UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

"DISEÑO DE MODELOS DE PREDICCIÓN DE AUSENTISMO Y ATRASOS LABORALES PARA UNA EMPRESA CONSULTORA EN RECURSOS HUMANOS"

MEMORIA PARA OPTAR AL TITULO DE INGENIERO CIVIL INDUSTRIAL

SEBASTIÁN RICARDO MUÑOZ ARRIAZA

PROFESOR GUÍA:
PABLO MARÍN VICUÑA

MIEMBROS DE LA COMISIÓN:
JUAN PABLO ROMERO
LORETO MARTÍNEZ GIMENEZ

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR AL
TITULO DE: Ingeniero Civil Industrial
POR: Sebastián Muñoz Arriaza
FECHA: 17/08/2020
PROFESOR GUÍA: Pablo Marín Vicuña

DISEÑO DE MODELOS DE PREDICCIÓN DE AUSENTISMO Y ATRASOS LABORALES PARA UNA EMPRESA CONSULTORA EN RECURSOS HUMANOS

SCM LATAM, es una consultora boutique, que está enfocada en el desarrollo de *software*, sistemas y servicios altamente tecnológicos, orientados al área de Recursos Humanos y Operaciones de sus empresas cliente. Durante los últimos años, esta compañía ha registrado un constante crecimiento, desde 2014 hasta 2019 incrementó sus ventas en un 276%.

Con el objetivo de diversificar su línea de negocios para atraer y fidelizar clientes, SCM LATAM ha explorado los problemas de sus empresas cliente de la industria de combustibles. Entre éstos destaca el ausentismo que, en promedio, representa cerca de un 1,5% de pérdida de la mano de obra programada, mientras que el atraso laboral un 1%. Esta situación puede llegar a ser muy costosa, pues las compañías deben pagar horas extra a los trabajadores no planificados, con el fin de cubrir los espacios vacíos provocados por ausentismos y atrasos.

Los factores mencionados anteriormente impactan directamente la productividad de una compañía. En la empresa analizada se calcula que el trabajador promedio falta un 1,5% de los días del año de manera injustificada, mientras que se atrasa 12,2% de sus días planificados. A esto se suma la incertidumbre de los atrasos y las ausencias no planificadas, por lo que la disminución y capacidad de predecir dichos fenómenos facilitaría la programación efectiva de trabajadores adicionales y ayudaría a aumentar la productividad, junto con reducir los costos de horas extra.

Frente a este escenario, esta tesis aborda la necesidad de diseñar y desarrollar dos modelos de predicción, uno que calcule atrasos y otro las ausencias injustificadas de una empresa cliente de la industria de combustibles. La metodología usada se basa en *Knowledge Discovery in Databases (*KDD*)*, debido a que su estructura permite retomar los pasos lógicos anteriores mientras se mejoran los modelos de predicción. Tras el desarrollo de este trabajo, el mejor modelo de predicción de atrasos corresponde a un *Random Forest* con un *AUC* de 91%, lo que lo convierte en un modelo de clasificación robusto y con alto poder estadístico. Bajo las condiciones y factores planteados en este estudio, este modelo ayudaría a mitigar las pérdidas de rentabilidad por atrasos en un 32%. Por otro lado, el mejor modelo de predicción de ausentismos también corresponde a un *Random Forest* con un *AUC* de 84%.

Se concluye que el ausentismo y atraso laboral se relacionan de manera similar con las variables estudiadas y en alto grado con el género, edad, antigüedad en la empresa, cargo, distancia al trabajo, estado civil, nacionalidad y precipitaciones. En el caso de los ausentismos también hay otras variables estacionales con alta importancia como el mes y día de la semana.

Los resultados sugieren que los *outputs* de los modelos pueden ser usados en herramientas que apoyen a los jefes de tienda a planificar de mejor manera la fuerza laboral y tareas a realizar, como por ejemplo un Scheduler con restricciones blandas que permita minimizar las probabilidades de atraso y ausentismo. También, se desprenden como producto comercial afín modelos de predicción de tendencias de atraso y ausentismos que serviría de apoyo en épocas de contratación.

Finalmente, se puede indicar que el tipo de modelos construidos en este trabajo, pueden dar origen a trabajos futuros que busquen el mejoramiento mediante modelos de aprendizaje profundo (*Deep Learning*) que servirían como punto de comparación o mejora.

A mis padres, a mis hermanas y a mi Nalita, porque tu efímera existencia me hizo un hombre más feliz.

Agradecimientos

Con este trabajo culmina un enorme proceso de aprendizaje que estuvo compuesto de momentos y personas que cumplieron un rol fundamental en esta etapa.

En primer lugar, quiero agradecer a mi familia; a mi mamá, Rebeca, porque me inculcaste el hábito de estudio, me entregaste tu amor y cariño incondicional; a mi papá, Ricardo, porque me aportaste con tu sabiduría y ejemplo que el trabajo duro y la dedicación son la única forma de lograr lo que uno se propone; a mi hermana mayor, Paula, por ser mi protectora, consejera y mejor amiga; a mi hermanita chica, Amanda, por ser mi compañera de juegos, locuras y demencias; y a mi perrita Nala por enseñarme el amor perruno y levantarme el ánimo en la última milla de mi trabajo.

A los “cabros” de la Universidad, Agustín, Rockets, Matías, Morales, Cristóbal, Lucca, Gastón (mechón 1), Coloro (mechón 2), Seba-Me y Pablito que me acompañaron en muchas locuras, viajes y carretes, sin duda no podría haber tenido un mejor grupo de aventuras.

A mis profesores Pablo y Juan Pablo, por su dedicación e interés, por guiarme y darme la oportunidad de aprender de ellos en cada encuentro virtual. La disposición que mostraron para resolver mis dudas es admirable y demuestra el compromiso de los profesores de esta Universidad.

Finalmente, a toda la gente de SCM que desde el día uno me incluyeron como uno más del equipo, en especial a Claudio, Seba y Jose por ayudarme a recolectar datos.

Tabla de contenido

1. Antecedentes generales

1.1 Características de la organización

1.1.1 Identificación y descripción del sector industrial

En sus inicios, la firma SCM LATAM, desde ahora SCM, se instaló como una consultora enfocada en el desarrollo de *software* y sistemas para recursos humanos. Con el tiempo ha migrado para constituirse como una *dealer*[1] de marcas internacionales de gestión de fuerza laboral, encargándose de la asesoría e implementación de estas soluciones. De esta manera, la compañía se ha insertado en el rubro de la consultoría.

En la industria de la consultoría se pueden establecer varias divisiones. La clasificación que predomina por sobre las demás es la que segmenta la industria de la consultoría de acuerdo al siguiente orden: Operaciones, Estrategia, Finanzas, Tecnología y Recursos Humanos. De acuerdo con este criterio, la compañía se clasifica como una consultora de nicho sobre tres de los cinco puntos descritos anteriormente: operaciones, tecnología y recursos humanos.

Entre estos nichos podrían hacerse más subdivisiones. Por ejemplo, para consultoras enfocadas en Gestión de la Fuerza Laboral, hay algunas dedicadas a la captación de personal, y otras, como SCM, están especializadas en proveer soluciones tecnológicas para apoyar las áreas de recursos humanos y operaciones de las empresas cliente.

1.1.2 Visión y Misión

Visión: "Modernizar la industria latinoamericana, siendo el puente entre las mejores TICS a nivel mundial y las múltiples realidades locales en Latinoamérica".

Misión: "Implementar y soportar tecnologías de información de clase mundial en las grandes empresas de Latinoamérica".

1.1.3 Organigrama

El diseño organizacional de SCM se puede analizar desde dos enfoques distintos. Por un lado, se tiene una estructura geográfica al tener presencia en Chile, Argentina y Perú, donde cada país se rige de acuerdo a sus respectivas gerencias generales. Estas gerencias deben responder al CEO, quien viaja constantemente entre los tres países.

En tanto, la organización interna de cada país se define por una estructura funcional. Bajo el mandato principal del gerente general respectivo, cada oficina local se subdivide en tres áreas fundamentales. La primera es el área comercial (ventas), liderada por un gerente comercial, encargado de los directores comerciales que llevan a cabo las tareas de venta directa, preventa y de relación con el cliente. En segundo lugar, se encuentra el área de administración y

[1] Proveedor de productos internacionales.

finanzas, que cuenta con un gerente dedicado a toda la gestión del personal legal, administrativo/cobranza y de contabilidad, para mantener el funcionamiento interno y equilibrar las inversiones que se realizan con los proyectos. Esta área lleva la estrategia a largo plazo de la empresa. En tercer lugar, el área de servicios se relaciona con el ámbito técnico de implementación y mantención de los sistemas dentro de las empresas clientes.

La estructura funcional se muestra de manera gráfica en el siguiente diagrama:

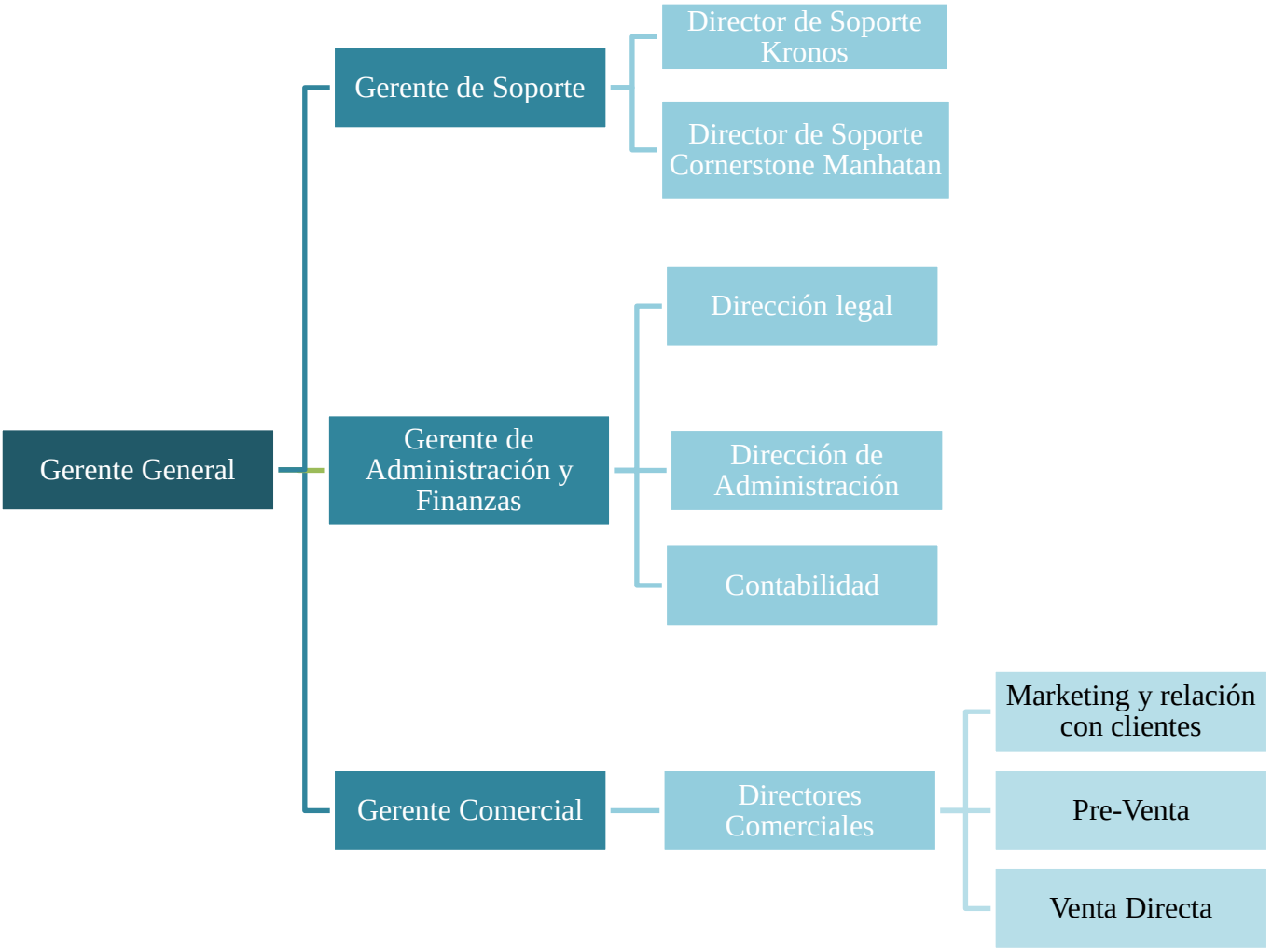


Diagrama 1: Estructura organizacional de SCM LATAM.

1.1.4 Productos y servicios

SCM es la única empresa latinoamericana *partner* de *KRONOS Incorporated*[2], *partnership* al que se le puede ofrecer la mayoría de sus productos y servicios.

[2] *KRONOS Incorporated* es una empresa con alto prestigio que administra un software diseñado por el *MIT* llamado *KRONOS*, para la gestión de fuerza laboral y capital humano, actualmente es considerado uno de los programas más robustos y potentes del mercado mundial.

Gran parte de las actividades de la firma, se enfocan en la solución de *Timekeeper*[3] que consiste en la recopilación de marcas en tiempo real, gestión de asistencia y horas laborales, junto con otros servicios extras que agregan valor, como soporte especializado, reportería personalizada y análisis de métricas e indicadores.

Otra de las actividades que generan valor en la firma, corresponde a la solución *Scheduling and Forecasting*[4], que consiste en determinar las necesidades del negocio mediante la generación de pronósticos de demanda; dotación óptima de trabajadores part-time y full-time; consultoría en productividad y operaciones; planificación de horarios; y optimización de recursos de la compañía.

Con respecto al servicio, primero se importan los equipos de hardware necesarios para la solución propuesta. Posteriormente, se ejecutan las tareas de implementación; configuración de equipos; configuración de las formas de pago, reportes y otras necesidades; y enrolamiento del personal e instalación. Finalmente, se abre paso a la capacitación del personal. Además, se cuenta con un servicio de soporte para la capacitación remota, resolución de problemas y desarrollo de nuevas funciones según las necesidades específicas del cliente.

Para los clientes de empresas más pequeñas se ha desarrollado una solución simple: *MyPalTime*[5], con el fin de reducir costos y ser accesible para sus necesidades específicas. En estos casos, el proceso de servicios es bastante similar, pero quitando algunos espacios de personalización de la solución.

Existen otros servicios y productos, pero no son el foco de este trabajo y tienen menor relevancia para la empresa.

1.1.5 Clientes y usuarios

Las oficinas de SCM en Argentina, Chile y Perú dirigen sus operaciones a otros países de la región, como Bolivia, Brasil y Uruguay, superando los 200 mil usuarios registrados en Latinoamérica. Esta expansión hacia nuevos mercados ha sido una constante durante los últimos años, dando buenos frutos, posicionando a la marca y llevando sus propuestas de solución a nuevos lugares. Asimismo, la mejora de los servicios ha permitido que los clientes capitalicen su confianza de mejor manera en el quehacer que impulsa la empresa en el rubro de la consultoría.

[3] Una de las funcionalidades más básicas de *KRONOS*.

[4] Una de las funcionalidades más potentes de *KRONOS*.

[5] Software desarrollado por SCM, que sirve como herramienta para que los trabajadores de las empresas cliente de SCM marquen su entrada y salida a trabajo.

1.1.6 Dimensionamiento de actividad realizada por la organización

SCM LATAM S.A es, como su nombre lo dice, una Sociedad Anónima compartida por dos socios. La razón social de la firma es "Sistemas de Clase Mundial", y el giro donde se mueve es "Asesorías en sistemas computacionales / Comercialización de *hardware* y *software*". Actualmente, cuenta con casi 30 trabajadores repartidos en las distintas oficinas de Chile, Argentina y Perú. Durante los últimos años ha tenido un crecimiento sostenido, ilustrado en la siguiente tabla:

Año	Crecimiento
2014	-
2015	22%
2016	35%
2017	30%
2018	30%
2019	30%

Tabla 1: Ventas y crecimiento anual de la compañía.

1.1.7 Ventaja competitiva en el mercado

SCM es una empresa con más de 20 años de experiencia en el mercado, con presencia en seis países latinoamericanos y una gran trayectoria que la respalda como una de las mejores consultoras chilenas de su categoría.

La ventaja competitiva más fuerte de SCM es su *partnership* con *KRONOS Incorporated*, empresa de alto prestigio y con más de 40 años de presencia en diversos países del mundo, caracterizada por contar con equipos de consultores especializados, generar valor a través de la innovación y mejorar constante sus productos y servicios. Por tanto, este *partnership* convierte a SCM en la única empresa latinoamericana que puede proveer el *software Kronos* en la región, otorgándole cierto poder monopólico.

1.2 Mercado y/o marco institucional

El mercado de la consultoría en Chile y Latinoamérica no recauda mucha información, pues se trata de una industria en desarrollo. Sin embargo, a nivel global presenta algunos antecedentes (Consultancy, 2016):

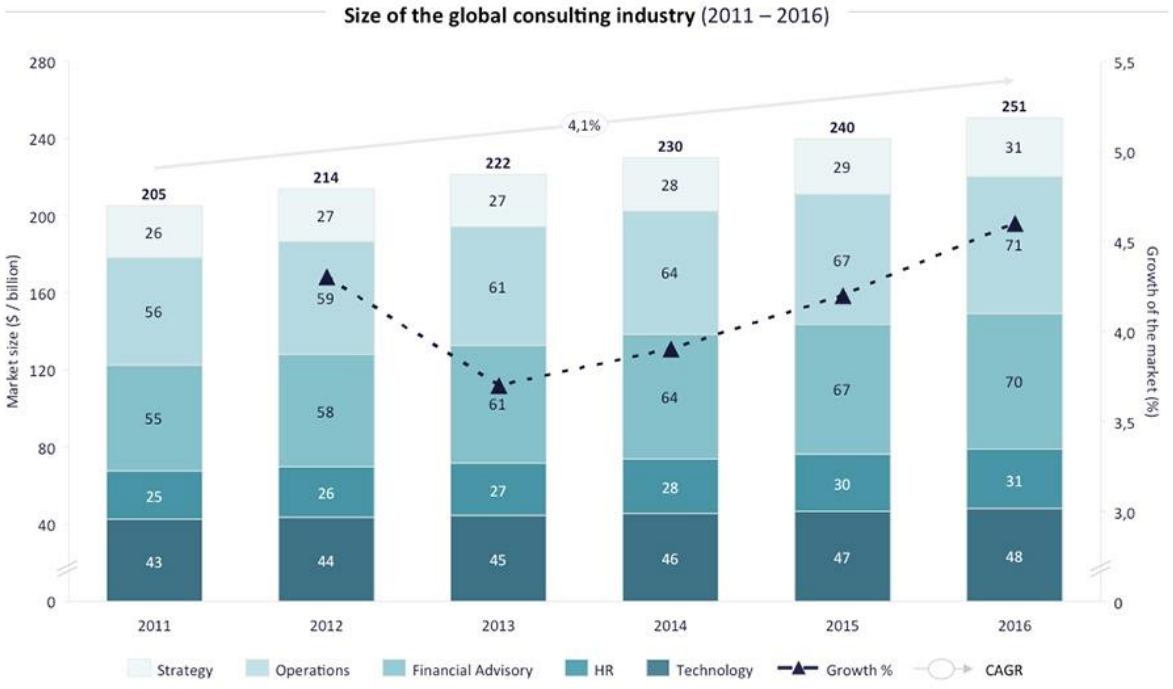


Diagrama 2: Tamaño (en billones de USD) y crecimiento de la industria mundial de la consultoría 2011- 2016.

