H	154	2%	87.60%
9	I	120	2%	89.50%
10	J	112	2%	91.20%

Fuente: Elaboración propia

La jornada asignada al trabajador representa el periodo operativo durante el cual realiza sus actividades. En la siguiente figura se observa que el ausentismo se concentra principalmente en la jornada tipo JT -1, la cual corresponde a un ciclo de 10 días de actividad seguidos por 10 días de descanso programado.

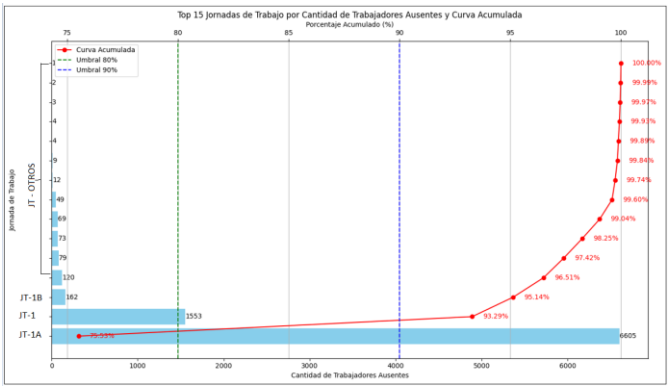

Figura 3. Distribución acumulada del ausentismo laboral por Jornada
Fuente. Elaboración propia

El universo de análisis está conformado por el conjunto de eventos en los que se ejecutó la planificación de actividades y que pudieron haber registrado ausencias efectivas o, en su defecto, el cumplimiento de la jornada por parte del trabajador operario. Para este estudio, se consideró únicamente a los trabajadores pertenecientes al top 6 de superintendencias con mayor proporción de ausentismo y que operan bajo la jornada JT-1.

Aplicando este criterio, en la figura siguiente se observa que esta población representa el 56.6 % del total de trabajadores, concentrando el 83.4 % del ausentismo laboral. De esta manera, se define y justifica el foco del análisis.

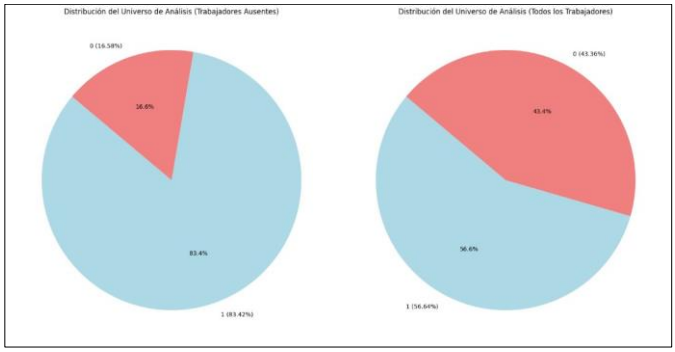

Figura 4. Proporción del ausentismo laboral por el Universo de Análisis
Fuente. Elaboración propia

Se procedió a realizar un análisis bivariado con el objetivo de explorar la relación entre los factores potenciales identificados y la ocurrencia de ausentismo. Este análisis permite evaluar cómo varía el comportamiento del ausentismo en función de distintas características de los trabajadores.

En la figura siguiente, un diagrama de caja (box plot) muestra la comparación de edades entre trabajadores con y sin registros de ausencias. Los resultados evidencian que aquellos que presentan ausentismo tienden a tener una edad ligeramente superior respecto a quienes no se ausentan, lo que sugiere una posible asociación entre la edad y la propensión al ausentismo.

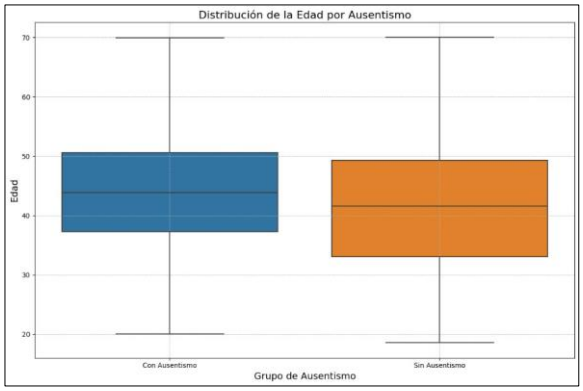

Figura 5. Comparativa de edad entre grupos con y sin ausentismo.
Fuente. Elaboración propia

Se aplicó una categorización por grupos generacionales basada en la definición de generación digital, clasificando a los trabajadores según su año de nacimiento. Las categorías consideradas fueron: Generación Z (2001–2017), Generación Y o Millennials (1980–2000), Generación X (1965–1979) y Baby Boomers (1945–1964).

En la figura siguiente, se muestra una línea verde que representa el porcentaje de ausentismo dentro de cada grupo generacional. Se observa que este porcentaje tiende a incrementarse conforme se avanza hacia generaciones de mayor edad, lo que sugiere una posible relación entre la antigüedad generacional y la propensión al ausentismo.

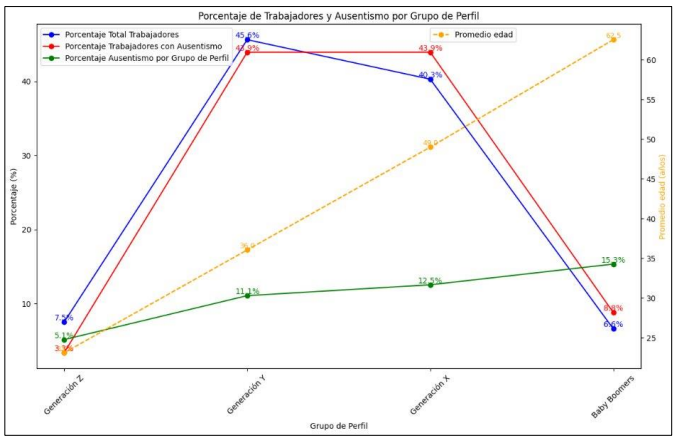

Figura 6. Comparación de perfiles de Generación digital por Incidencia de ausentismo
Fuente. Elaboración propia

El tiempo de servicio, medido en años, corresponde al período transcurrido desde el ingreso del trabajador a la compañía. Este indicador permite analizar la relación entre la antigüedad laboral y la ocurrencia de ausentismo.

En la figura siguiente se presenta una comparativa entre las distribuciones del tiempo de servicio de trabajadores con y sin registros de ausentismo. Se observa una diferencia clara entre ambos grupos, lo que sugiere que la antigüedad podría estar asociada a una mayor o menor propensión a ausentarse.