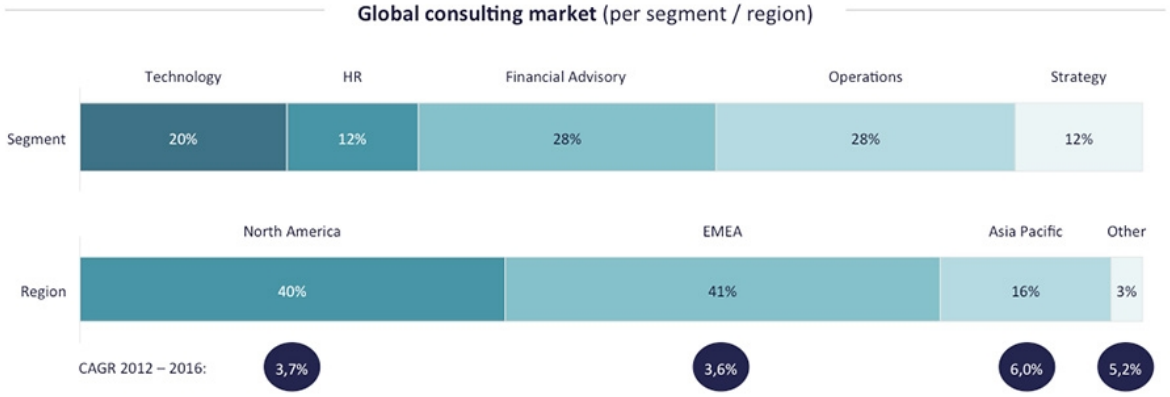


Diagrama 3: Segmentación del mercado mundial de la consultoría según tipo y región.

Respecto al mercado local en el que se encuentra la firma, no existen cifras para explicitar la participación de mercado. Hoy en día se presentan varios submercados que convocan a las marcas con las SCM compiten, como: Qwantec, GeoVictoria, Vortec, Punto Seguro, Genera, ADP, Shiftlabor y Reflexis, entre otros.

Otro punto a destacar es el marco institucional y regulatorio, que no posee apartados especializados en el rubro, pero si en la calidad y condiciones que deben cumplir los productos incluidos en las soluciones ofrecidas. Ejemplo de ello es la nueva normativa laboral de la Dirección del Trabajo (ORD. 1140/27 (Dirección de Trabajo, s.f.)), referente al registro de horas y asistencia.

1.3 Desempeño organizacional

Como se ve en el punto I.I.IV, desde 2015 SCM ha tenido una evolución importante, registrando un crecimiento del 30% en los últimos cuatro años. Esto refleja que aún se encuentra en etapa de desarrollo.

Por otro lado, entre los planes de expansión de SCM, se busca cubrir otros países del cono sur y explorar la factibilidad de abrir oficinas en alguna de las naciones con mayor presencia, como Brasil y Uruguay. En el mediano plazo se proyecta la ampliación de la oficina y el equipo de trabajo, integrando a sociólogos que ayuden a entender el comportamiento de los trabajadores; a ingenieros matemáticos que apoyen en el manejo de *Big-Data*; y a diseñadores gráficos y publicistas que apoyen en el plan de marketing[6].

[6] Todos estos planes se encuentran congelados por la pandemia del covid-19, se espera retomarlos con la vuelta a la normalidad.

2. Descripción del proyecto y justificación

2.1 Información del área de la empresa

La Gerencia General de SCM, tiene por objetivo investigar y explicar los patrones de comportamiento de los trabajadores de las empresas cliente. Esto permitirá evaluar la posible construcción e implementación de herramientas que mitiguen el efecto negativo de malas prácticas como ausentismos y atrasos laborales.

El organigrama de la Gerencia General y sus ramificaciones se muestra en el diagrama 1 del punto I.I.III.

2.2 Identificación del problema u oportunidad y su relevancia, con sus efectos y posibles causas

SCM es una empresa que entrega apoyo en el área de Gestión de Operaciones y Recursos Humanos a sus empresas cliente. De esta manera, gestiona y controla (en algunos casos) el tiempo de trabajo mediante dos softwares computacionales que marcan la entrada y salida de los trabajadores: *KRONOS* y *MyPalTime*.

Debido a la aplicación de estos programas se cuenta con una gran cantidad de información individual de los trabajadores, como su antigüedad en la empresa, edad, cargo, marcas en la plataforma, lugar de residencia y penalización por faltar o llegar atrasado, entre otros.

La mayoría de las empresas cliente de SCM se desarrollan en la industria del retail[7] (Adidas, Alvi, Supermercados Peruanos SA y Unilever). En este contexto, se detectan niveles de ausentismo y atrasos bastante grandes. A modo de ejemplo, se consideró una de las empresas clientes de SCM y se detectó que, en una ventana de tiempo de un año, el trabajador promedio falta de manera injustificada un 1,5% y se atrasa un 12,2% de los días del año. Esto se traduce en que el trabajador promedio falta aproximadamente seis días de manera injustificada y llega atrasado 35 días anualmente.

Las cifras descritas anteriormente resultan bastante problemáticas, debido a que estas faltas dejan vacíos en la planificación de horario que deben rellenarse con otros trabajadores o con horas extra. De lo contrario, esto impacta directamente en el nivel de ventas e, incluso, en la calidad de atención al cliente.

En muchas ocasiones las empresas deben tomar medidas drásticas, como descontar el salario a sus trabajadores o despedirlos si incurren en estas malas prácticas reiteradamente. Esta última medida está respaldada por la legislación chilena, en el artículo 160, inciso 3 del Codigo del trabajo (Edicion 2020):

[7] El retail es un sector económico que engloba a las empresas especializadas en la comercialización masiva de productos o servicios uniformes a grandes cantidades de clientes.

"Inciso 3.- No concurrencia del trabajador a sus labores sin causa justificada durante dos días seguidos, dos lunes en el mes o un total de tres días durante igual período de tiempo; asimismo, la falta injustificada, o sin aviso previo de parte del trabajador que tuviere a su cargo una actividad, faena o máquina cuyo abandono o paralización signifique una perturbación grave en la marcha de la obra".

Los trabajadores pueden ser despedidos por ausencia injustificada, sin embargo, la legislación no contempla los atrasos en el inicio de la jornada laboral como una causal de despido o caducidad de contrato. Este tipo de desvinculaciones no son extrañas en las empresas, pues los trabajadores están obligados a cumplir todas las estipulaciones detalladas en sus contratos laborales, como el horario de trabajo. En caso contrario, podría considerarse como un "incumplimiento grave de las obligaciones que impone el contrato de trabajo" lo que sí es una causa legal de despido según el Art. 160 N°7 del Código del Trabajo.

Por otro lado, desde los atrasos y ausencias laborales se desprenden diversos efectos negativos para las empresas:

- **Sensación de derecho o moral negativa:** la literatura afirma que la falta o atraso de una persona puede provocar que el resto del personal comience a pensar que no hay problema al actuar de esta forma (Natter, 2018; Investopedia, 2013; Bell, 2018). Además, según Natter (2018) los trabajadores pueden comenzar a hablar sobre la injusticia de la situación con sus pares. Este tipo de oposiciones podría generar resentimiento, afectando la moral, el rendimiento y la eficiencia.
- **Efecto en cadena:** Según Bell (2018) cuando un empleado llega constantemente tarde, otros miembros del personal pueden comenzar a pensar que a los jefes no les importa esta situación. De esta forma, cuando varios trabajadores empiezan a llegar tarde porque pueden, se produce una actitud desinteresada hacia el lugar de trabajo.
- **Pérdida de respeto hacia el jefe:** este punto puede ser una causa y también un efecto de los atrasos en el lugar de trabajo que, junto a la ausencia laboral, evidencian una falta de respeto importante ante la administración. Esto también podría considerar una falta de liderazgo que suele ocasionar un efecto demoledor para la empresa y el resto del equipo (Natter, 2018).
- **Descontento del cliente:** Según Cucchiella (2014), la degradación de la moral del personal, las ausencias y los atrasos, pueden afectar el área de servicio al cliente. Si no se presenta un empleado encargado de abrir una locación en un tiempo específico, puede ocasionar un disgusto y también pérdida de clientes.
- **Pérdida de productividad:** un empleado que no trabaja cuando debería implica una pérdida de productividad inmediata. Esta interrupción también puede afectar a otros empleados, especialmente a aquellos que dependen del otro para realizar sus tareas. Por ejemplo, un trabajador que llega 12 minutos tarde todos los días de la semana genera una hora de trabajo perdida. Por tanto, si algún colega lo necesita se dificultan los plazos y el cumplimiento de los quehaceres planificados (Va, 2014).
- **Impacto financiero:** empleados que tienen problemas de ausencias y atrasos probablemente sigan siendo pagados, incluso sin haber cumplido su parte del

contrato. Un cheque emitido a cambio de un trabajo incompleto provoca pérdidas económicas para la empresa. Además, reemplazar a los trabajadores actuales puede resultar más costoso que mantenerlos. La compañía deberá compensar las horas perdidas a través de horas extras, o invertir dinero y tiempo extra en el entrenamiento de nuevos empleados (Va, 2014; Investopedia, 2013).

Considerando los puntos señalados, las ausencias y los atrasos son un problema real para las empresas, significando un gran costo económico y social. Por ello, SCM busca entender cuáles son los factores que inciden en el comportamiento de falla al horario laboral, para evaluar el poder estadístico de un modelo que tenga como objetivo predecir estos fenómenos. En caso de que el modelo tenga alto poder estadístico, el área de desarrollo de productos tecnológicos construirá un módulo que se conecte al *output* del modelo, creando una herramienta para que los jefes de tienda ajusten de mejor manera su fuerza laboral.

De esta manera, surgen las siguientes preguntas de investigación:

- ¿Cuáles son los factores que deben ser considerados que causan mayor ausentismo y atraso en los trabajadores?
- ¿Qué tan bien pueden predecir modelos de *Machine Learning* el ausentismo y el atraso laboral?
- ¿Qué algoritmos son más robustos en cuanto a sus métricas de AUC para predecir ausentismo y atraso laboral?

2.3 Identificación de hipótesis y posibles alternativas de solución para resolver el problema

El ausentismo y atraso laboral suelen ser actitudes habituales entre los trabajadores. A pesar de que la experiencia e intuición permite que los empleadores estimen cierto número de ausentismo y atrasos a lo largo del año, estas malas prácticas se traducen en una disminución de la productividad. Además, impactan las finanzas, la moral y otros factores importantes en el desarrollo de la empresa.

Se cree que los trabajadores cometen faltas y atrasos por una variedad de razones. Algunas pueden ser consideradas legitimas, pero en la práctica no todas lo son. Siguiendo a Korkki (2007), Badubi (2017), Cucchiella, Gastaldi y Ranieri (2014), estas causas son las siguientes:

- **Acoso:** empleados que son acosados u hostigados por sus compañeros de trabajo tienen mayores probabilidades de reportarse enfermos para evitar la situación.
- **Cuidado de otros:** empleados pueden verse obligados a faltar al trabajo para quedarse en casa cuidando a los enfermos, niños u otros familiares.
- **Agotamiento, estrés y baja moral:** grandes cargas laborales, sentimientos de no ser apreciados y otras situaciones estresantes, como reuniones largas, presentaciones importantes, fechas de entregas ajustadas y asuntos personales, podrían afectar el horario laboral de una persona, llevándola a elevar su tasa de atrasos y ausencias.

- **Depresión:** según el Instituto de Salud Mental de Estados Unidos, la principal causa del ausentismo es la depresión.
- **Bajo compromiso:** empleados sin sentimiento de compromiso o pertenencia en la empresa tienen más probabilidades de fallar en su responsabilidad con su horario laboral.
- **Lejanía al trabajo:** empleados que viven lejos del trabajo, o que tienen problemas con encontrar una conexión expedita, tienen mayores probabilidades a ser afectados por el tránsito y, en consecuencia, llegar atrasados.
- **Enfermedades y heridas causadas por accidentes:** estas estas son las razones más comunes en los reportes de falta al trabajo. Se ha demostrado que muchas veces los trabajadores fingen y ocupan documentos médicos falsos. Además, durante las temporadas de resfriados y gripe aumenta la tasa de ausentismo de los empleados.
- **Cargo y cantidad de integrantes en la estructura de la organización:** se cree que ciertos cargos son más propensos a faltar que otros. Es muy probable que la cantidad de integrantes en el equipo tenga relación con los ausentismos y atrasos. Por ejemplo, es más probable que en un equipo grande los trabajadores falten o lleguen atrasados, pues es más difícil detectarlo. En cambio, en los equipos más pequeños es más fácil dar cuenta de integrantes que no se presentan.
- **Búsqueda de otro trabajo:** empleados en búsqueda de otra oportunidad laboral pueden reportarse como enfermos para ir a las entrevistas. Puede existir una correlación entre la antigüedad en la empresa y la edad de la persona, ya que los trabajadores más jóvenes tienden a cambiar más frecuentemente de trabajo que las generaciones más antiguas (Linkedin, 2016).
- **Tiempo:** condiciones atmosféricas, como bajas temperaturas y altas precipitaciones, aumentan la probabilidad de que el empleador se enferme. Además, estas condiciones obstaculizan el tránsito afectando los tiempos de traslado de los trabajadores.
- **Festivos o eventos especiales:** cumpleaños, celebraciones patrias o eventos masivos, pueden aumentar la probabilidad que los empleados falten por extender sus vacaciones; o que lleguen tarde tras ingerir bebidas alcohólicas durante instancias de recreación.

Las causas de ausentismo y atrasos pueden ser varias y muy diversas. Con el fin de entender estos fenómenos, se aplican técnicas de extracción de datos para trabajar con las bases de datos recopiladas por *KRONOS*. Se recopilan los marcajes diarios de los empleados (fecha y hora de entrada y salida al trabajo), antigüedad en la empresa, RUT, horas semanales según contrato, edad, cargo, género, nacionalidad, penalización por atraso/ausencia y cantidad de integrantes en el equipo, entre otros.

Frente a lo expuesto, como hipótesis central del trabajo se sostiene que, a partir de la información histórica y las variables características de los trabajadores, se pueden construir modelos para predecir ausentismo y atraso laboral diario a nivel individual.

2.4 Propuesta de valor de las posibles soluciones o impacto del cambio propuesto

Abordar este problema, puede significar una enorme oportunidad de entendimiento de los factores que influyen en el comportamiento de asistencia de los trabajadores. Este análisis permitirá:

- Identificar los factores que afectan los atrasos y ausentismo laboral.
- Evaluar la posibilidad de utilizar el *output* de los modelos como un factor de seguridad para tener en cuenta los días más críticos en cuanto a demanda.
- Evaluar la posibilidad de utilizar el *output* de los modelos como una restricción blanda que penalice en un *Scheduler*, de manera de asignar horarios a los trabajadores minimizando la probabilidad de atraso y ausencia injustificada.
- Determinar probabilidad de ausentismo y atraso de trabajadores a nivel diario.
- Evaluar la posibilidad de crear una herramienta de visualización que describa, bajo cierta probabilidad, cuántos trabajadores faltarán y/o llegarán atrasados, para que el jefe de tienda pueda tomar acciones correctivas en cada caso.

3. Objetivos

3.1 Objetivo general

Desarrollar un modelo de predicción que estime inasistencias y un modelo que estime atrasos de trabajadores de una empresa de la industria de combustibles, para robustecer el sistema de planificación.

3.2 Objetivos específicos

Para llevar a cabo el cumplimiento del objetivo general se declaran los siguientes objetivos específicos:

3.2.1 Identificar las variables relevantes que expliquen el comportamiento de atraso y ausentismo laboral.

3.2.2 Determinar distintos modelos predictivos basados en Machine-Learning para estimar ausentismo injustificado.

3.2.3 Determinar distintos modelos predictivos basados en Machine-Learning para estimar atrasos.

3.2.4 Determinar aplicación de los *outputs* de los modelos como factor para potenciar la resiliencia en la planificación.

3.2.5 Identificar acciones para mitigar los efectos negativos de ausentismos y atrasos laborales.

4. Resultados esperados

Los resultados esperados del trabajo son coherentes a los objetivos planteados previamente. Por tanto, se esperan los siguientes puntos:

4.1 Identificación de cuáles son las variables que más afectan el ausentismo y atraso laboral (condiciones climáticas, variables características del trabajador, marcaje histórico, estacionalidad, festivos y eventos deportivos, entre otros).

4.2 Determinar base de datos con variables que permitan estimar los atrasos y ausentismos laborales.

4.3 Un modelo predictivo que estime probabilidad de ausentismo a nivel individual y diario.

4.4 Un modelo predictivo que estime probabilidad de atraso a nivel individual y diario.

4.5 Determinar usabilidad de los *outputs* de los modelos desarrollados.

4.6 Determinar acciones que ayuden a mitigar los efectos negativos de ausentismos y atrasos laborales.

5 Marco conceptual

5.1 KDD y minería de datos (*data mining*)

KDD, *Knowledge Discovery in Databases*, es un proceso de extracción no trivial de información, con el objetivo de encontrar patrones, asociaciones, anomalías y estructuras a partir de bases de datos masivas (Fayyad, Piatetsky-Shapiro, & Smyth, 1991).

En este proceso se engloban otras áreas, como el análisis estadístico de datos, redes neuronales, técnicas de representación del conocimiento y razonamiento aproximado y basado en casos, entre otros.

La minería de datos o *data mining*, es una parte fundamental del proceso de KDD. En esta etapa en particular, se descubren patrones en el set de datos a partir de algoritmos.

KDD se divide en una secuencia iterativa de pasos lógicos. Estos son los siguientes:

1. **Estudio de la problemática:** Corresponde al entendimiento del problema a estudiar, para identificar los factores que pueden influir en el fenómeno objetivo. Se descubre cuál es el conocimiento previo que se tiene en relación al objeto de estudio, basándose en trabajos y literatura existente. En esta etapa también se definen los límites y el objetivo de la investigación.
2. **Selección de datos:** En esta fase se seleccionan los datos disponibles para realizar el estudio y se integran en una base de datos con el fin de ejecutar el resto del proceso.

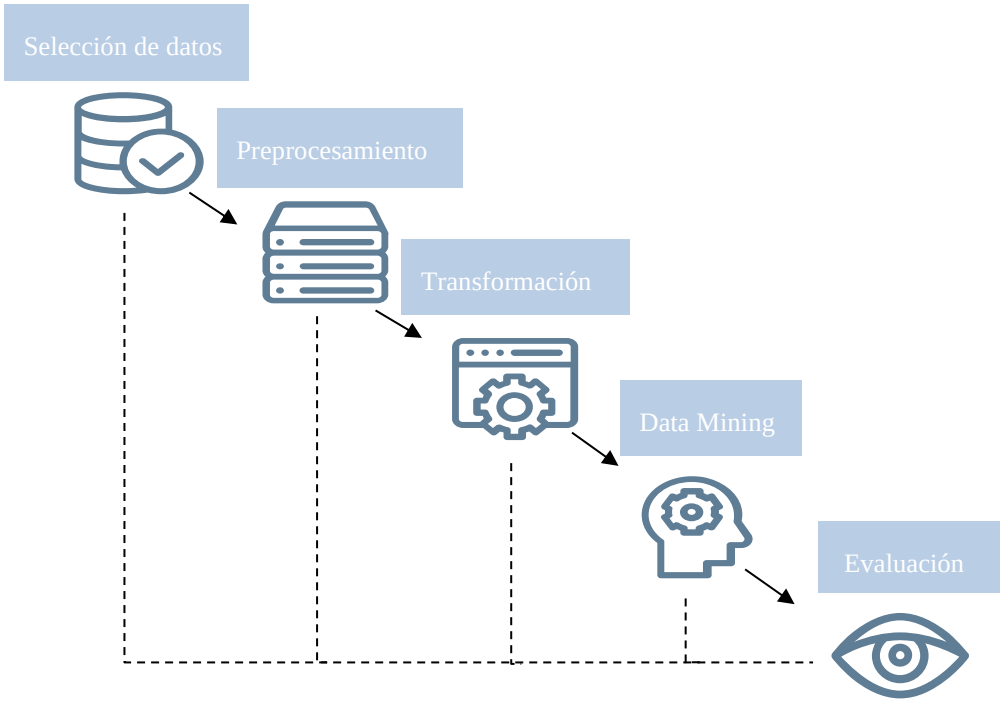


Diagrama 4: Metodología KDD.

3. **Limpieza y preprocesamiento:** Se determina la confiabilidad de la información mediante la realización de tareas que garanticen la utilidad de los datos. Para esto se hace una limpieza de los datos, es decir, tratamiento de *missing values* y *outliers*.

4. **Transformación de los datos:** En esta etapa se mejora la calidad de los datos con transformaciones para reducir dimensionalidad y/o se aplican conversiones a las variables.

5. **Minería de datos o *Data Mining*:** Se analizan los datos y se seleccionan los algoritmos para encontrar patrones en la data. Aquí se decide qué modelos, parámetros y configuración, son los más apropiados según el tipo de datos disponibles y el objetivo de estudio.

6. **Interpretación y evaluación de resultados:** Corresponde a la etapa final del proceso, donde se interpretan los algoritmos y patrones encontrados. En esta etapa se retrocede a pasos anteriores, pues es un proceso iterativo que busca los modelos y variables que mejor se ajusten a la realidad de la información. También se seleccionan las métricas de evaluación de los modelos. Por último, se evalúa el conocimiento descubierto tras la finalización de este proceso, mediante la incorporación en sistemas y/o aplicaciones, o documentándolo para las partes interesadas y trabajos futuros.

5.2 *Machine Learning*

Machine Learning se define como una disciplina del área de estadísticas e inteligencia artificial, cuyo objetivo es otorgar a las computadoras la habilidad de aprender sin ser explícitamente programados para algo en particular. Se basa en la búsqueda de algoritmos que razonan sobre los datos, con el fin de ayudar al investigador a probar sus hipótesis y predicciones futuras con nueva data (ACEC; Accenture; AED, 2019).

5.2.1 Aprendizaje Supervisado y No Supervisado

El aprendizaje supervisado es una técnica para deducir una función mediante datos de entrenamiento. Estos modelos se alimentan de observaciones, con datos de entrada y con los resultados deseados que pueden ser un valor (regresiones), o una etiqueta de clase (clasificación).

En tanto, el aprendizaje no supervisado es un método donde un modelo se ajusta a las observaciones, es decir, no hay un conocimiento a priori. El modelo se alimenta de los datos de entrada y los trata como variables aleatorias, construyendo un modelo de densidad para el conjunto (Géron, 2019).

5.2.2 Métodos de clasificación

En minería de datos, un clasificador es una función que asigna observaciones no etiquetadas a una clase o etiqueta, utilizando las estructuras de datos interna de la observación. Existen diversos métodos de clasificación, basados en distintos campos de investigación.

5.2.2.1 Regresión Logística

La regresión logística modela la probabilidad logarítmica de una variable dependiente binaria Y, utilizando una combinación lineal de covariables independientes X:

$$\log\left(\frac{\boldsymbol{P(Y=1|X)}}{\boldsymbol{1-P(Y=1|X)}}\right)=\boldsymbol{X\beta}$$

Que se reordena para obtener la probabilidad de ocurrencia de Y:

$$\boldsymbol{P(Y=1|X)}=\frac{\boldsymbol{1}}{\boldsymbol{1+e^{(-x\beta)}}}$$

Los coeficientes de la regresión logística, β, representan el efecto de la variable independiente en la probabilidad de ocurrencia de las clases de estudio. Generalmente son ajustados utilizando máxima verosimilitud.

Este modelo se usa como modelo de clasificación cuando se quiere estudiar la probabilidad de un suceso con solo dos posibles resultados. Las ventajas de este clasificador se basan en que es bastante fácil de implementar e interpretar, ofreciendo un buen ajuste y un desempeño similar a algoritmos más complejos (Fernández, 2011).

5.2.2.2 *Naïve Bayes*

Naïve Bayes es una técnica de clasificación basada en el teorema de Bayes, donde se asume independencia entre los atributos predictores, es decir, asume que la presencia de una variable particular en una clase no está relacionada con la presencia de las demás variables (Zhang, 2004).

Este método resulta fácil de construir e interpretar y tiene bajo costo computacional, siendo bastante útil cuando se trabaja con bases de datos masivas. En términos simples, el Teorema de Bayes plantea que la probabilidad posterior de ocurrencia de una clase, dado cierto atributo, es proporcional a la multiplicación de la probabilidad condicional del atributo dada la clase por la probabilidad previa de la clase:

$$\boldsymbol{P(X|c)}=\frac{\boldsymbol{P(c|X)P(c)}}{\boldsymbol{P(X)}}$$

Donde:

6 $P(X|c)$ es la probabilidad del atributo dado la clase.

7 $P(c|X)$ *e*s la verosimilitud, probabilidad de la clase dado el atributo.

8 $P(c)$ es la probabilidad anterior de la clase.

9 $P(X)$ es la probabilidad del atributo.

5.2.2.3 *Gradient Boosting* (*GBM*)

Gradient Boosting es una técnica para problemas de clasificación y regresión. Produce un modelo predictivo en forma de un conjunto de modelos de predicción débiles, típicamente árboles de decisión. De esta forma, construye el modelo de forma escalonada y los generaliza permitiendo la optimización arbitraria de una función de pérdida diferenciable.

Similar a otros modelos de *Boosting*, GBM combina modelos débiles en un modelo fuerte de manera iterativa. El objetivo es enseñarle al modelo $\boldsymbol{F}$ a predecir valores de la forma $\widehat{\boldsymbol{y}} = \boldsymbol{F}(\boldsymbol{x})$, minimizando la función de pérdida. Si se considera el ejemplo de la regresión de mínimos cuadrados, la función a minimizar es la media de los errores cuadráticos:

$$\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}}(\widehat{\boldsymbol{y}_\iota} - \boldsymbol{y_i})^{\mathbf{2}}$$

Donde i son los índices del conjunto de entrenamiento de tamaño n, $\widehat{\boldsymbol{y}_\iota}$ es el valor predicho $\boldsymbol{F}(\boldsymbol{x})$ e $\boldsymbol{y_i}$ el valor real.

Ahora, si se considera un GBM de M etapas, con cada etapa m ($1 \leq m \leq M$), suponiendo un modelo imperfecto $\boldsymbol{F_m}$, con el objetivo de mejorar $\boldsymbol{F_m}$ el algoritmo agrega un nuevo estimador, $\boldsymbol{h_m}(\boldsymbol{x})$:

$$\boldsymbol{F_{m+1}} = \boldsymbol{F_m}(\boldsymbol{x}) + \boldsymbol{h_m}(\boldsymbol{x}) = \boldsymbol{y} \overset{\boldsymbol{equivalente\ a}}{\Longleftrightarrow} \boldsymbol{h_m}(\boldsymbol{x}) = \boldsymbol{y} - \boldsymbol{F_m}(\boldsymbol{x})$$

Por consiguiente, GBM ajusta h según $\boldsymbol{y} - \boldsymbol{F_m}(\boldsymbol{x})$, por lo que cada $\boldsymbol{F_{m+1}}$ intenta corregir los errores de su predecesor $\boldsymbol{F_m}$. Está misma lógica se aplica cuando se opera con otros modelos débiles (Hastie, Tibshirani , & Friedman, 2009).

Importancia de variables en *Gradient Boosting*

Para los árboles de *Gradient Boosting*, las medidas de importancia se basan en el número de veces que una variable es seleccionada para hacer divisiones, ponderadas por la mejora al cuadrado como resultado de cada división y promediada por todos los árboles (Elith, Leathwick, & Hastie, 2008). Lo que se calcula como:

$$\widehat{\boldsymbol{I}_{\boldsymbol{J}}^{\mathbf{2}}}(\boldsymbol{T}) = \sum_{\boldsymbol{t}=\mathbf{1}}^{\boldsymbol{J}-\mathbf{1}} \widehat{\boldsymbol{\iota}_{\boldsymbol{t}}^{\mathbf{2}}}\mathbf{1}(\boldsymbol{v_t} = \boldsymbol{j})$$

Donde la sumatoria es sobre los nodos no terminales $\boldsymbol{t}$ del árbol $\boldsymbol{T}$ de nodo terminal $\boldsymbol{J}$, $\boldsymbol{v_t}$ es la variable de división asociada con el nodo t, y $\widehat{\boldsymbol{\iota}_{\boldsymbol{t}}^{\mathbf{2}}}$ es la mejora empírica correspondiente en el error acuadrado como resultado de la división, definida como $\boldsymbol{i^2}(\boldsymbol{R_l}, \boldsymbol{R_r}) = \frac{\boldsymbol{w_l w_r}}{\boldsymbol{w_l + w_r}}(\overline{\boldsymbol{y_l}} - \overline{\boldsymbol{y_r}})^{\mathbf{2}}$, donde $\overline{\boldsymbol{y_l}}$, $\overline{\boldsymbol{y_r}}$ son las medias de respuesta hija izquierda y derecha respectivamente, y $\boldsymbol{w_l}, \boldsymbol{w_r}$ son las sumas correspondientes de los pesos (Friedman, 2001).

5.2.2.4 *Extreme Gradient Boosting (XGBoost)*

En la literatura reciente, este algoritmo ha estado dominando por la resolución de problemas mediante modelos de *Machine-Learning.* En esta discusión, *XGBoost* es una implementación de árboles de decisión con *Gradient Boosting* diseñado para minimizar la velocidad de ejecución y maximizar el rendimiento.

XGBoost utiliza el hessiano empírico de la función de pérdida para construir los árboles y calcular los pesos de los nodos terminales u hojas. Esto mejora la calidad de los árboles construidos y aumenta la eficiencia y escalabilidad del modelo, ya que no debe resolver problemas de optimización para calcular el peso de las hojas. De esta manera, aporta robustez al modelo al agregar variabilidad en la construcción de cada árbol, mediante el muestreo aleatorio en las variables del modelo, en la partición de los nodos y en la muestra de datos con la que se entrena el modelo.

Evita *over-fitting* al agregar regularización, es decir, tiende a cero los pesos de las hojas (Chen & Carlos , 2016).

5.2.2.5 Árbol de decisión *(Decision Tree)*

Los árboles de decisión son modelos de clasificación de observaciones, ordenándolas según las características de cada una de éstas. Se componen por un nodo principal o raíz; nodos internos con arcos de entrada y salida; y nodos terminales u hojas que no tienen arcos salientes. Estos últimos representan una característica de la instancia que será clasificada, mientras que cada rama o arco representa el valor que cada nodo puede tomar.

Las observaciones son clasificadas en clases predefinidas, subdividiendo el árbol secuencialmente. Se comienza por el nodo principal, asignando cada observación a una clase de acuerdo a su nodo final. El diagrama 5 es un ejemplo de un árbol de decisión para el conjunto de entrenamiento mostrado en la tabla 2.

Atributo 1	Atributo 2	Atributo 3	Atributo 4	…	Atributo n	Clase
a1	a2	a3	a4	…	an	1
a1	a2	d3	a4	…	bn	0
a1	a2	b3	b4	…	cn	2
b1	c2	c3	c4	…	dn	1
c1	c2	a3	a4	…	en	0
c1	a2	d3	c4	…	fn	0
a1	c2	a3	d4	…	gn	1
c1	b2	b3	d4	…	dn	2

Tabla 2: Conjunto de entrenamiento del árbol de decisión.

En el nodo principal, o nodo raíz, se considera el atributo que mejor divide el conjunto de entrenamiento. En el caso del ejemplo descrito en el diagrama 5, éste corresponde al atributo 1.

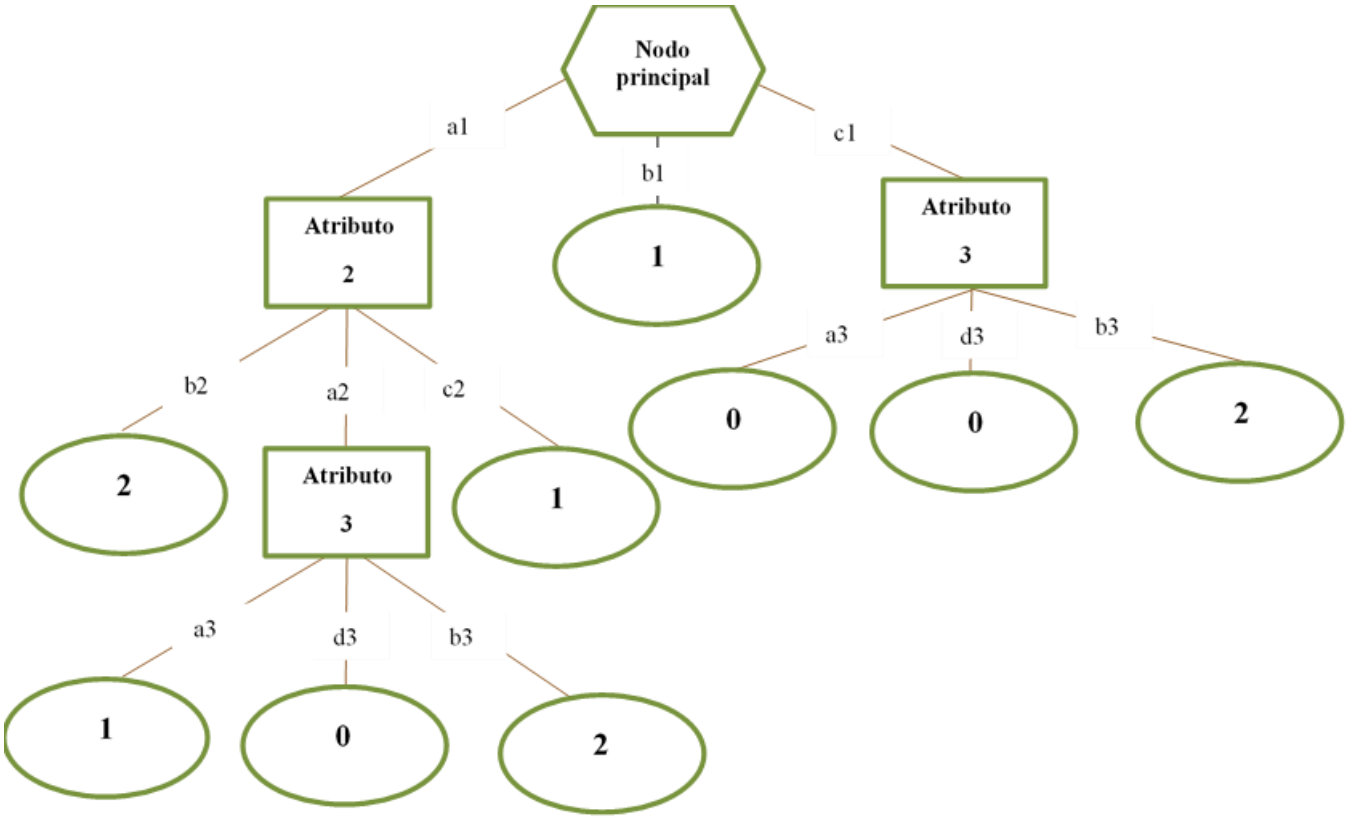


Diagrama 5: Árbol de decisión.

Los árboles de decisión representan una herramienta altamente efectiva en diversas áreas para el reconocimiento de patrones. Se caracterizan por su alta versatilidad, permitiendo su uso en la resolución de problemas de clasificación, regresión, *clustering* y selección de atributos. Asimismo, son altamente interpretables (su misma estructura permite entender el problema), son flexibles al admitir diferentes tipos de datos (continuos, binarios y categóricos), y se adaptan fácilmente en el procesamiento de datos con *missing values* y *outliers* (Breiman, Friedman, Olshen, & Stone, 1984).