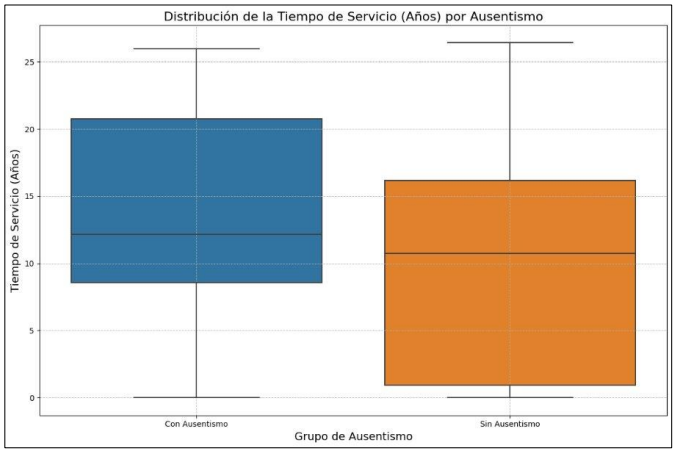

Figura 7. Comparativa de Tiempo de servicio entre grupos con y sin ausentismo
Fuente. Elaboración propia

Para facilitar el análisis del tiempo de servicio, se categorizaron los trabajadores en niveles de crecimiento dentro de la compañía, según su antigüedad laboral. Los grupos definidos fueron: Nuevo (hasta 6 meses), Inicio (más de 6 meses hasta 1 año), Crecimiento (de 1 a 5 años), Madurez (de 6 a 10 años) y Veterano (más de 10 años).

En la figura siguiente se muestra el nivel de incidencia de ausentismo (porcentaje dentro de cada grupo), observándose una relación directa entre el tiempo de servicio y la probabilidad de ausentismo. A medida que aumenta la antigüedad, se incrementa también la tasa de ausencias registradas.

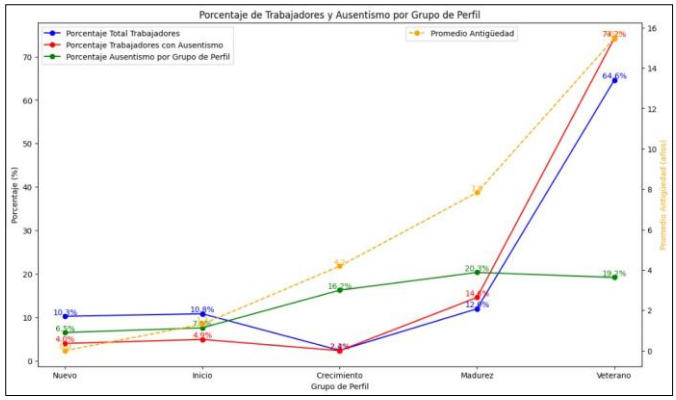

Figura 8. Comparación de perfiles de Crecimiento por Incidencia de ausentismo
Fuente. Elaboración propia

El turno de la jornada de trabajo puede estar asignado según la guardia, ya sea en turno de día o de noche. Esta variable permite analizar si existen diferencias en la incidencia del ausentismo en función del horario laboral.

En la figura siguiente se observa que los trabajadores con planificación en el turno de día presentan una mayor incidencia de ausentismo en comparación con los del turno noche. Esta diferencia podría estar influenciada por diversos factores externos, como obligaciones personales, condiciones ambientales o eventos sociales que afectan más a quienes trabajan durante el día.

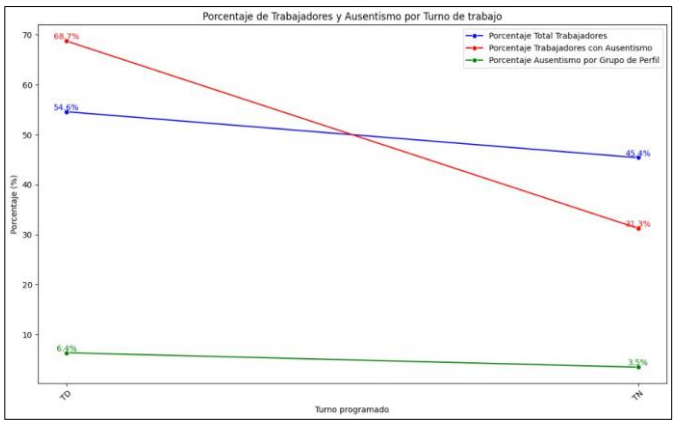

Figura 9. Comparación de perfiles del Turno asignado por Incidencia de ausentismo
Fuente. Elaboración propia

Se realizó una comparativa del nivel de incidencia de ausentismo entre las distintas superintendencias. Esta comparación permite identificar áreas específicas con mayores niveles de ausencias, lo cual es clave para orientar acciones focalizadas.

En la figura siguiente se observa que, si bien la mayor proporción de trabajadores se encuentra en la superintendencia A, su nivel de incidencia de ausentismo es del 16 %. En contraste, la superintendencia E, con menor proporción de trabajadores, presenta una incidencia significativamente mayor, alcanzando el 22 %.

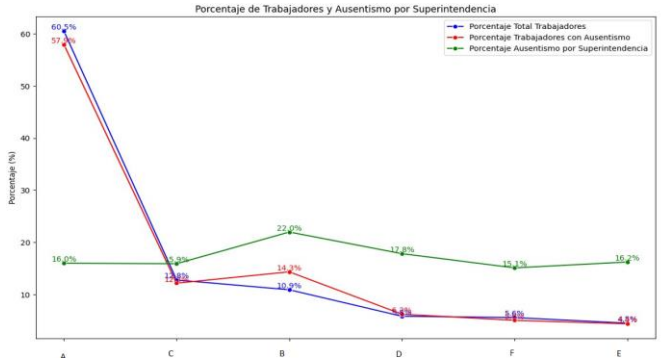

Figura 10. Comparación de perfiles del top 6 de Superintendencias por Incidencia de ausentismo
Fuente. Elaboración propia

En la figura siguiente se analizan los periodos con mayores picos de ausentismo registrados entre 2023 y 2024. Se identificaron fechas específicas en las que se concentran estos aumentos, las cuales coinciden con celebraciones clave como fiestas patrias, festividades patronales, pagos de bonos y otras fechas significativas. Estos hallazgos están alineados con las hipótesis previamente formuladas sobre la influencia de eventos externos en el comportamiento del ausentismo.

Las fechas con mayor concentración de ausencias incluyen:

- 19 de diciembre de 2022 y días circundantes (Navidad),
- Del 7 al 12 de mayo de 2023 (Aniversario del distrito de San Marcos),
- Del 26 al 31 de julio de 2023 (Fiestas Patrias),
- 14 y 15 de octubre de 2023 (Fiesta patronal del distrito de San Marcos),
- 23 de diciembre de 2023 y días circundantes (Navidad).

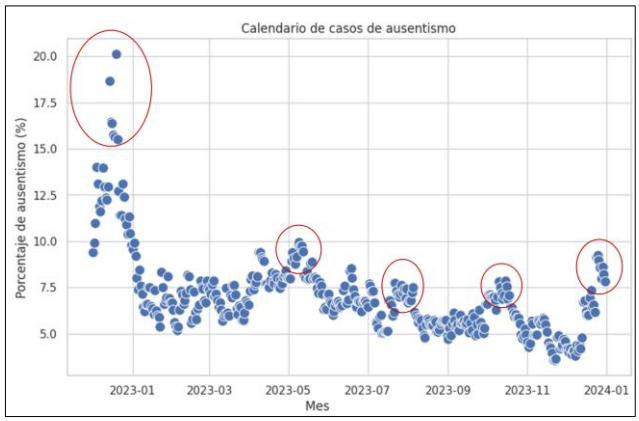

Figura 11. Ausentismo registrado por fechas
Fuente. Elaboración propia

Preparación de los datos:

En la siguiente tabla se presentan las distribuciones correspondientes a los distintos periodos definidos para cada "foto de análisis". Cada foto representa un punto de corte temporal desde el cual se analiza la información histórica del trabajador y se proyectan los eventos futuros. Específicamente, se consideran cuatro meses de historial previos al mes de corte, así como un horizonte de seis meses hacia adelante, correspondiente a la planificación de actividades.

A partir del conjunto de datos disponibles, se definió una lista de fotos de análisis. Cada una de estas fotos constituye un subconjunto de datos que permite entrenar modelos predictivos y, a su vez, validar su desempeño en periodos posteriores. Este enfoque facilita una evaluación robusta del rendimiento del modelo bajo distintos escenarios temporales.

Tabla 4. Distribución de meses considerados para las fotos de análisis

Foto	Set de Datos	Historia del Trabajador			Mes Corte	1	2	3	4	5	6
		MH4	MH3	MH2	MH1	MF1	MF2	MF3	MF4	MF5	MF6
1	Training	202201	202202	202203	202204	202205	202206	202207	202208	202209	202210
2	Training	202202	202203	202204	202205	202206	202207	202208	202209	202210	202211
3	Training	202203	202204	202205	202206	202207	202208	202209	202210	202211	202212
4	Training	202204	202205	202206	202207	202208	202209	202210	202211	202212	202301
5	Training	202205	202206	202207	202208	202209	202210	202211	202212	202301	202302
6	Training	202206	202207	202208	202209	202210	202211	202212	202301	202302	202303
7	Training	202207	202208	202209	202210	202211	202212	202301	202302	202303	202304
8	Training	202208	202209	202210	202211	202212	202301	202302	202303	202304	202305
9	Training	202209	202210	202211	202212	202301	202302	202303	202304	202305	202306
10	Training	202210	202211	202212	202301	202302	202303	202304	202305	202306	202307
11	Training	202211	202212	202301	202302	202303	202304	202305	202306	202307	202308
12	Training	202212	202301	202302	202303	202304	202305	202306	202307	202308	202309
13	Training	202301	202302	202303	202304	202305	202306	202307	202308	202309	202310
14	Training	202302	202303	202304	202305	202306	202307	202308	202309	202310	202311
15	Training	202303	202304	202305	202306	202307	202308	202309	202310	202311	202312
16	Training	202304	202305	202306	202307	202308	202309	202310	202311	202312	202401
17	Training	202305	202306	202307	202308	202309	202310	202311	202312	202401	202402
18	Training	202306	202307	202308	202309	202310	202311	202312	202401	202402	202403
19	Training	202307	202308	202309	202310	202311	202312	202401	202402	202403	202404
20	Training	202308	202309	202310	202311	202312	202401	202402	202403	202404	202405
21	Training	202309	202310	202311	202312	202401	202402	202403	202404	202405	202406
22	Training	202310	202311	202312	202401	202402	202403	202404	202405	202406	
23	Training	202311	202312	202401	202402	202403	202404	202405	202406		
24	Training	202312	202401	202402	202403	202404	202405	202406			
25	Training	202401	202402	202403	202404	202405	202406				
26	Training	202402	202403	202404	202405	202406					
26	Testing	202403	202404	202405	202406	202407	202408	202409	202410	202411	202412

Fuente. Elaboración propia

Se prepararon los datos históricos correspondientes al periodo 2022–2024, integrando los registros de asistencia con los factores potenciales identificados. Para cada guardia, se habilitó la variable de interés que indica si el trabajador presentó o no una ausencia, permitiendo construir una base de datos estructurada para el análisis y modelado predictivo del ausentismo.

Se calcularon las variables explicativas asociadas al ausentismo laboral a partir de los hallazgos del análisis exploratorio efectuado en la fase de comprensión de datos. En Anexo se puede visualizar el detalle de los campos que se usan para los features, los cuales se empleó posteriormente para entrenar el modelo predictivo.

La unidad de estudio se definió en función de la asignación de actividades ejecutadas y planificadas por guardia del trabajador. De esta manera, los registros diarios se consolidaron en una estructura por turnos, permitiendo una representación más adecuada para el análisis del ausentismo.

Para el modelado se consideraron dos tipos de características (features): por un lado, aquellas calculadas a partir de los antecedentes de los tres últimos meses del trabajador; y por otro, las características contextuales, definidas en función de la información asociada a la fecha de planificación de la actividad.

El siguiente gráfico muestra el ranking de features categóricos asociado al ausentismo, este análisis permite identificar las variables explicativas relevantes con el ausentismo.

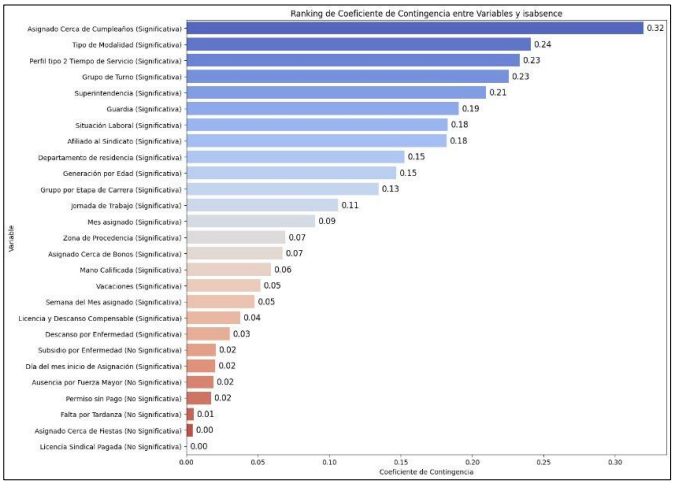

Figura 12. Ranking de asociación bivariada con el ausentismo (variables categóricas).
Fuente. Elaboración propia

La figura siguiente muestra las correlaciones entre variables, lo que permite identificar aquellas que están altamente relacionadas entre sí. Este análisis facilita la simplificación del conjunto de features, reduciendo la dimensionalidad sin perder información relevante.