5.2.2.6 Bosque aleatorio *(Random Forest)*

Esta modelo de clasificación, como su nombre lo indica, es una combinación de árboles de decisión. Cada uno de los árboles depende de los valores de un vector seleccionado aleatoriamente, con la misma distribución para todos los árboles en el bosque.

En el entrenamiento, esta técnica utiliza una muestra del conjunto de entrenamiento original, buscando sólo a lo largo de un subconjunto de los atributos de entrada escogido aleatoriamente para determinar la división en cada nodo. Para la clasificación, cada árbol del bosque entrega una respuesta respecto a qué clase debiera pertenecer cada instancia que entró como input. La decisión definitiva está determinada por la respuesta más frecuente entre los árboles individuales.

La variación de los árboles de decisión puede manejar datos de mayor dimensionalidad y utilizar una mayor cantidad de árboles en conjunto. Esto, combinado con la minimización de error que provoca la aleatorización de los árboles de decisión escogidos, permite que los resultados obtenidos sean comparables con otras técnicas más sofisticadas, pero siendo computacionalmente más ligero (Breiman, 2001).

Importancia de variables en *Random Forest*

En los árboles de decisión, cada nodo es una condición de separación según una característica, de tal manera que valores similares de la variable dependiente terminen en un mismo conjunto después de la separación. La condición está basada en la impureza, donde en caso de un problema de clasificación corresponde a impureza de Gini o ganancia de información (entropía).

Al entrenar un árbol se calcula cuánto contribuye cada característica a disminuir la impureza ponderada. Para los *Random Forest*, este cálculo se hace promediando la disminución de la impureza a lo largo de los árboles.

Las ventajas de este enfoque es que es relativamente fácil de implementar, no requiere alto poder computacional y su interpretación es directa. No obstante, este enfoque esta sesgado ya que tiende a inflar la importancia de las variables continuas y variables categóricas con alta cardinalidad.

Impureza de Gini (*Gini Impurity*)

La impureza de Gini es la probabilidad de clasificar de manera incorrecta un elemento seleccionado de manera aleatoria en el conjunto de datos si se etiquetara aleatoriamente de acuerdo con la distribución de la clase en el dataset. Se calcula como:

$$\boldsymbol{C = \sum_{i=1}^{C} p(i) * (1 - p(i))}$$

Donde $\boldsymbol{C}$ corresponde al número de clases y $\boldsymbol{p(i)}$ la probabilidad de seleccionar de manera aleatoria el elemento de la clase i.

Al entrenar un árbol de decisión, la mejor separación es elegida maximizando la ganancia de Gini (*Gini Gain*), que se calcula restando las impurezas ponderadas de las ramas de la impureza original (Louppe, 2014).

Ganancia de información (Entropía)

En el contexto de árboles de decisión, la entropía se puede definir como cuanta varianza tiene la data. Se calcula según:

$$\boldsymbol{E = -\sum_{i=1}^{C} p(i) * \log_2 p(i)}$$

Donde $\boldsymbol{C}$ corresponde al número de clases y $\boldsymbol{p(i)}$ la probabilidad de seleccionar de manera aleatoria el elemento de la clase i.

Posteriormente, la ganancia de información se calcula para cada división restando las entropías ponderadas de cada rama de la entropía original.

Al entrenar un árbol de decisión con estas métricas, la mejor división se elige maximizando la ganancia de información (Louppe, 2014).

5.2.3 *Tunning* de hiperparámetros (*Hyperparameter tunning*)

Los hiperparámetros son partes importantes de los modelos de *Machine Learning* y la correcta configuración de estos puede terminar en mejoras considerables. Algunos ejemplos de hiperparámetros pueden ser la cantidad de nodos hijos de un árbol de decisión o el número de estimadores de un *Random Forest*.

En este contexto, el *tunning* de hiperparámetros es la selección de un set óptimo (local o global) de hiperparámetros para un algoritmo de *Machine Learning*. La mejor forma de entender este proceso es la configuración de un algoritmo para optimizar su desempeño; similar a cuando se ajustan las perillas de una radio para encontrar la mejor señal.

Encontrar los mejores hiperparámetros es imposible de determinar usualmente, debido a la gran cantidad de combinaciones posibles y el tiempo que demanda cada una de estas.

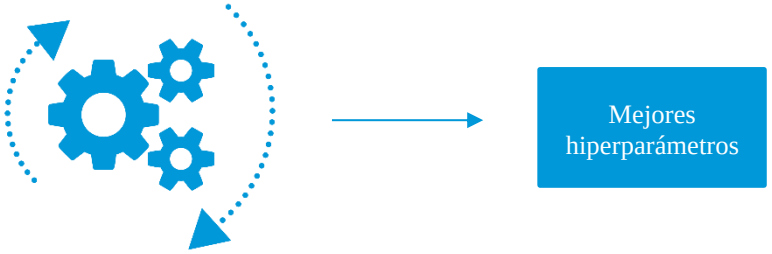


Diagrama 6: Hyperparameter tunning.

En este trabajo se utilizan dos técnicas de optimización de hiperparámetros: *Random Search with Cross Validation* y *Grid Search with Cross Validation.*

5.2.3.1 *Random Search with Cross Validation*

El *Random Search* es un enfoque donde se prepara un set de hiperparámetros candidatos del que aleatoriamente se seleccionan subconjuntos de hiperparámetros para luego ejecutar *K-Fold Cross Validation*, este proceso se repite un número determinado de veces.

Las ventaja de este método es que permite controlar el tiempo computacional al elegir el número de búsquedas de parámetros. La desventaja es que, si el espacio de hiperparámetros

es muy grande, puede que existan ciertos parámetros no explorados suficientemente (Bengio & Bergstra, 2012).

5.2.3.2 *Grid Search with Cross Validation*

El *Grid Search* es un enfoque donde se elige un set de hiperparámetros candidatos y se hace una búsqueda exhaustiva con todas las combinaciones posibles, donde para cada una de estas combinaciones se aplica *K-Fold Cross Validation.*

La ventaja de este método es que cubre todas las combinaciones posibles de hiperparámetros y alcanza el óptimo global si se selecciona una muestra representativa de hiperparámetros. La desventaja es que el tiempo computacional es considerablemente más grande que el *Random Search* y a veces no vale la pena, ya que la mayoría de las veces el *Random Search* alcanza un buen set de hiperparámetros (Bengio & Bergstra, 2012).

5.2.4 Métricas de evaluación de los modelos

Para evaluar el desempeño de un modelo se deben considerar tres aspectos fundamentales:

1. **Conjuntos de entrenamiento y testeo del modelo:** Para construir estos tipos de conjuntos, generalmente se dividen las observaciones de manera aleatoria en dos partes. Una para el entrenamiento del modelo, donde se definen los parámetros del clasificador, mientras que la otra, para validar el modelo entrenado y estimar su error de generalización, con el objetivo de minimizar este error evitando el sobreajuste (*over-fitting*).
2. **Error:** Corresponde a la relación entre las observaciones mal clasificadas sobre el total de observaciones del conjunto de validación.
3. **Métricas de desempeño:** Son medidas usadas para calcular la calidad del modelo y el error del modelo. Existe una gran variedad y según la problemática a estudiar, ciertas métricas resultan preferibles a otras. Más adelante se describen las medidas usadas en este estudio según la referencia de Géron (2019).

5.2.4.1 Entrenamiento y testeo del modelo

Con el objetivo de testear y validar los modelos, se separa la base de datos en una base de entrenamiento y en una base de testeo. Usualmente se "baraja" de manera aleatoria la base de datos y se separa en 80% para entrenar el modelo y 20% para testearlo.

Base de entrenamiento	Base de testeo

Tabla 3: Representación de base de datos de entrenamiento y testeo.

Al dividir la base de datos, se debe asegurar que ambas partes (entrenamiento y testeo) sean conjuntos representativos. El problema es que al realizar la partición de forma aleatoria es muy probable que las clases de ocurrencia no ocurran de forma equitativa en ambas bases,

entonces, existe un *trade-off* entre el número de observaciones para entrenar el modelo y el número de ocurrencias de las clases en cada parte.

Lo anterior, puede traer ciertas desventajas, pues es difícil que ambas particiones sean efectivamente representativas de la muestra inicial. Esto provoca que los resultados obtenidos puedan variar significativamente, dependiendo de cómo se seleccionan las particiones y reflejando una gran varianza en las métricas de evaluación.

Para solucionarlo se usa el método de validación cruzada de *k folds* (*K-Fold cross-validation*). Se caracteriza por ser iterativo, pues la muestra total se divide en k subconjuntos distintos y disjuntos. El modelo se entrena con k-1 subconjuntos y se evalúa con el subconjunto que se deja fuera del entrenamiento. Este proceso se repite k veces, generando k modelos.

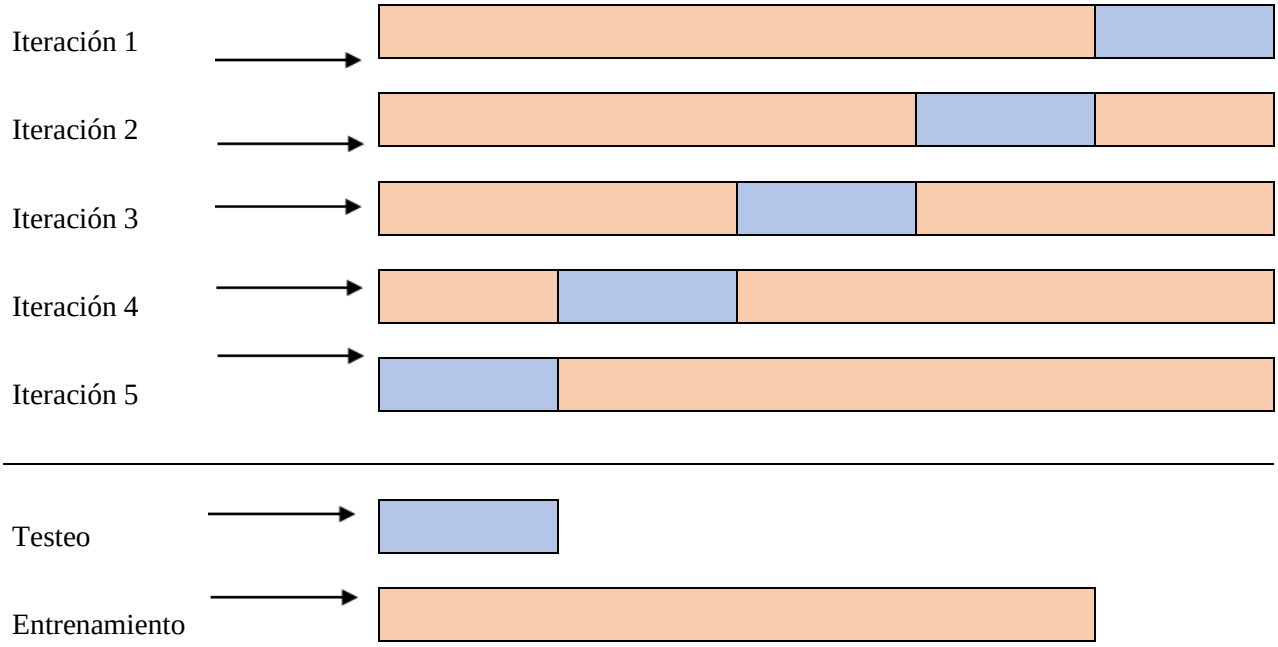


Diagrama 7: Diagrama de división de conjuntos k-fold cross-validation.

De esta manera, el desempeño se calcula como la media aritmética de los valores obtenidos en cada uno de los k modelos. La idea principal de este método es que el desempeño promedio de los k modelos generados, en cada una de las iteraciones, es un buen estimador del desempeño del modelo original al evaluarlo en un set de datos nuevos (Géron, 2019).

5.2.4.2 Matriz de confusión

La matriz de confusión de un modelo muestra información sobre la clasificación real del set de testeo y la predicha por el modelo.

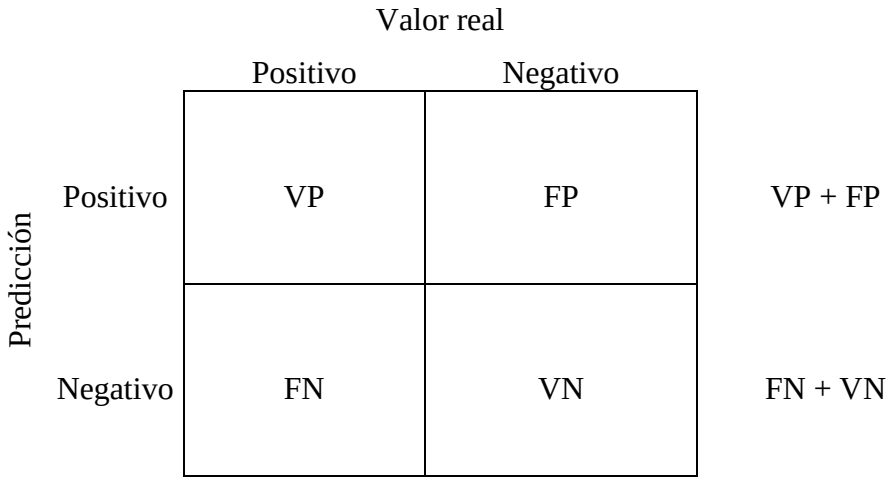


Diagrama 8: Matriz de confusión

La matriz de confusión contiene la siguiente información:

1. **Verdaderos Positivos (VP):** Número de casos en los que el clasificador predijo correctamente la etiqueta positiva.
2. **Falsos Positivos (FP):** Número de casos donde el clasificador predijo incorrectamente la etiqueta positiva.
3. **Falsos Negativos (FN):** Número de casos donde el clasificador predijo incorrectamente la etiqueta negativa.
4. **Verdaderos Negativos (VN):** Número de casos donde el clasificador predijo correctamente la etiqueta negativa.

A partir de estas medidas se construyen las métricas que se describen en los puntos posteriores.

5.2.3.2.1 Métrica de exactitud (*accuracy*)

Esta métrica indica el número de observaciones clasificadas correctamente (VP y VN), en comparación al número total de observaciones clasificados en cualquier clase (VP, VN, FP y FN).

$$\boldsymbol{accuracy = \frac{VP + VN}{VP + VN + FP + FN}}$$

Cabe destacar que esta métrica tiene la desventaja de ser poco representativa cuando las clases están muy desequilibradas, pues algunas tendrán muchos más elementos que otras.

5.2.3.2.2 Métrica de especificidad (*specificity*)

Esta métrica indica la probabilidad de obtener un resultado negativo cuando la observación está realmente etiquetada como negativa (VN).

$$\boldsymbol{Specifity = \frac{VN}{VN + FP}}$$

5.2.3.2.3 Métrica de sensibilidad (*recall o sensitivity*)

Esta métrica corresponde a la probabilidad de obtener un resultado positivo cuando la observación está realmente etiquetada como positiva (VP).

$$\boldsymbol{recall = \frac{VP}{VP + FN}}$$

5.2.3.2.4 Métrica de precisión (*precision*)

Esta métrica representa el número de verdaderos positivos que son realmente positivos, en comparación con el número total de valores positivos predichos.

$$\boldsymbol{precision = \frac{VP}{VP + FP}}$$

5.2.3.2.5 Puntuación F1 (*F1-Score*)

Esta métrica corresponde a la media harmónica entre la precisión y la sensibilidad, donde la mejor puntuación posible es 1 y la peor es 0.

$$\boldsymbol{F_1 = 2 * \frac{precision * recall}{precision + recall}}$$

5.2.3.2.6 Curva *ROC* (*Receiver Operating Characteristics*)

El análisis de la curva *ROC* es una metodología para evaluar, comparar y seleccionar clasificadores en base a su desempeño. Se define como un gráfico en dos dimensiones, que representan la sensibilidad en el eje Y, en función de la especificidad en el eje X. De esta manera, el desempeño de un clasificador, representado por su sensibilidad y especificidad, es simbolizado por un solo punto en el gráfico *ROC*.

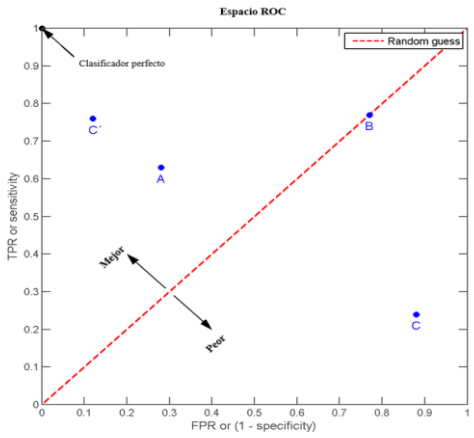


Diagrama 9: Espacio ROC

El punto de origen del espacio *ROC* (0,0) corresponde a un clasificador que nunca predice una observación positiva, mientras que tampoco clasifica erróneamente una observación negativa como positiva (*sensitiviy*=0, *specifity*=0). El otro extremo, en el punto (1,1) es un clasificador de todas las observaciones como positivas (produciendo un número grande de falsos positivos) y ninguna como negativa.

Es fácil notar que un clasificador perfecto se situaría en el puto (0,1), donde la sensibilidad y la especificidad son 1. Si bien no es realista esperar obtener un clasificador perfecto, es un punto al cual se quiere acercar para encontrar el mejor clasificador.

La diagonal de línea punteada de color rojo representa un clasificador aleatorio, donde los puntos que están por arriba representan resultados de clasificación mejor que el azar, mientras los puntos que están por debajo simbolizan a los peores clasificadores.

5.2.3.2.7 *AUC* (*Area Under the Curve*)

AUC corresponde al "área bajo la curva *ROC*". Esto significa que el *AUC* mide toda el área bidimensional por debajo de la curva *ROC* completa de (0,0) a (1,1).

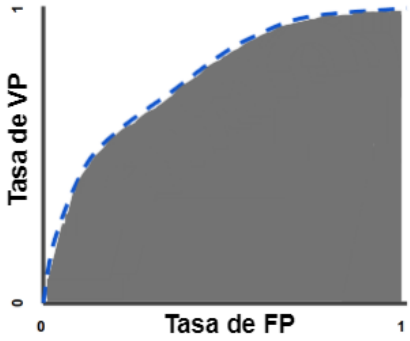


Diagrama 10: AUC

De esta forma, un clasificador perfecto tendría un *AUC* de 1.0 y un clasificador aleatorio tendría *AUC* 0.5.

6 Revisión de la literatura

Para encontrar y entender modelos de problemas similares al de este estudio, se estudiaron diversos informes que se describen a continuación:

En Martiniano et al. (2012) desarrollaron una red *neuro-fuzzy* usando un *multilayer perception* con un algoritmo de propagación de error para predecir ausentismo. En este estudio, recolectaron observaciones de 3 años de una compañía de Courier brasileña y después de tabular y filtrar la data se quedaron con gran parte de la base de datos con datos nulos. A pesar de esto, procedieron con el estudio y clasificaron las ausencias justificadas según la Clasificación Internacional de Enfermedades *(International Classification of Diseases, ICD*) en 21 categorías distintas con la intención de obtener el impacto de cada una de estas ausencias en el ausentismo injustificado. En este informe, descubrieron que solo seis categorías de la base de datos son la razón del 78,65% de las ausencias de la ICD. Identificaron que este modelo sería útil para la toma de decisiones de la empresa, por lo que decidieron aumentar su base de datos para hacer otro estudio que los resultados aún no han sido publicados.

En Taylor et al. (2013), recolectaron observaciones de 4 años de una empresa de ferrocarriles, usando variables climáticas, feriados, estacionales y características para predecir el ausentismo. En este estudio, usaron una regresión de *Poisson* para identificar los factores que influyen estadísticamente en el número de ausencias injustificadas. A pesar de que con el modelo fueron capaces de encontrar ciertos factores que inciden de manera estadísticamente significativa en el ausentismo, el modelo final registró un Pseudo R^2 de solo 0,0179.

Nunung et al. (2014) desarrolló un clasificador basado en un Árbol de Decisión para encontrar las características especiales de grupos de empleados que se ausentan más frecuentemente en el lugar de trabajo. Recolectaron 14.400 observaciones de marcaje de empleados desde 2009 hasta 2011 de una empresa privada de Indonesia. Usaron las reglas del departamento de recursos humanos para clasificar las observaciones en tres categorías: frecuentemente ausente, raramente ausente y frecuentemente presente. Posteriormente, entrenaron y testearon el árbol con un 80% y 20% respectivamente, donde descubrieron que las empleadas de 33-39 años con al menos 3 hijos tienden a faltar más que cualquier empleado con otras características. El modelo tuvo un *accuracy* de 95%, pero las demás métricas bastante deficientes.

En el informe de Gayathri (2018) utilizaron la misma base de datos que en Martiniano et al. (2012) y crearon un modelo de clasificación para predecir horas de ausentismo, dado ausentismo, en períodos cortos y largos. En este estudio, aplicaron *Naive Bayes, Multilayer Perceptron* y clasificadores *J48*, donde concluyeron que el *Multilayer Perceptron* entrega mejores resultados con un error de 0.1%. En este artículo usaron *SMOTE* como técnica de *Oversampling*, sin embargo, fallaron en aplicar esta técnica, pues la aplicaron a la base de entrenamiento y de testing, por lo que sus resultados no representan el desempeño real de sus modelos, sin embargo, sirven como línea base para no cometer los mismos errores metodológicos.

En Ferreira et al. (2018) aplicaron Redes Neuronales Artificiales (ANN) para predecir atraso en el trabajo. Para esto, usaron una base de solo 2243 observaciones y 38 atributos. Posteriormente, mediante el uso de ANN, lograron un error mínimo de 1,44 minutos y un error máximo de 8,8 días por trabajador, lo que es bastante aceptable considerando la baja cantidad de observaciones.

Debido a los bajos resultados de los pocos estudios que existen sobre el fenómeno de estudio, en esta tesis se usarán distintos modelos para estimar el ausentismo y atrasos laborales, regresión logística, *Naive Bayes*, *Gradient Boosting*, *eXtreme Gradient Boosting*, *Decision Tree* y *Random Forest*.

7 Metodología

La metodología a utilizar se basa en KDD (*Knowledge Discovery in Databases*), ya que el problema a abordar requiere del estudio de bases de datos para encontrar un modelo válido e interpretable, que describa y prediga patrones de conducta de los trabajadores. Esta metodología también permite dividir el problema en diversos pasos lógicos con el fin de estructurar el problema de manera óptima y así encontrar un modelo de minería que se ajuste al objeto de estudio:

1. **Estudio de la problemática:** En primer lugar, se debe entender la problemática, es decir, comprender aquellos factores que influyen en el comportamiento de falta y atraso laboral de los trabajadores. También se estudia cómo el ausentismo y atraso laboral afectan a las empresas en materia de costos económicos, productividad y cultura empresarial. En esta etapa se realiza una investigación sobre el estado del arte de los modelos estadísticos que pueden servir para predecir el fenómeno objetivo.
2. **Selección de datos:** Se trabaja con las bases de datos de *KRONOS* para extraer los marcajes históricos de una de las empresas cliente de SCM. Posteriormente, se cruzan estos marcajes con otras variables que permiten entender el comportamiento de los trabajadores. Por ejemplo: edad, cargo, penalización de sueldo por atraso e inasistencia, clima (temperatura y precipitaciones), tiempos de traslado hacia el lugar de trabajo, entre otros. Posteriormente, se identifican y seleccionan aquellas variables que realmente influyen en el atraso y ausentismo laboral.
3. **Limpieza y preprocesamiento:** En esta etapa se determina la confiabilidad de la información, por lo que se tratan los valores perdidos y se eliminan los valores atípicos. También se descartan ciertos trabajadores que no tienen información suficiente para construir los modelos predictivos.
4. **Transformación de los datos:** Se mejora la calidad de los datos usando transformaciones que permiten una mejor interacción entre variables y modelos. Por ejemplo, eliminación de variables y reducción de dimensionalidad de ciertas variables, entre otros.
5. **Minería de datos:** En esta fase se procede a trabajar con herramientas computacionales para la minería de datos, se generan clústers de trabajadores mediante aprendizaje no supervisado y se crean distintos modelos de predicción de ausentismo y atraso laboral mediante aprendizaje supervisado de *Machine Learning*.
6. **Evaluación de resultados:** En esta etapa final se realiza un análisis completo de los resultados de la etapa anterior, para posteriormente elegir los modelos que mejor se ajustan al fenómeno a estimar.

Cabe destacar que se escoge este enfoque por la flexibilidad a posibles cambios e integración de nuevas variables que permiten iterar con el objetivo de alcanzar un mejor modelo de estimación.

7.1 Herramientas

Se usa *Python 3.7* y una de sus distribuciones libre, *Anaconda*, debido a la facilidad de su despliegue, administración e instalación de paquetes y librerías de software (*conda*[8]).

Para ejecutar los scripts formulados, se usa el entorno de desarrollo *Spyder*, dado que funciona como multiplataforma a través de *Anaconda*. *Spyder* es una plataforma altamente robusta e interactiva para programación científica. Además, permite ejecutar los códigos en varios *Kernels*[9] distintos y una ventana adicional para visualizar los resultados.

Las librerías utilizadas para el proyecto son: *pandas, scipy, numpy, matplotlib, math, datetime, scikit-learn, seaborn, ast, imblearn, xgboost, graphviz, skopt* y *functools*.

[8] Sistema de gestión de paquetes de *Anaconda*.

[9] Núcleo donde se ejecuta un algoritmo.

8 Alcances

Los alcances del trabajo a realizar se pueden separar en varios aspectos, los cuales permitirán manejar el problema de forma más dinámica sin perder el foco en el proceso. En primer lugar, se debe considerar el alcance del trabajo en términos de tiempo, pues debido a la limitación de éste, es una limitación en esta tesis. El desarrollo del modelo, la obtención de los resultados y el análisis de los beneficios comprenden un poco más de un semestre académico.

En segundo lugar, dentro de los intereses de SCM, está el desarrollo de un módulo extra denominado Predictor de Ausencias/Atrasos (PA). El *output* del modelo a desarrollar se conectará directamente a este módulo para ofrecer un nuevo servicio diferenciador a las empresas clientes de SCM, instalándose como herramienta para los jefes de tienda, al entregar la probabilidad de falta o de atraso un día antes, con el fin de tomar medidas con anticipación.

Asimismo, en este trabajo se desarrolla el modelo con foco en la escalabilidad, pues en primera instancia será adaptado solo para una empresa de SCM, que es una empresa idónea para usarla como piloto del modelo a producir, pues la buena relación con el cliente facilita el uso de sus bases de datos, que cuentan con una gran cantidad de datos y pocos *outliers*. Se espera que en un futuro este modelo sea adaptable a cualquier otra empresa cliente.

Por último, existen ciertas limitaciones en el alcance de este trabajo. La “crisis social”, la pandemia originada por el CoVid-19, la existencia de limitada información característica de los trabajadores de las bases de datos disponibles, la complejidad de predicción del fenómeno propuesto y las restricciones de presupuesto para hacer consultas por medio de otras APIs (*Google Weather, Google Maps, Traffic Scrapping*, entre otros). Sin embargo, a futuro se buscará aumentar la complejidad y poder estadístico del modelo, a medida que existan más empresas interesadas, solicitándoles más información característica de los trabajadores para agregarla al modelo y así entender mejor su comportamiento.

9 Desarrollo metodológico

9.1 Selección de datos

Para este proyecto se usan datos de una empresa cliente de SCM de la industria del petróleo. Las bases de datos se describen a continuación:

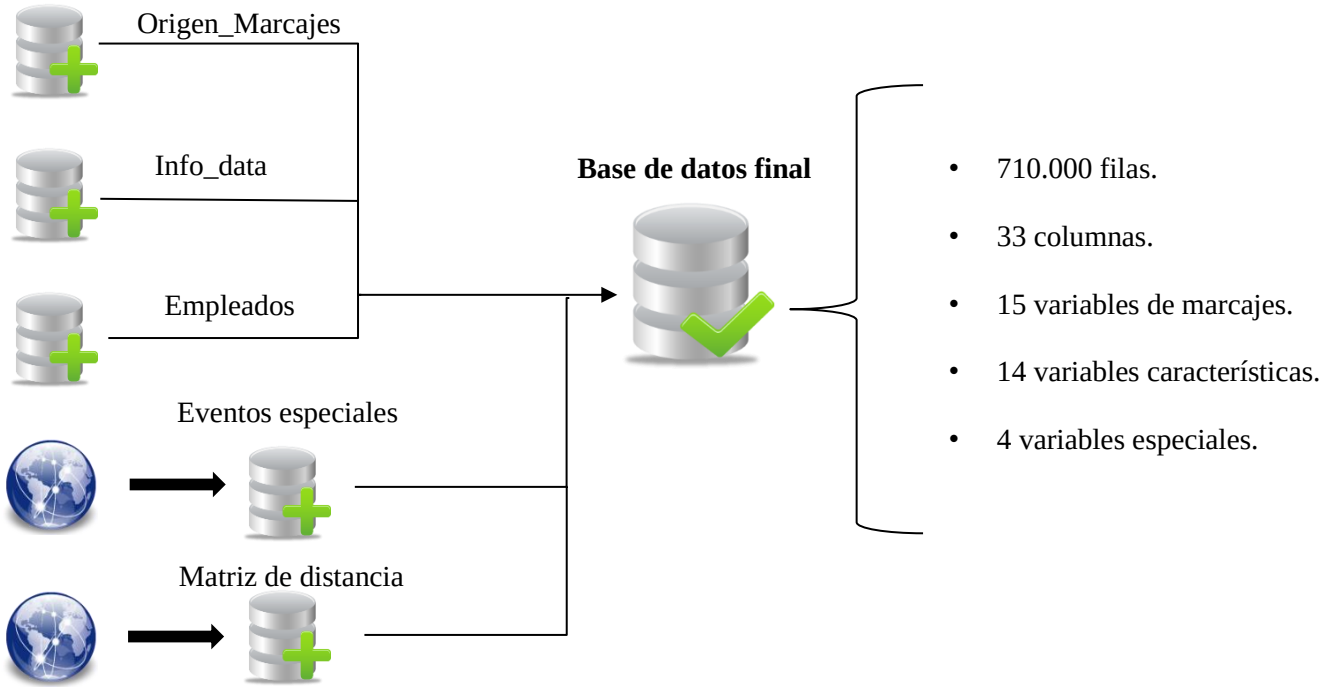


Diagrama 11: Selección y extracción de bases de datos.

1. **Origen_Marcajes:** Esta base contiene los marcajes históricos de los trabajadores desde el 1 de enero de 2017 hasta el 2 de abril de 2020. Consiste en 1.226.720 filas, con cinco columnas: ID empleado, fecha de marcaje, hora de marcaje, si la marca fue modificada (eliminada o actualizada) y el nombre del terminal físico donde marcó. En muchos casos, el nombre del terminal físico no aparece, por lo que se imputa la moda de los datos para ese trabajador.

   En esta base de datos, 4.435 son datos duplicados (0,3%) y 104.435 (8,4%) corresponden a marcas eliminadas, de los cuales cerca del 50% corresponden a marcaciones eliminadas o actualizadas desde el 18 de octubre de 2019 hasta el 02 de abril de 2020. Esto puede explicarse por el estallido social[10] y el Covid-19[11], por lo

[10] Corresponden a una serie de manifestaciones y disturbios originados en Santiago y propagados a todas las regiones de Chile

[11] Virus que causa una pandemia en todo el mundo, registrándose más de 3 millones de casos a la fecha (25/04/2020).

que posteriormente se intenta rescatar la primera marca para la mayoría de estos casos hasta el 11 de octubre de 2019.

2. **Info_data:** Esta base de datos contiene el turno de todos los trabajadores junto con el motivo de ausentismo desde el 1 de enero de 2017 hasta el 2 de abril de 2020. Consiste en 697.613 filas y cuatro columnas: ID, fecha de turno, entrada y salida de turno y motivo de ausencia. Se detectan y eliminan 12.080 datos duplicados y nulos, es decir, el 1,7% de la base de datos.

   Las razones de ausentismo se describen en la A.1 del apéndice.

3. **Empleados:** Esta base de datos contiene el ID de todos los trabajadores que ha tenido la empresa, desde mayo del 2000 hasta marzo del 2020. Consiste en 3.133 filas, con 40 atributos. Entre ellos destaca: ID del empleado, nombre, Rut, estado civil, dirección, código AFP y tipo de contrato, entre otros. Cabe destacar que la gran mayoría de estos atributos están mal ingresados y no reflejan ningún tipo de información concreta.

   En el apéndice A, A.1, se exhiben los datos de un trabajador modificando su nombre, Rut e identificador para resguardar su privacidad.

4. **Eventos especiales:** Base de datos de autoría propia con tres hojas: Partidos (fecha, hora, lugar del partido y resultado), Manifestaciones (fecha, hora y lugar) y Feriados obligatorios (fecha), desde el 17 de enero de 2017 hasta el 11 de octubre de 2019.

5. **Matriz de distancia:** A través de la base de datos Origen_Marcaje, se determina la localización geográfica de los terminales y, posteriormente, desde una fuente de datos externa se extraen las direcciones de los empleados a través del Rut y se calcula mediante una API de *Google Maps*[12] la distancia hasta el terminal de marcaje.

9.1.1 Base de datos final

En primer lugar, debido al estallido social y a la pandemia causada por el COVID-19, las bases de datos se cortan una semana antes del estallido social (11 de octubre del 2019), lo que deja las bases de datos Origen_Marcajes e Info_data con 1.064.045 y 591.456 filas respectivamente.

Posteriormente, se junta la información de todas las bases de datos mencionadas y se hace una primera normalización[13] a las variables mencionadas anteriormente, dando origen a una base de datos con 714.838 filas y 33 columnas, que se describen a continuación:

[12] https://developers.google.com/maps/documentation/distance-matrix/intro?hl=es

[13] Se agrupan ciertos datos para un mismo atributo, por ejemplo: Soltero (a), soltera, soltero(a), entre otros.

Variables de marcajes:

Variable	Descripción	% datos nulos	Promedio	Desviación estándar	Mínimo	Máximo
ID	Identificador del empleado	-	-	-	-	-
Fecha	Fecha de la observación	-	-	-	01-01-2017	16/18/2019
PunchIn	Fecha y hora de marca de entrada	3,25%	-	-	-	-
PunchOut	Fecha y hora de marca de salida	3,30%	-	-	-	-
BreakIn	Fecha y hora de entrada a colación	86,20%	-	-	-	-
BreakOut	Fecha y hora de salida de colación	86,20%	-	-	-	-
Fecha_retiro	Fecha de término de contrato/retiro del empleado	-	-	-	-	-
Turno	Lista con el turno del empleado (entrada y salida)	5%	-	-	-	-
D_Atraso	Si el empleado se atrasó o no (binaria)	-	0,122	0,318	0	1
AusenciaInjustificada	Si el empleado se ausento injustificadamente o no (binaria)	-	0,015	0,127	0	1
AusenciaJustificada	Si el empleado se ausento de manera justificada (binaria)	-	0,137	0,343	0	1
Paycode (justificación)	Justificación de la ausencia del empleado	-	-	-	-	-
Reloj	Nombre del reloj donde marca el empleado	30%	-	-	-	-
Horas_trabajadas	Horas trabajadas por el empleado	0,10%	7,76	2,1	2,46	16,98
Min_Atraso	Minutos de atraso del empleado	3,25%	3,642	18.694	0	190

Tabla 4: Variables de marcaje.

Variables características:

Variable	Descripción	% datos nulos	Promedio	Desviación estándar	Mínimo	Máximo
Sexo	Género del empleado	0,50%	-	-	-	-
Cumpleaños	Fecha de cumpleaños	-	-	-	-	-
Nación	Si es chileno o no (binaria)	0.10%	-	-	-	-
Estado_Civil	Estado civil del empleado	-	-	-	-	-
Jornada laboral	Tipo de jornada laboral	-	-	-	-	-
Jefe	Jefe del empleado	-	-	-	-	-
Tipo_turno	Tipo de turno (mañana, tarde, noche)	-	-	-	-	-
Tipo_Contrato	Tipo de contrato (45 horas, articulo 22, entre otros)	0,1%	-	-	-	-
Forma_Pago	Categórica nominal (cheque o abono)	-	-	-	-	-
Edad	Edad del empleado	-	39,49	11328	18	67
Meses_contrata	Meses de antigüedad en la empresa	-	69	110,2	3	503
Distancia	Distancia del domicilio al reloj de marcaje	15%	14.823	11.34	349	49.504
Temperatura	Temperatura mínima registrada en el día	0,1%	8,17	4,43	-13	23
Precipitaciones	Precipitaciones acumuladas del día	0,1%	0,14	0,57	0	9,27

Tabla 5: Variables características.