Sobre este análisis permitió identificar las siguientes variables categóricas.

Socio Demográfico:

- Generación Digital (Grupo por Edad)
- Etapa de Carrera (Grupo por Tiempo de Servicio)

Temporalidad de asignación:

- Mes, Semana, Dia de semana
- Mano calificada
- Jornada
- Guardia

Historial:

- Enfermedad, Subsidios, Permisos, Licencias, Vacaciones.

Eventos claves:

- Cumpleaños, Feriados, Bonos.

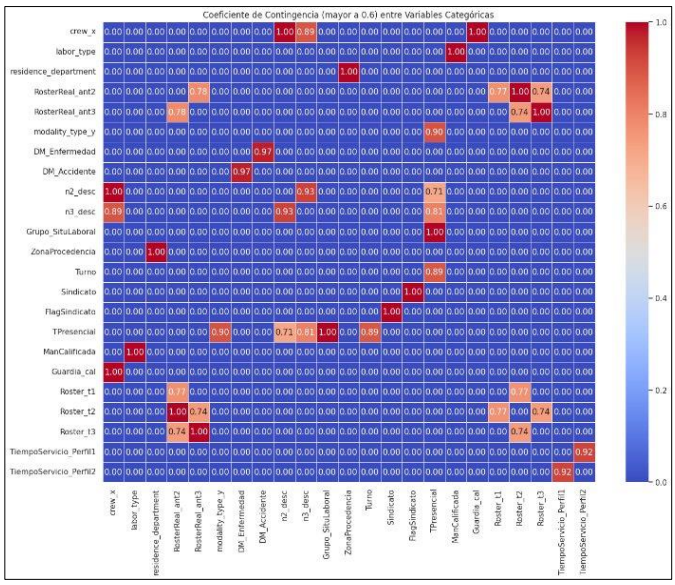

Figura 13. Matriz de correlaciones entre las variables categóricas
Fuente. Elaboración propia

Se construyó un mapa de correlaciones con el objetivo de identificar pares de variables numéricas altamente correlacionadas entre sí, utilizando un umbral de 0.7. En los casos donde existía una alta correlación, se descartó la variable con menor relación frente a la variable objetivo (target), con el fin de reducir la multicolinealidad y mejorar la eficiencia del modelo.

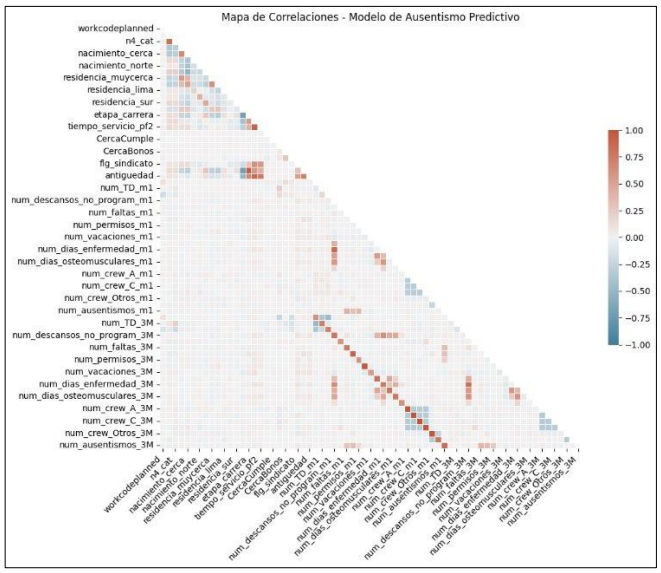

Figura 14. Matriz de correlaciones variables numéricas
Fuente. Elaboración propia

Modelado:

En esta etapa se probaron distintos algoritmos de machine learning, entre ellos regresión logística, random forest y XGBoost. Como resultado, los modelos de regresión logística y XGBoost destacaron por ofrecer un buen desempeño en términos de precisión y capacidad predictiva, siendo considerados como los más adecuados para el problema de ausentismo.

Para cada algoritmo se realizó un proceso de búsqueda del mejor modelo que lograra una adecuada convergencia con los datos. Esto permitió generar la mejor versión de cada algoritmo, conocida como modelo ganador, y posteriormente compararlos entre sí para evaluar cuál ofrecía el mejor desempeño en la predicción del ausentismo, en la siguiente tabla se presentan los hiperparametros usados y óptimos obtenidos.

Tabla 5. Hiperparámetro para la experimentación de algoritmos

Algoritmo	Hiperparámetro	Rango de hiperparámetro	Valor de Hiperparámetro
Regresión logistica	C	[0.0005, 0.001, 0.01, 0.1, 0.5, 1, 1.15]	0.001
	penalty		l1
	solver		liblinear
XGBoost	objective		binary:logistic
	eval_metric		auc
	max_depth	[1, 10]	8
	colsample_bytree	[0.01, 1.0]	0.913068646
	min_child_weight	[1, 20]	14
	learning_rate	[1e-5, 1e-1]	0.006299092
	n_estimators		800
	n_jobs		-1
	scale_pos_weight		1
RandomForest	n_estimators	[50, 100, 200]	50
	max_depth	[10, 15,20]	20
	min_samples_split	[2, 5, 10]	2
	min_samples_leaf	[1, 2, 4]	1
Decision Tree	criterion	['gini', 'entropy']	gini
	max_depth	[5,7,10]	10
	min_samples_split	[2, 5, 10]	2
	min_samples_leaf	[4,8]	4

Fuente: Elaboración propia

Las métricas obtenidas luego de la experimentación presentan a Xgboost como el algoritmo que mayor AUC ha logrado, seguido a Regresión logística, estos resultados nos indican que estos dos algoritmos han convergido mejor con la naturaleza de sus datos y logrando un recall de 71 y 74.

Figura 15. Resultado de experimento entre distintos algoritmos
Fuente. Elaboración propia

Ahora para reforzar el rendimiento obtenido se experimenta en un ensamble de ambos algoritmos, comparando el promedio vs máxima probabilidad de ausentismo, obteniendo mejores resultados, así mismo estos modelos presentan entre 73 a 75 de AUC en modelos entrenados para cada mes durante los 6 meses siguientes, puede visualizarse en la figura 18.