Variables especiales:

Variable	Descripción	% datos nulos
Feriado	*Dummy* si la observación es en un día feriado o no	-
Partido	*Dummy* si hay un partido en la fecha	-
Movilización	*Dummy* si hay movilizaciones que pasan por el camino del domicilio del trabajador al trabajo	-
Cumpleaños	*Dummy* si está de cumpleaños	-

Tabla 6: Variables especiales

9.2 Limpieza y preprocesamiento de los datos

En esta etapa los datos recolectados se preparan adecuadamente para el proceso de minería de datos posterior. Esto es, tratamiento y eliminación de *missing values*[14] y *outliers*[15].

Tratamiento de *missing values*:

En primer lugar, se tratan los valores perdidos de las variables anteriormente mencionadas. A continuación, se resume el tratamiento realizado:

1. Los marcajes válidos corresponden a los que tienen al menos una marca de entrada y una marca de salida, con mínimo tres horas de diferencia entre las marcas. También se consideran válidos los marcajes con dos marcas de entrada y dos marcas de salida. De la base de datos hay un 4% que no cumplen estos requisitos, por lo que se eliminan.
2. Se identifica solo 1 ID sin jornada laboral. Por ello, se desconoce los días que debe trabajar por contrato, por lo que se imputa la moda de la jornada de la sucursal donde trabaja.
3. Se desconoce la edad de dos personas, por lo que se imputa la edad mediante regresiones lineales.
4. Se desconoce la fecha de ingreso de solo 1 persona con 144 observaciones por lo que se elimina de la base de datos.
5. Se desconoce el estado civil de solo una persona por lo que se imputa la moda.
6. Se desconoce la nacionalidad de tres personas, pero gracias a su Rut se identifica que son extranjeras.
7. Había una gran cantidad de casos donde personas marcaban su entrada y salida, sin embargo, el sistema no identificaba el lugar donde marcó la persona, por lo que se tomó la moda de la localización de sus marcaciones. Se podría pensar que el sistema no reconoció el lugar porque la persona marcó en otro terminal, pero es un error bastante común, pues el sistema no siempre guarda el origen de la marca.

Tratamiento de *outliers*:

En segundo lugar, se tratan los valores atípicos o inconsistencias de la base de datos. A continuación, se resume el tratamiento realizado:

1. 1,2% de la base de datos corresponden a casos en los que los trabajadores marcaban en un turno que no les correspondía, es por esto que dichos casos se tomaron como un cambio de turno no mapeado en el sistema, tomando como requisito mínimo que hubiesen trabajado el equivalente al 90% su jornada laboral por contrato. Los que no

[14] Corresponden a los valores faltantes cuando no se almacena ningún valor para una variable de observación.

[15] Corresponden a valores atípicos de una observación, es decir, un valor que es numéricamente distante del resto de observaciones

cumplen esta condición, son cerca del 20% de los casos detectados (1.720), por lo que se eliminan de la base de datos.

2. 0,5% de la base de datos corresponden a ausencias injustificadas sistemáticas, es decir, que ocurren más de 4 días seguidos, por lo que se eliminan porque son considerados como marcajes omitidos o ausencias justificadas no mapeadas.
3. Para las variables de edad, meses de contratación y distancia, se eliminan los outliers mediante el uso de *boxplots*.

9.3 Análisis descriptivo de los datos

A continuación, se explica la distribución de las variables más relevantes de la base de datos:

1. **Evolución de empleados en el tiempo:** Desde 2017 se observa que la cantidad de trabajadores ha ido creciendo en el tiempo, con bajas en los meses de febrero y septiembre de cada año. Además, se observa que el número de trabajadores comienza a descender a partir de mayo de 2019, lo que podría explicarse por la reducción de los resultados financieros de la empresa. Por último, debido al *estallido social*[16] y a la pandemia originada por el *coronavirus*[17], la base de datos se corta a partir del 11 de octubre.

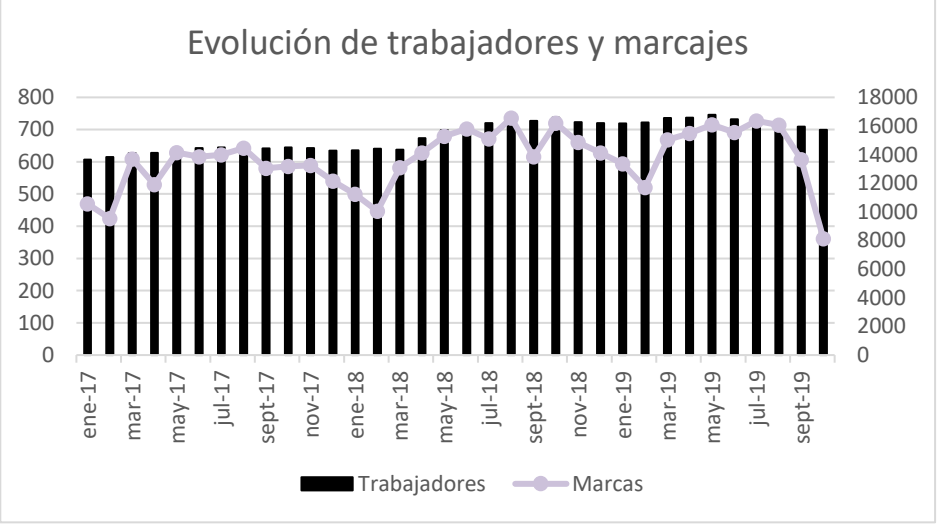


Gráfico 1: Evolución de número de trabajadores y marcajes.

2. **Ausentismo laboral:** El ausentismo laboral corresponden a todas aquellas ausencias que ocurren de manera inesperada y que no se justifican en la plataforma. A lo largo de los años, el ausentismo sigue una curva bastante atípica, aunque se logran detectar algunas similitudes. Por ejemplo, en junio el ausentismo tiende a subir, lo que coincide con el Mundial de Fútbol de 2018 y la Copa América de 2019. Contra intuitivamente, el ausentismo disminuyó el 2018 durante septiembre, mes de las fiestas patrias en Chile. En el gráfico 2, se observa que hay una leve similitud entre la cantidad de marcas y el ausentismo. En adición, el ausentismo corresponde al 1,5% del total de las observaciones y en promedio se ausentan nueve personas al día de manera injustificada.

[16] Serie de manifestaciones y disturbios originados el 18 de octubre en Santiago y propagados a todo el país, culminando en una crisis social.

[17] Pandemia mundial derivada por el virus *COVID-19*. A la fecha se registran 5,8 millones de casos a nivel mundial. El virus ha podido ser contenido por algunos países, pero aún no se ha podido eliminar, ni tampoco se ha encontrado una vacuna.

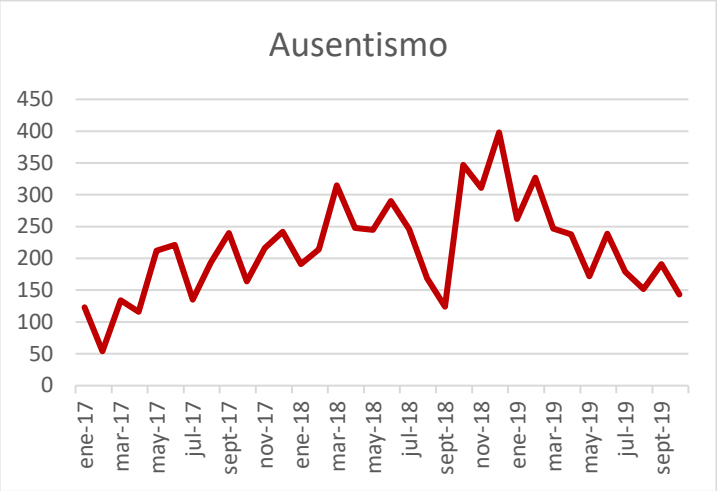


Gráfico 3: Evolución de ausentismo por mes.

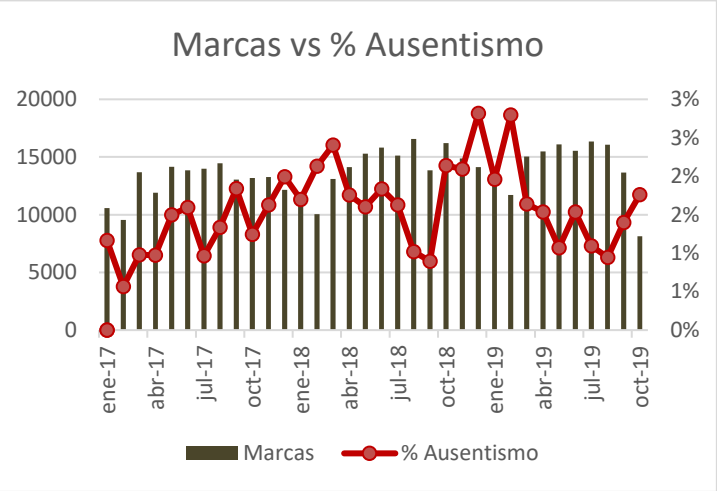


Gráfico 2: Número de marcas vs porcentaje de ausentismo.

3. **Atraso laboral:** El atraso laboral se calcula a partir de los marcajes de entrada, que ocurren al menos cinco minutos más tarde que el turno asignado. La política de la empresa define los descuentos por atrasos de esta forma, lo que también es de conocimiento común entre los trabajadores.

   De manera agregada, los atrasos suman el 12,2% del total de las observaciones, es decir, en promedio se atrasan 75 personas al día y el atraso promedio es de 32 minutos.

   En el grafico 4 se observa que los minutos de atraso son directamente proporcionales a la cantidad de atrasos. Además, en el grafico 5 se visualiza que la curva de atrasos es bastante regular, donde a priori no se identifica ningún patrón claro.

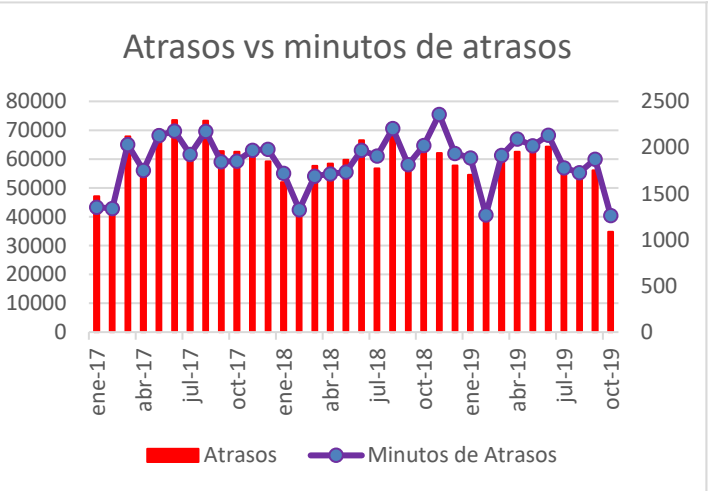


Gráfico 4: Atrasos vs minutos de atrasos.

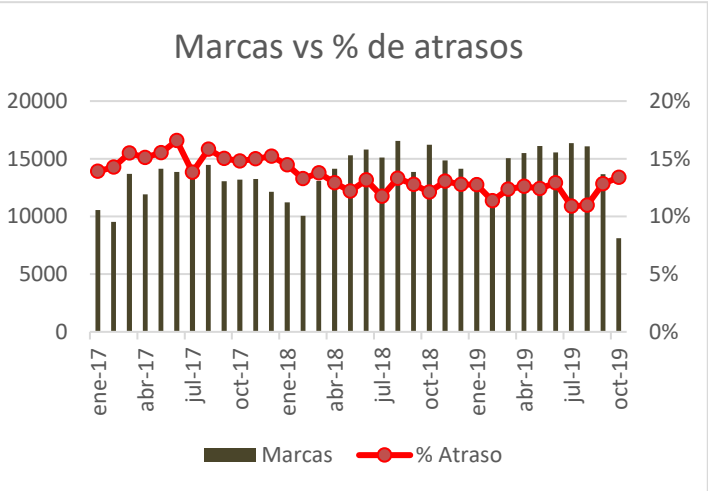


Gráfico 5: Número de marcas vs % de atrasos.

4. **Año**: En el gráfico 6 se observa que no hay una relación clara entre los ausentismos y los años. No obstante, los atrasos decrecen desde 2017, lo que se explica por la integración de políticas de descuento por atrasos en 2018.

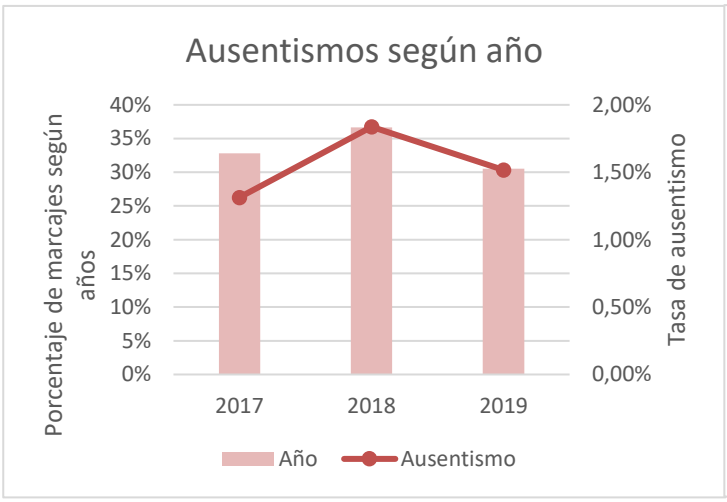


Gráfico 6: Marcaje y ausentismo según año.

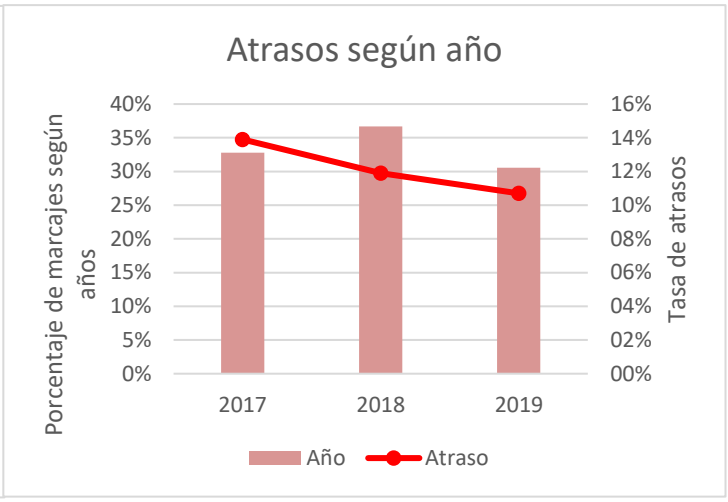


Gráfico 7: Marcaje y ausentismo según mes.

5. **Mes:** En el gráfico 8 se observa que los ausentismos tienden a crecer a partir de agosto y hasta fin de año. Por otro lado, los atrasos siguen una curva más regular, registrando mínimos en los meses de enero y febrero, lo que puede explicarse porque en estos meses típicamente la gente sale de vacaciones y, en consecuencia, se registra menos tráfico en las calles.

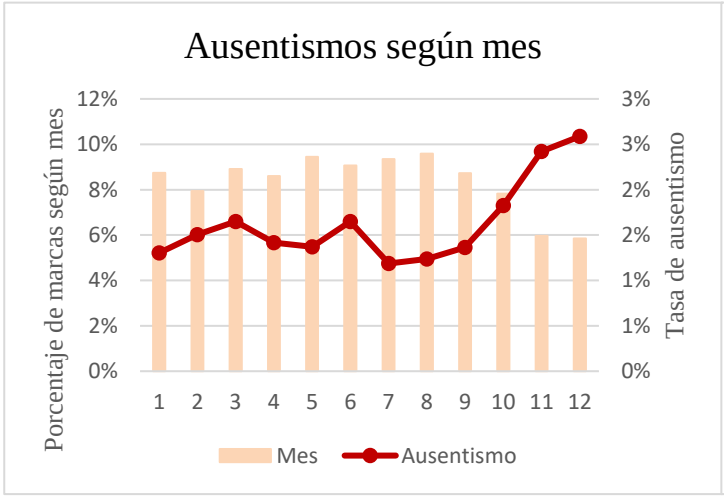


Gráfico 8: Ausentismos según mes

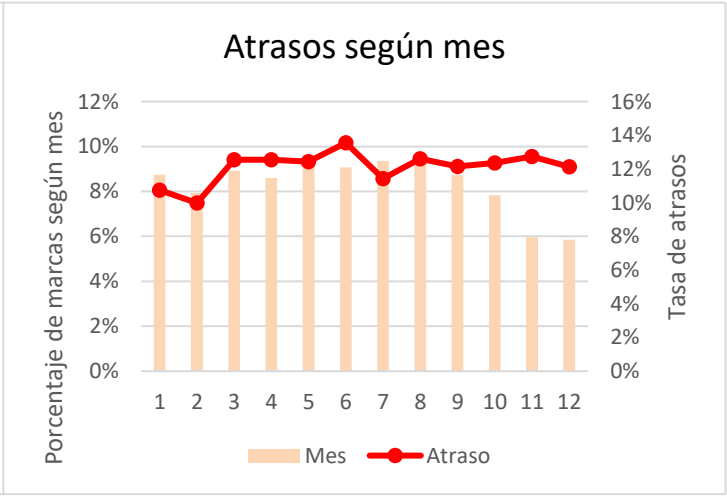


Gráfico 9: Atrasos según mes.

6. **Día de la semana y semana del mes:** En el gráfico 10 se observa que las marcas descienden para el sábado y el domingo, pues para estos días la demanda es inferior. Además, se ve que el ausentismo alcanza su máximo el sábado, mientras que su mínimo se registra el domingo. Por otro lado, en el gráfico 11 se muestra que el ausentismo no varía mucho con respecto a las semanas del mes.

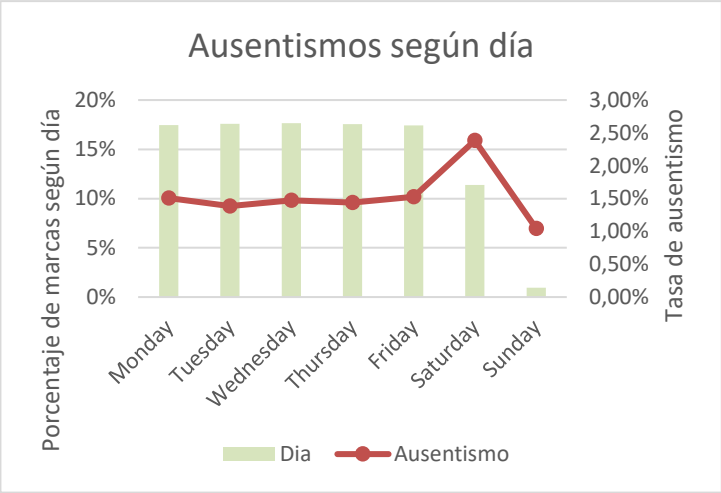


Gráfico 10: Ausentismo según día de la semana.