Evaluación:

La evaluación de los modelos se realizó utilizando validación cruzada, junto con la búsqueda de hiperparámetros dentro de un espacio finito de combinaciones. Para ello, se empleó una estrategia de búsqueda aleatoria optimizada, con el objetivo de encontrar la mejor configuración para cada algoritmo. Esta metodología fue aplicada a los modelos de Xgboost, Random Forest, regresión logística y árboles de decisión, permitiendo comparar su rendimiento de manera justa y robusta.

En las figuras siguientes se muestra la curva ROC correspondiente al conjunto de prueba para los modelos de Xgboost y regresión logística, así mismo el ensamble, combinando sus predicciones mediante el cálculo del promedio o del valor máximo de probabilidad, con el objetivo de mejorar la capacidad predictiva del modelo.

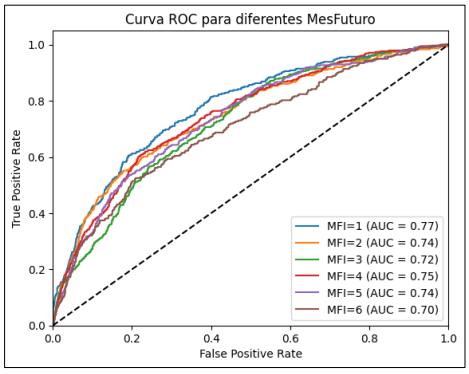

Figura 16. Resultado de Curva ROC para Xgboost
Fuente. Elaboración propia

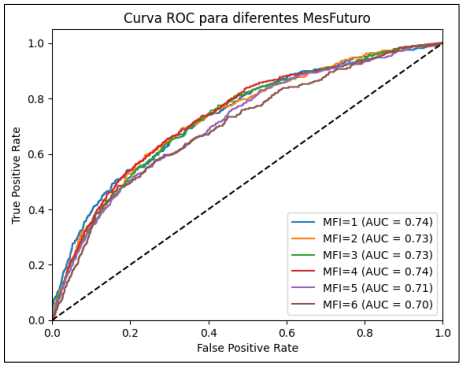

Figura 17. Resultado de Curva ROC para Regresión logística
Fuente. Elaboración propia

En la figura siguiente se presenta el resultado del ensamblado de los algoritmos, el cual muestra una mejora en el valor del AUC en comparación con los modelos individuales. Esta mejora se evidencia en distintos meses del horizonte de predicción, lo que indica un mejor desempeño general del modelo ensamblado.

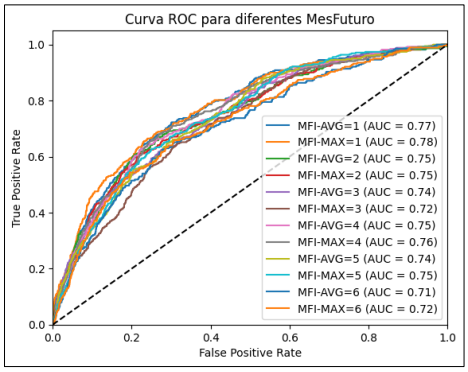

Figura 18. Resultado de Curva ROC para el Ensamble
Fuente. Elaboración propia

Los resultados obtenidos indican que el modelo alcanza un desempeño moderado. Para la clasificación, se definió un umbral de 0.5, seleccionado por conveniencia debido a su capacidad para alcanzar un recall mínimo del 70 %. Este valor permite capturar una proporción significativa de los casos reales de ausentismo, lo que lo hace adecuado para fines preventivos.

En la tabla siguiente se muestran los resultados de la ejecución del modelo bajo distintos umbrales y algoritmos, destacando la coincidencia entre los casos reales de ausentismo y los trabajadores identificados como propensos. Este comportamiento consistente valida el uso del modelo como una herramienta efectiva para

anticipar y mitigar el impacto del ausentismo en las operaciones.

Tabla 6. Resultados del modelo con distintos umbrales

modelo	umbral	accuracy	precision	recall	f1	roc_auc
1	0.30	0.303711	0.080808	0.951220	0.148962	0.771351
1	0.35	0.438867	0.094874	0.908537	0.171807	0.771351
1	0.40	0.493945	0.102564	0.890244	0.183937	0.771351
1	0.45	0.551562	0.108592	0.832317	0.192118	0.771351
1	0.50	0.673633	0.132458	0.737805	0.224594	0.771351
1	0.55	0.778320	0.163470	0.597561	0.256713	0.771351
1	0.60	0.828516	0.188209	0.506098	0.274380	0.771351
2	0.30	0.307754	0.076716	0.946844	0.141932	0.748909
2	0.35	0.464243	0.088665	0.847176	0.160529	0.748909
2	0.40	0.515267	0.092593	0.797342	0.165918	0.748909
2	0.45	0.590599	0.102881	0.747508	0.180868	0.748909
2	0.50	0.747891	0.142429	0.631229	0.232416	0.748909
2	0.55	0.816593	0.170968	0.528239	0.258327	0.748909
2	0.60	0.834472	0.176761	0.475083	0.257658	0.748909
3	0.30	0.343449	0.092458	0.933486	0.168251	0.741306
3	0.35	0.417523	0.100255	0.901376	0.180441	0.741306
3	0.40	0.457987	0.104007	0.869266	0.185784	0.741306
3	0.45	0.552456	0.111485	0.759174	0.194420	0.741306
3	0.50	0.694240	0.139418	0.637615	0.228807	0.741306

Fuente: Elaboración propia

En la siguiente tabla se presentan los resultados de la predicción correspondientes al mes de corte noviembre de 2024 (202411), considerando un horizonte de seis meses, desde diciembre de 2024 (202412) hasta mayo de 2025 (202505). Este análisis permite anticipar con antelación a los trabajadores con mayor probabilidad de ausentarse en ese periodo.

La métrica utilizada para evaluar el desempeño del modelo es el recall, que refleja la capacidad de identificar correctamente los casos positivos de ausentismo. En este caso, el modelo logró detectar un 75.36 % de los trabajadores propensos a ausentarse, lo que evidencia un buen poder predictivo y refuerza su utilidad en la planificación de acciones preventivas.

El análisis del ranking de importancia de variables revela que los factores históricos de salud y ausentismo son los más determinantes para predecir la probabilidad de que un trabajador presente ausencias en los próximos seis meses.

En particular, el número de días de descanso médico del mes anterior representa la variable más influyente (8 %), seguida del número de días de suspensión (6 %) y de descansos médicos por lesiones osteomusculares (4 %). Estos hallazgos confirman que el comportamiento reciente del trabajador es un fuerte predictor de su posible ausentismo futuro.

Por otro lado, las características estructurales del trabajador también tienen un peso considerable en la predicción. Variables como el tiempo de servicio (7 %).

El modelo también destaca la relevancia de factores acumulativos de comportamiento, como el número de días de ausentismo, cambios de guardia y descansos médicos en los últimos tres meses, con una importancia promedio de 3 %. Esta información complementa la visión de corto plazo, aportando una perspectiva más amplia sobre los patrones de comportamiento del trabajador en periodos anteriores.

Asimismo, se identifican como relevantes diversos factores contextuales vinculados al calendario y entorno personal, como la cercanía a fiestas patronales, cumpleaños y feriados nacionales. Aunque su peso individual es menor (alrededor de 2 %–3 %), su impacto es consistente con hipótesis previamente planteadas sobre el efecto de fechas especiales en la probabilidad de ausentismo.

Finalmente, el modelo también reconoce la influencia de variables organizacionales y geográficas, como la superintendencia de origen, el tipo de guardia asignada. Si bien estas variables tienen un peso individual moderado, su presencia recurrente en el ranking evidencia que el contexto operativo y la logística personal del trabajador pueden influir en su disponibilidad para asistir al trabajo.

Tabla 7. Factores potenciales relacionado con el ausentismo laboral

Variable	importancia
Num.dias.descanso.médico.mes.anterior.predict	8%
Grupo.tiempo.servicio	7%
Num.dias.suspensiones.mes.anterior.predict	6%
Num.dias.dm.osteomuscular.mes.anterior.predict	4%
Num.dias.dm.traumatismo.mes.anterior.predict	4%
Num.dias.ausentismo.3meses.anterior.predict	4%
Num.guardias.distintas.3meses.anterior.predict	3%
Cercanía.fiestapatronal.rango7dias	3%
Num.dias.licencia.3meses.anterior.predict	3%
Pct.cumplimiento.trabajo.programado.mes.anterior.predict	3%
Cercanía.cumpleaños.rango7dias	2%
Superintencias.top.ausentismo	2%
Num.dias.faltas.justi.nojust.3meses.anterior.predict	2%
Num.dias.ausencia.fuerzamayor.mes.anterior.predict	2%
Num.dias.ausencia.fuerzamayor.3meses.anterior.predict	2%
Zona.residencia.norte	2%
Guardias.tipo.V1.V2	2%
Num.dias.vacaciones.mes.anterior.predict	2%
Residentes.Prov.Huari	2%
Num.dias.permiso.mes.anterior.predict	2%
Num.dias.vacaciones.3mesese.anterior.predict	2%
Zona.nacimiento.norte	2%
Zona.Residentes.Sur	2%
Num.dias.permiso.3meses.anterior.predict	2%
Pct.cumplimiento.trabajo.programado.3meses.anterior.predict	2%
Guardia.tipo.C	2%
Num.turnosNoche.mes.anterior.predict	2%
Grupo.generacion.edad	2%
Lugar.Nacimiento.Lima	2%
Cercanía.feriados.rango7dias	2%
Guardia.tipo.D	2%
Guardia.tipo.A	2%
Residentes.Lima	2%

Fuente: Elaboración propia

Despliegue:

Para el despliegue de la herramienta predictiva se han definido tres artefactos clave. El primero es un proceso de reentrenamiento mensual, que permite actualizar el modelo incorporando los datos más recientes. Esto garantiza que el modelo se mantenga vigente y ajustado a los cambios en el comportamiento del ausentismo a lo largo del tiempo.

El segundo artefacto corresponde al proceso de inferencia, que se ejecuta mensualmente y genera predicciones para los siguientes seis meses a partir del mes de corte. Finalmente, se ha desarrollado un dashboard interactivo orientado al usuario funcional, donde se visualizan los resultados del modelo, permitiendo identificar a los trabajadores con mayor probabilidad de ausentarse y así activar campañas de mitigación oportunas.

Para los modelos seleccionados, se desplegaron flujos automatizados que ejecutan de forma mensual las distintas etapas del proceso, incluyendo la ingesta de datos, el entrenamiento del modelo, la generación de inferencias y la actualización del dashboard, asegurando así un ciclo continuo de actualización y entrega de valor al usuario.

Todo este proceso automatizado sigue los principios de Machine Learning Operations el cual permite automatizar los despliegues, garantizando que la operación de los modelos estará disponible en cada actualización del modelo.

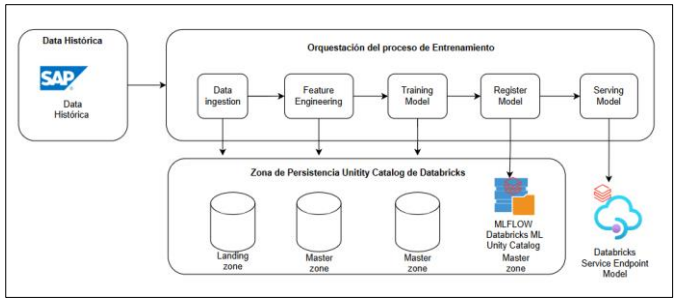

Figura 18. Flujo de Entrenamiento del Modelo
Fuente. Elaboración propia

La disponibilización de la información predictiva se realiza al inicio de cada mes, tomando como referencia el mes calendario que acaba de finalizar. Este proceso se conoce como proceso de inferencia, y tiene como objetivo generar las predicciones mensuales de ausentismo laboral a futuro.

Durante este flujo, se realiza la ingesta de nuevos datos, que incluyen tanto la información del mes recientemente cerrado como la planificación de los seis meses siguientes. A partir de estos datos, se construyen las características actualizadas de cada trabajador y se generan las predicciones correspondientes para el periodo futuro. En la figura siguiente se muestra el proceso completo que permite automatizar esta operación.

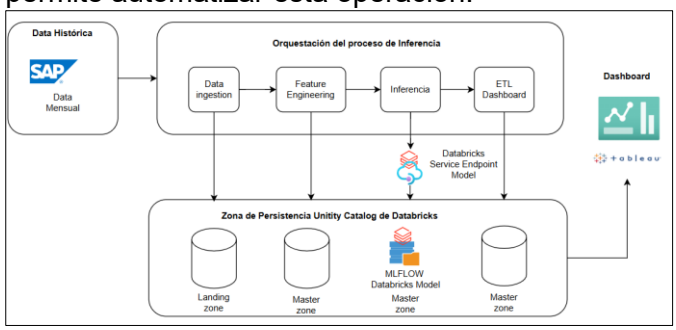

Figura 19. Flujo de Inferencia del modelo
Fuente. Elaboración propia

La siguiente figura presenta la evidencia de las ejecuciones satisfactorias del reentrenamiento del modelo, así mismo el despliegue de este,

cumpliendo con el concepto de Machine Learning Operations.

Figura 20. Flujo de ejecución mensual del reentrenamiento del modelo
Fuente. Elaboración propia

l

4. Presentación y discusión de resultados

El análisis evidenció que seis superintendencias concentran al menos el 80 % del ausentismo registrado. Este hallazgo facilitó la focalización del estudio en los perfiles de mayor riesgo, optimizando el enfoque preventivo y predictivo del modelo.