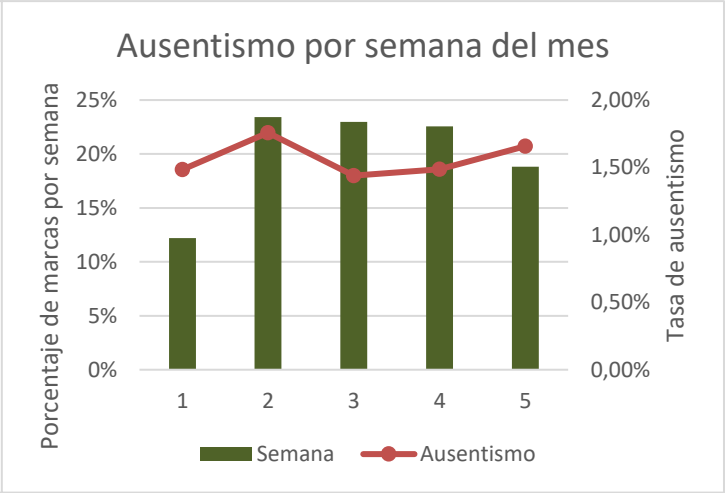


Gráfico 11: Ausentismo según semana del mes.

Para la tasa de atraso, se observa que es levemente superior los domingos, mientras que a fin de mes tiende ser un poco mayor y desciende la primera semana de cada mes.

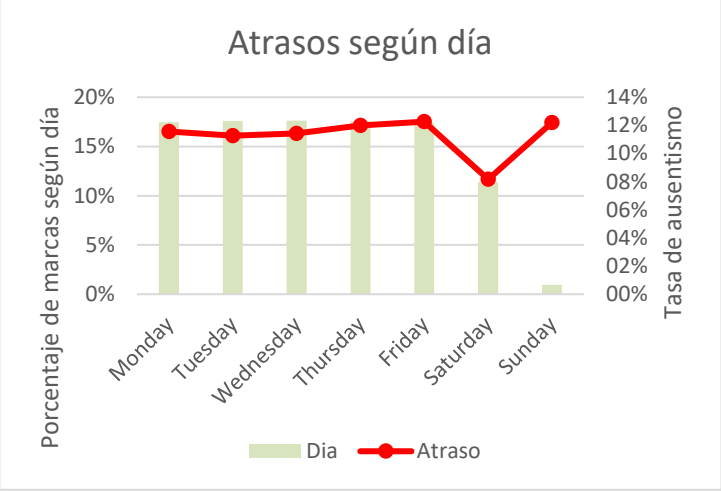


Gráfico 13: Atrasos según día de la semana.

Gráfico 12: Atrasos según semana del mes.

7. **Terminal:** Los terminales son las plataformas donde los trabajadores marcan su entrada y salida, hay 13 de estos, asignados cada uno a una sucursal diferente. Como se observa, coinciden que los terminales con mayor ausentismo también tienen tasas de atraso muy altas, por ejemplo: las terminales 3 y 12 son las terminales con ausentismo más altas y también tienen tasas de atraso muy por encima del promedio, esto se relaciona a una de las hipótesis de la dirección que sostienen que hay sucursales menos estrictas en cuanto al cumplimiento del horario.

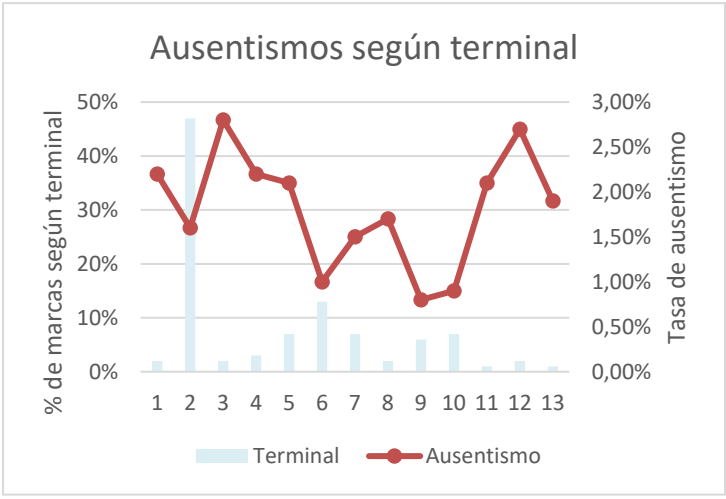


Gráfico 14: Ausentismos según terminal

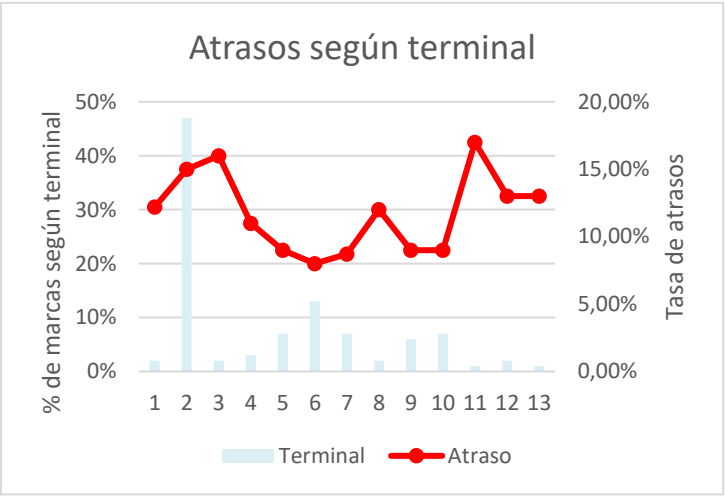


Gráfico 15: Atrasos según terminal

8. **Sexo:** Del total de trabajadores, 72% corresponden a hombres, mientras que el 28% son mujeres. En los gráficos también se observa que los hombres son más propensos a ausentarse, mientras que las mujeres tienden a atrasarse más. Según uno de los jefes de tienda, esta relación ocurre porque las mujeres tienden a ser las principales encargadas de transportar a sus hijos al jardín y/o colegio.

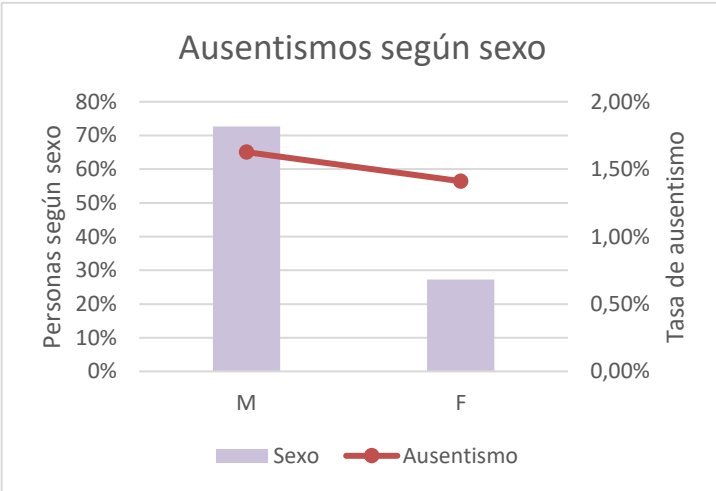


Gráfico 16: Ausentismo según sexo.

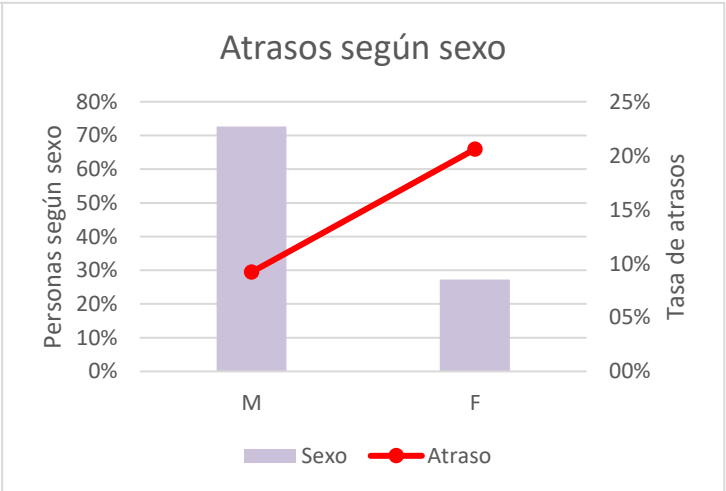


Gráfico 17: Atraso según sexo.

9. **Edad:** Se observa que el ausentismo es más alto en los dos primeros grupos etarios, mientras que el atraso laboral se concentra en los tres primeros. Esto se puede explicar porque los grupos etarios de mayor edad suelen tener un mejor comportamiento, ya que en caso de despido es más difícil reinsertarse en el mundo laboral.

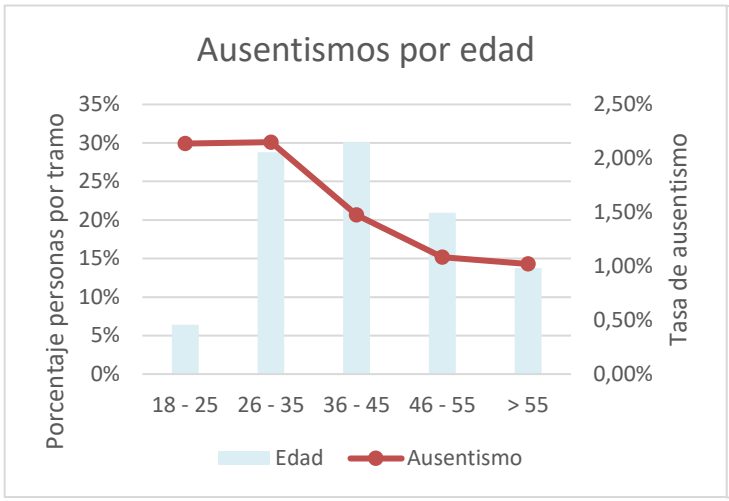


Gráfico 19: Ausentismo por edad.

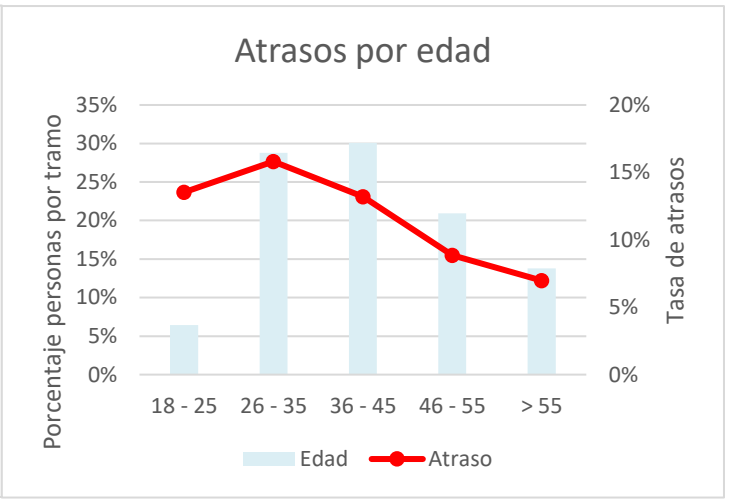


Gráfico 18: Atraso por edad.

10. **Antigüedad en la empresa:** En los gráficos se observa que los trabajadores más nuevos tienden a ausentarse en mayor tasa que los más experimentados. Similar ocurre para los atrasos, donde la tasa de atrasos es mayor para los trabajadores más nuevos. Las cifras caen para el segundo tramo (1 – 3 años) y vuelven a subir.

    En conversaciones con jefes de tienda de la empresa, explican que estas relaciones se dan porque los trabajadores comprenden con la experiencia que los ausentismos pueden culminar en causal de despido, mientras que rara vez son despedidos por atrasarse de manera reiterada.

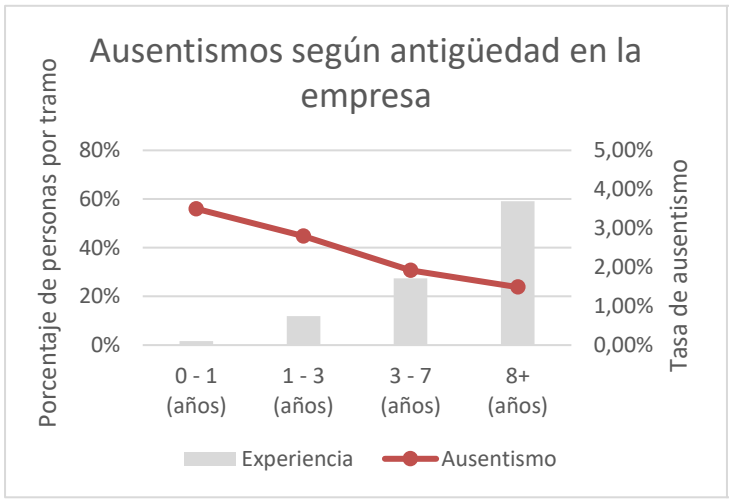


Gráfico 21: Ausentismo según antigüedad en la empresa.

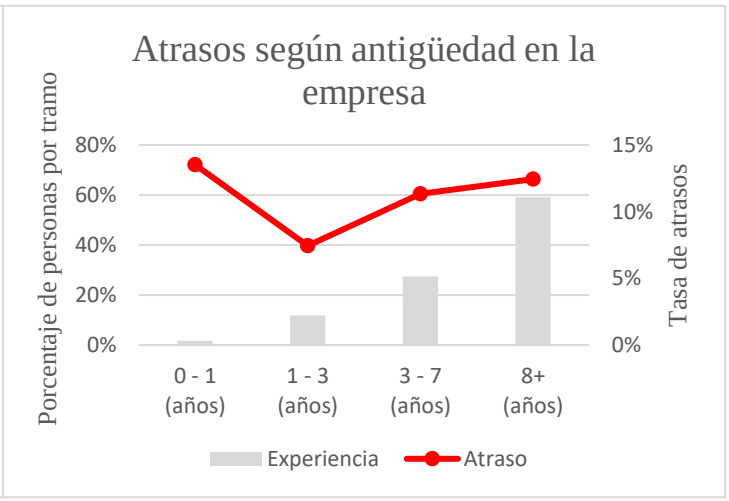


Gráfico 20: Atrasos según antigüedad en la empresa.

11. **Cargo:** En los gráficos se observa que los cargos más predominantes corresponden a los de operador y administrativo, mientras que el ausentismo y el atraso laboral tiende a ser más elevado para los grupos con menores trabajadores (chofer, contabilidad, ejecutivo y analistas, entre otros). Esto valida una de las hipótesis iniciales, que destaca que el cargo y cantidad de integrantes en la estructura de la organización afectan el comportamiento de ausentismo y atraso laboral.

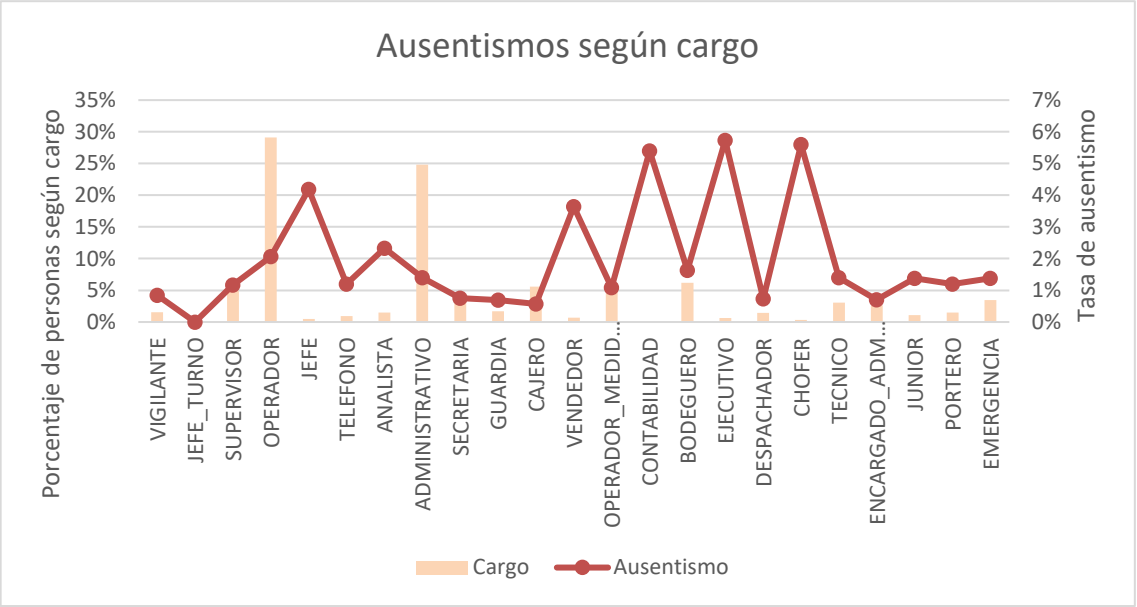


Gráfico 22: Ausentismo según cargo.

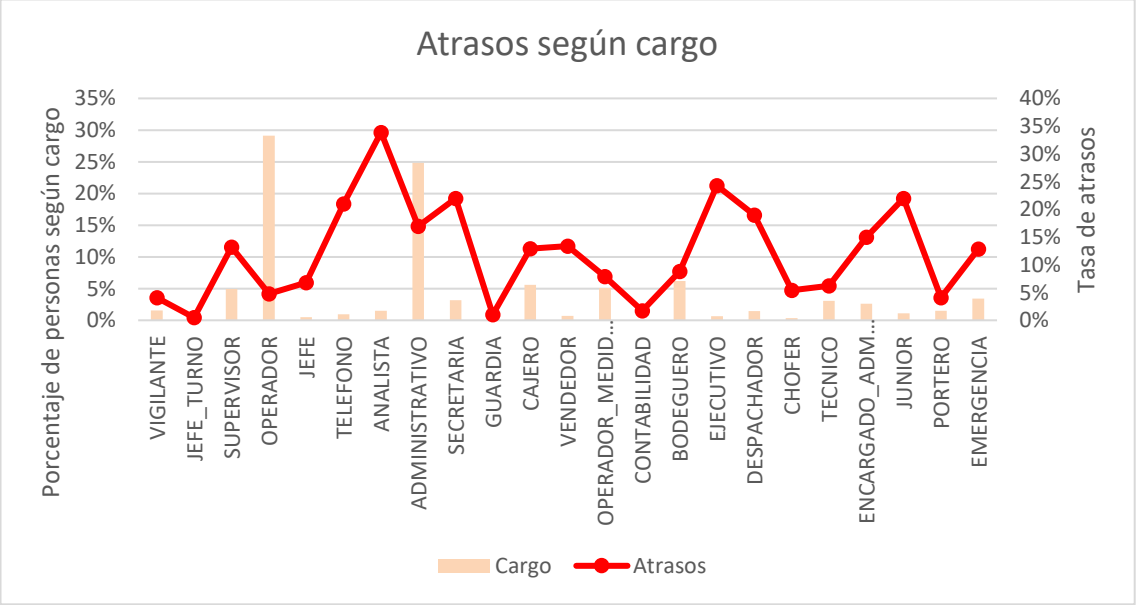


Gráfico 23: Atrasos según cargo.