El análisis exploratorio identificó varios factores asociados a una mayor propensión al ausentismo entre los trabajadores. En primer lugar, se observó que los trabajadores con registros de ausencias tienden a tener una edad ligeramente superior, así como una mayor antigüedad laboral, en comparación con quienes no presentan ausentismo. Esta tendencia se refuerza al analizar las categorías generacionales, donde se evidencia un incremento progresivo en los niveles de ausentismo conforme aumenta la edad.

El modelo identificó como variable influyente importante, el antecedente de las dolencias con el concepto de descanso médico, así mismo las dolencias presentadas en los últimos tres meses, brinda oportunidad para acciones del tipo concientización médica.

De manera complementaria, al clasificar el tiempo de servicio por etapas de crecimiento dentro de la organización, se observó un patrón similar: los trabajadores con mayor antigüedad presentan mayores niveles de ausencias.

Asimismo, se observó una diferencia notable en la incidencia de ausentismo según el turno asignado. Los trabajadores programados en turnos diurnos registraron mayores niveles de ausencias en comparación con aquellos que laboran en turnos nocturnos, lo cual podría estar relacionado con factores personales, sociales o ambientales que afectan más a quienes trabajan durante el día.

Por otro lado, el análisis organizacional evidenció una concentración del ausentismo en seis superintendencias específicas, siendo la Superintendencia E la que registró el nivel más alto con un 22 % de incidencia. Finalmente, el análisis temporal permitió identificar picos de ausentismo que coinciden con celebraciones festivas, pagos y otros eventos clave, lo que confirma la influencia de factores externos en los patrones de asistencia laboral.

Los algoritmos que mostraron mejor desempeño en la predicción del ausentismo fueron XGBoost y Regresión Logística. El rendimiento del modelo se incrementó al combinar ambos enfoques mediante un esquema de ensamblado, alcanzando valores de AUC entre 72 % y 77 % en los seis modelos desarrollados. Estos resultados son prometedores, ya que permiten identificar hasta el 70 % de los trabajadores con alta probabilidad de ausentarse, lo que representa una base sólida para implementar acciones preventivas de forma anticipada.

Los modelos predictivos fueron operacionalizados mediante un proceso de Machine Learning Operations (MLOps), lo que permitió garantizar su estabilidad, actualización continua y disponibilidad de la información. Como parte del despliegue, se implementó un dashboard que se actualiza mensualmente, brindando información clave para la planificación de turnos y recursos. Esta herramienta permite identificar a los trabajadores con mayor probabilidad de ausentismo, facilitando la adopción de medidas preventivas y personalizadas para mitigar su impacto operativo.

5. Conclusiones

Gracias al uso de datos y herramientas analíticas, fue posible identificar el perfil de los trabajadores con mayor probabilidad de presentar ausentismo, enfocando el análisis en los trabajadores de las seis superintendencias que concentran el 80 % de los casos registrados.

La recopilación y validación de información por parte de las áreas clave permitió reconocer los principales factores asociados al ausentismo. Estos hallazgos fueron fundamentales para sustentar el

desarrollo de modelos predictivos basados en datos reales.

Se entrenaron modelos de machine learning, específicamente mediante algoritmos ensamblados, que lograron identificar hasta el 70 % de los trabajadores con alta propensión al ausentismo, lo que representa un avance significativo para la toma de decisiones preventivas.

Los modelos fueron desplegados en un entorno automatizado bajo prácticas de Machine Learning Operations (MLOps), lo cual garantiza su actualización periódica y la disponibilidad mensual de las predicciones para los usuarios finales.

Finalmente, la integración de los resultados en un dashboard desarrollado en Tableau permitió disponibilizar las predicciones de ausentismo para los seis meses siguientes, facilitando su uso en la planificación de turnos, asignación de recursos y definición de acciones de mitigación personalizadas.

6. Anexos

Tabla 8. Datos disponibles a partir de los factores potenciales

Campos	Descripción
id_encrypted	ID trabajador encriptado
workdate	Fecha del registro.
idperiod	Mes del registro (Formato AAAAMM).
workcodeplanned	Roster planificado en códigos.
workcode	Roster real en códigos. (registra la categoría del ausentismo)
dm_contingencydesc	Categoría de descanso médico cuando ocurre.
dm_diagnosisdesc	Diagnóstico de descanso médico cuando ocurre.
isabsence	Flag que indica ausentismo.
roster_id	Régimen de trabajo del empleado.
Gerencia	Posición en el organigrama Gerencia
Superintendencia	Posición en el organigrama Superintendencia
originalstartdate	Fecha de inicio de labores del trabajador.
emplstatus	Status del empleado (Activo)
birth_date	Fecha de nacimiento del trabajador.

Tabla 9. Features usados para el entrenamiento de modelos

feature	descripcion
edad	Edad actual del trabajador
antiguedad	Antigüedad del trabajador en la empresa (en años)
CercaCumple	Indicador si está cerca a la fecha de cumpleaños
CercaFiestas	Indicador si está cerca de un feriado nacional
CercaBonos	Indicador si está cerca a la fecha de entrega de bonos
CercaPatronal	Indicador si está cerca de fiestas patronales
CercaUtilidades	Indicador si está cerca a la fecha de pago de utilidades
num_DP_m1	Número de días de presencia el mes anterior
num_TN_m1	Número de turnos noche realizados el mes anterior
num_descansos_no_program_m1	Días de descansos no programados el mes anterior
num_faltas_m1	Número de faltas injustificadas el mes anterior
num_licencias_m1	Días con licencia el mes anterior

num_permisos_m1	Días con permisos personales el mes anterior
num_suspensiones_m1	Días de suspensión el mes anterior
num_vacaciones_m1	Días de vacaciones el mes anterior
num_dias_enfermedad_m1	Días con descanso médico por enfermedad el mes anterior
num_dias_traumatismos_m1	Días con descanso médico por traumatismos el mes anterior
num_dias_osteomusculares_m1	Días con descanso médico por lesiones osteomusculares el mes anterior
crew_A_m1	Número de asignaciones a guardia A el mes anterior
crew_B_m1	Número de asignaciones a guardia B el mes anterior
crew_C_m1	Número de asignaciones a guardia C el mes anterior
crew_D_m1	Número de asignaciones a guardia D el mes anterior
nunique_guardias_m1	Número de guardias distintas asignadas el mes anterior
flg_recordperfecto_m1	Indicador de cumplimiento perfecto de su jornada el mes anterior
num_DP_3M	Número de días de presencia en los últimos 3 meses
num_TD_3M	Número de turnos día en los últimos 3 meses
num_TN_3M	Número de turnos noche en los últimos 3 meses
num_descansos_no_program_3M	Descansos no programados en los últimos 3 meses
num_ausencias_3M	Número de ausencias totales en los últimos 3 meses
num_faltas_3M	Número de faltas injustificadas en los últimos 3 meses
num_licencias_3M	Número de días con licencia en los últimos 3 meses
num_permisos_3M	Número de días con permisos personales en los últimos 3 meses
num_dias_accidente_3M	Número de días por accidentes en los últimos 3 meses
num_dias_enfermedad_3M	Días con descanso médico por enfermedad en los últimos 3 meses
num_dias_traumatismos_3M	Días con descanso médico por traumatismos en los últimos 3 meses
num_dias_osteomusculares_3M	Días con descanso médico por lesiones osteomusculares en los últimos 3 meses
num_dias_digestivo_3M	Días con descanso médico por problemas digestivos en los últimos 3 meses
crew_A_3M	Cantidad de asignaciones a guardia A en los últimos 3 meses
crew_B_3M	Cantidad de asignaciones a guardia B en los últimos 3 meses
crew_C_3M	Cantidad de asignaciones a guardia C en los últimos 3 meses
crew_D_3M	Cantidad de asignaciones a guardia D en los últimos 3 meses
crew_Otros_3M	Cantidad de asignaciones a otras guardias en los últimos 3 meses
nunique_guardias_3M	Número de guardias distintas asignadas en los últimos 3 meses
flag_3um_isabsence	Indicador si hubo ausencia en los últimos 3 meses
WeekNumber_2	Pertenece a la segunda semana del mes
WeekNumber_4	Pertenece a la cuarta semana del mes
WeekDay_1	El turno empieza un lunes
WeekDay_7	El turno empieza un domingo
cured_workcodeplanned_TD	Asignado a trabajo planificado de turno día
cured_workcodeplanned_TN	Asignado a trabajo planificado de turno noche
n4_desc_Sup_de_C	Superintendencia de C
n4_desc_Sup_de_B	Superintendencia de B

n4_desc_Sup_de_D	Superintendencia de D
n4_desc_Sup_de_E	Superintendencia de E
n4_desc_Sup_de_F	Superintendencia de F
crew_group_Grupo_A	Guardia Grupo A
crew_group_Grupo_B	Guardia Grupo B
crew_group_Grupo_C	Guardia Grupo C
crew_group_Grupo_D	Guardia Grupo D
crew_group_V1	Guardia Grupo V1
crew_group_V2	Guardia Grupo V2
TiempoServicio_Perfil2_Crecimiento	Grupo Tiempo de Servicio en Crecimiento
TiempoServicio_Perfil2_Madurez	Grupo Tiempo de Servicio en Madurez
TiempoServicio_Perfil2_Veterano	Grupo Tiempo de Servicio en Veterano
Generacion_Generación_X	Perteneciente la Generación X
Generacion_Generación_Y	Perteneciente la Generación Y

Fuente: Elaboración propia

7. Referencias bibliográficas

Baltazar Chuqui, L. J. (2015). *Ausentismo laboral del personal del área de producción y su incidencia en la rentabilidad de la Empresa Sociedad Minera de Responsabilidad Limitada Condornegro de Chamana – Huamachuco* (Tesis de pregrado, Universidad Nacional de Trujillo). Repositorio institucional. https://hdl.handle.net/20.500.14414/2018

Berón, E. A., Mejía, D., & Castrillón, O. D. (2021). *Principales causas de ausentismo laboral: una aplicación desde la minería de datos*. **Información tecnológica**, 32(2), 11–20. https://doi.org/10.4067/S0718-07642021000200011

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. SPSS.

Databricks. (2024). MLflow documentation. Recuperado de https://docs.databricks.com

Databricks. (s.f.). *MLOps workflow*. Databricks Documentation. Recuperado de https://docs.databricks.com/aws/en/machine-learning/mlops/mlops-workflow

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3.ª ed.). Morgan Kaufmann.

Martínez, C. R. (2021). *Políticas de control de ausentismo: Manual para recursos humanos*. Editorial Universitaria.

Nisbet, R., Elder, J., & Miner, G. (2009). Handbook of statistical analysis & data mining applications. Academic Press.

Rista, A., Ajdari, J., & Zenuni, X. (2020). *Predicting and analyzing absenteeism at workplace using machine learning algorithms* (pp. 485–490). https://doi.org/10.23919/MIPRO48935.2020.9245118

Sarkar, D., & Natarajan, V. (2019). *Ensemble Machine Learning Cookbook*. Packt Publishing. Libro de recetas prácticas con ejemplos de implementación usando Keras, Scikit-Learn, H2O, XGBoost.

Tan, P.-N., Steinbach, M., & Kumar, V. (2021). *Introduction to Data Mining* (2ª ed.). Pearson.

Wexler, S., Shaffer, J., & Cotgreave, A. (2017). The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios. Wiley.

Zupančič, P., & Panov, P. (2024). *Predicting employee absence from historical absence profiles with machine learning*. *Applied Sciences, 14*(16), 7037. https://doi.org/10.3390/app14167037

8. Reseña Profesional

Ludmer Arcaya Arhuata

Supervisor de Ciencia de datos en Compañía Minera Antamina, Con más de 15 años de experiencia en Ciencia de datos e Inteligencia Artificial. Es Ingeniero en Informática y Sistemas por la Universidad Nacional Jorge Basadre Grohmnan (UNJBG) de Tacna, y una Licenciatura en Administración de Empresas por la Universidad de Tarapacá (Chile), con un MBA en Pacífico Business School (Perú), y una Maestría en Data Mining & knowledge Discovery por la Universidad de Buernos Aires (Argentina).

Mariana Corvera Ortiz

Supervisora de Administración de Personal en Compañía Minera Antamina. Cuenta con más de 15 años de trayectoria en el área de Administración de Personal, y como parte de sus funciones, cumplimiento legal y optimización de procesos de personal. Es Licenciada en Administración de Empresas por la Universidad Nacional Santiago Antúnez de Mayolo (UNASAM) en Áncash. Asimismo, cuenta con una Maestría en Desarrollo Organizacional y Dirección de Personas de la Universidad del Pacífico (Perú), obteniendo además un doble grado académico otorgado por la Universidad del Desarrollo (Chile).

Renzo Di Tolla

Superintendente de Compensación Total y Administración de Personal en Compañía Minera Antamina, Cuenta con más de 20 años de trayectoria en el área de Compensación total, Administración de Personal y Planeamiento de Recursos Humanos. Es Ingeniero Industrial por la PUCP (Pontificia Universidad Católica del Perú), con un MBA por CENTRUM PUCP.