12. **Turno:** Se observa que tanto el ausentismo como el atraso son más altos en los turnos de mañana y de tarde, lo que se explica por el tráfico que se genera en los horarios de entrada para estos turnos.

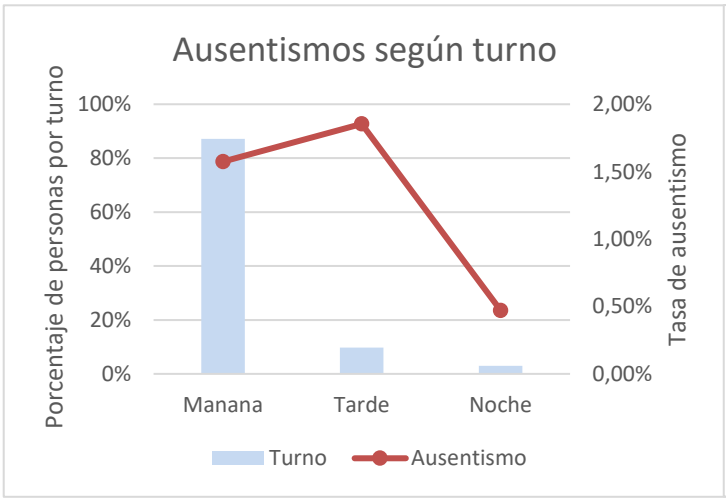


Gráfico 24: Ausentismo según tipo de turno.

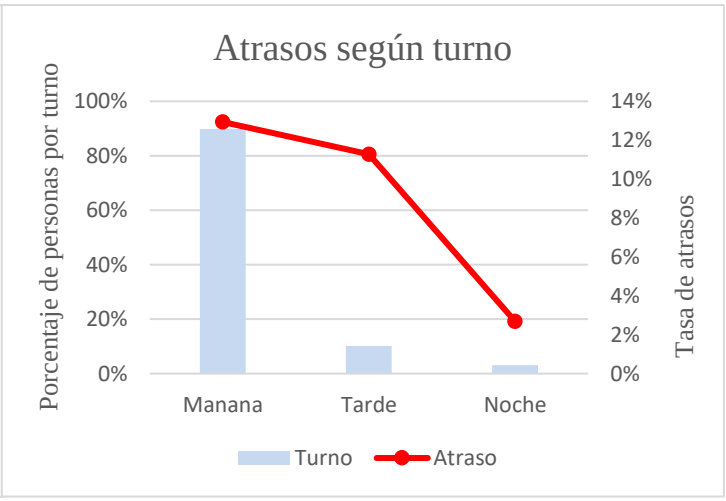


Gráfico 25: Atrasos según tipo de turno.

13. **Estado civil:** Se observa que los viudos y solteros tienden a ausentarse en mayor proporción, mientras que los solteros y divorciados son los más propensos a llegar tarde.

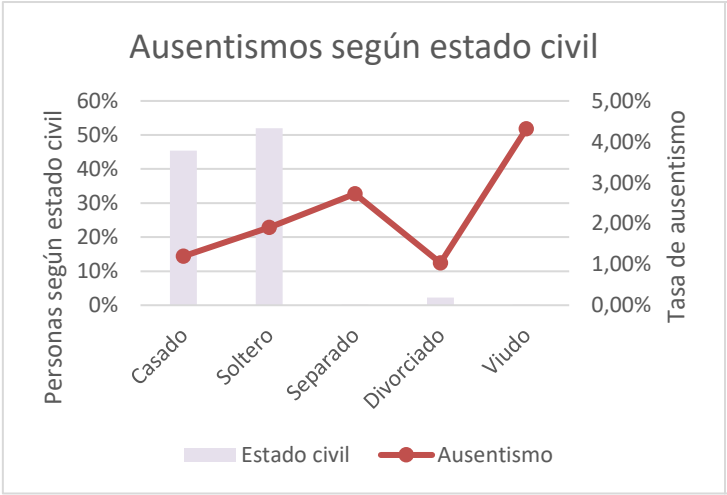


Gráfico 26: Ausentismo según estado civil.

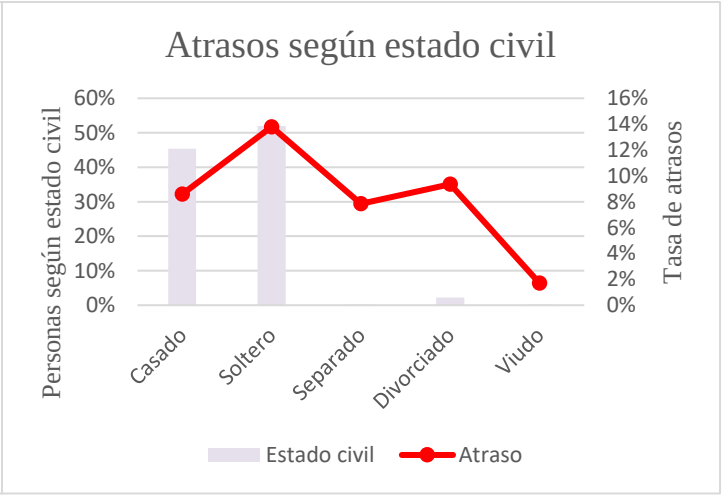


Gráfico 27: Atrasos según estado civil.

14. **Horario:** No hay diferencias significativas en la tasa de ausentismo, no obstante, la tasa de atrasos para los *full-time* es ligeramente superior, esta diferencia se puede explicar pues la mayoría de los *full-time* hacen turnos de mañana y en esta jornada la tasa de atraso es mayor debido a otras variables exógenas como el tránsito, congestión en el sistema público, tener que ir a dejar a los hijos al colegio, entre otras.

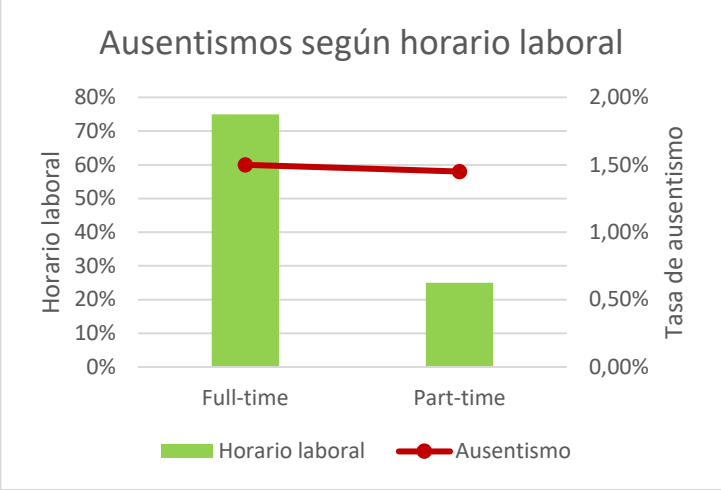


Gráfico 28: Ausentismo según horario laboral.

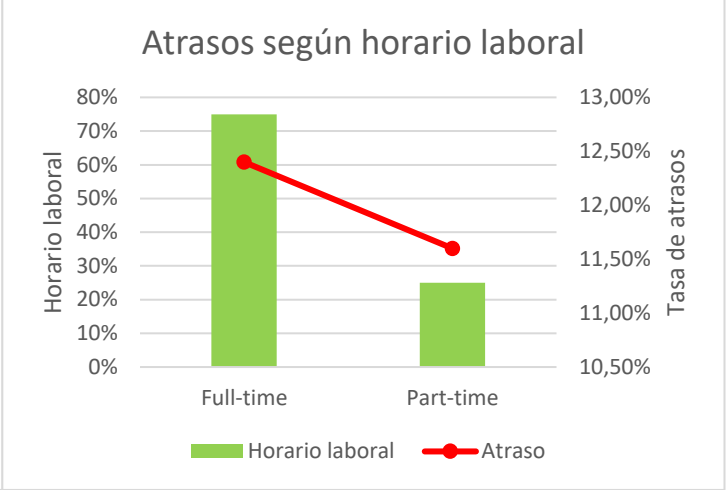


Gráfico 29: Atrasos según horario laboral.

15. **Nación:** No se observan diferencias significativas en las tasas de ausentismo según la nacionalidad. Por otro lado, los chilenos tienden a atrasarse más que los extranjeros, pues según jefes de tienda los trabajadores extranjeros tienden a cuidar más su trabajo.

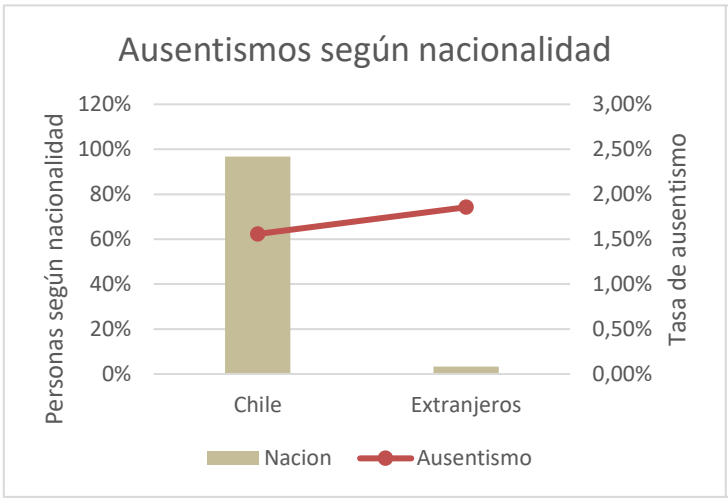


Gráfico 31: Ausentismo según nacionalidad.

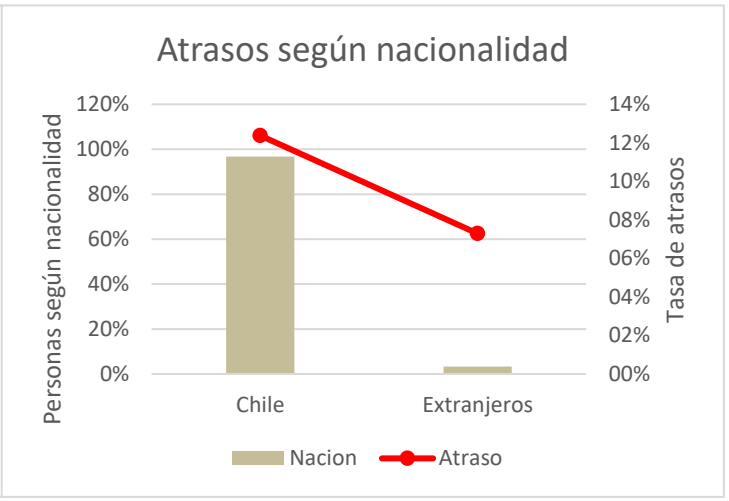


Gráfico 30: Atrasos según nacionalidad.

16. **Forma de pago:** Los trabajadores que reciben cheques como forma de pago tienden a ausentarse y atrasarse más. El atraso puede explicarse porque muchas veces los trabajadores se ausentan los primeros minutos de trabajo para cobrar sus cheques. Por otro lado, esta forma de pago también se relaciona con el perfil del trabajador, lo que se captura con la variable de "Cargo".

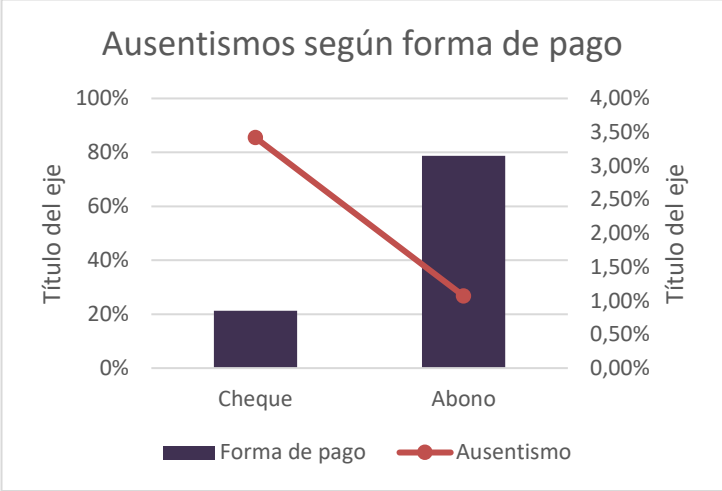


Gráfico 33: Ausentismo según forma de pago.

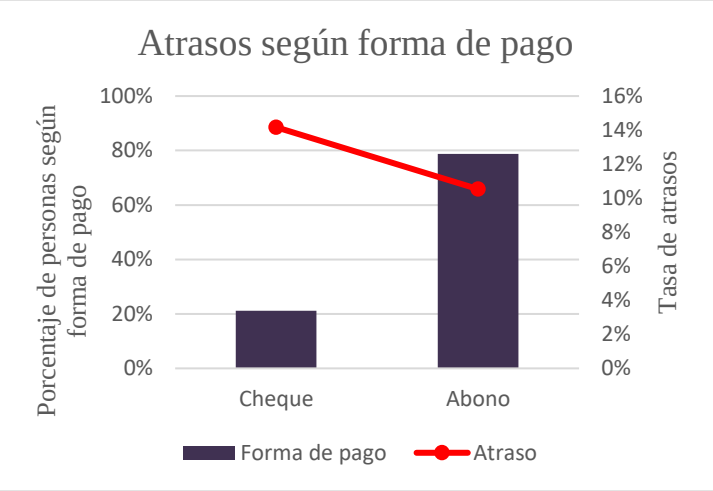


Gráfico 32: Atrasos según forma de pago.

17. **Tipo de contrato:** Se observa que los trabajadores con plazo fijo tienden a atrasarse y ausentarse en mayor proporción que aquellos con contrato indefinido. Esto se debe a que los trabajadores con contrato fijo saben que es más difícil que los amonesten, pues solo estarán en la empresa de manera temporal y tienden a comportarse peor al final de su período.

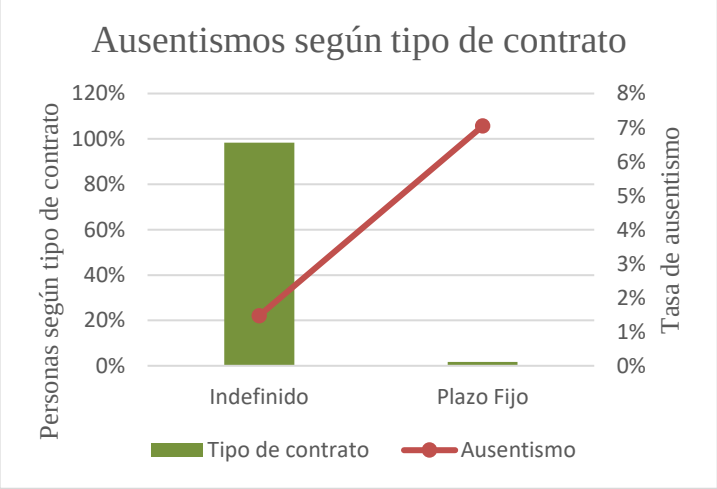


Gráfico 35: Ausentismo según tipo de contrato.

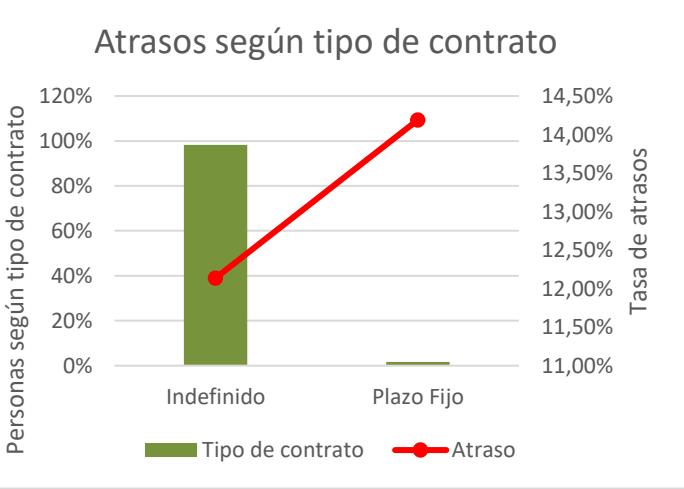


Gráfico 34: Atrasos según tipo de contrato.

18. **Jornada:** En el grafico 36 se observa que no hay diferencia en la tasa de ausentismo entre los colaboradores que trabajan de lunes a sábado y de lunes a viernes. No obstante, en la tasa de atrasos si hay diferencias bastante grandes, donde los trabajadores con jornada laboral de lunes a viernes tienden a atrasarse con mucha mayor frecuencia.

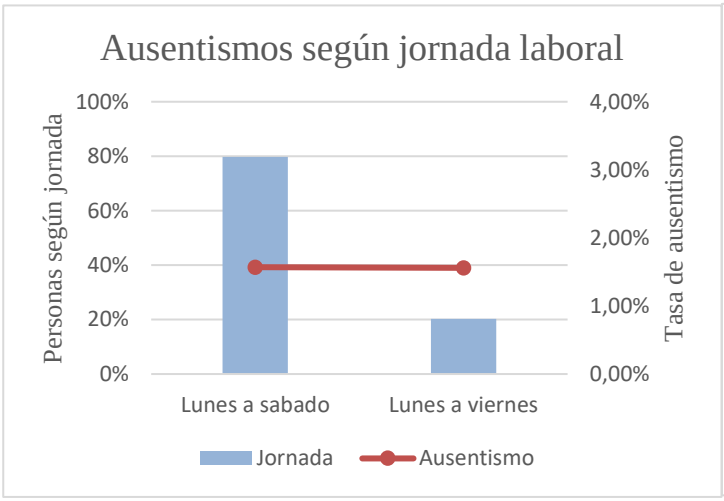


Gráfico 36: Ausentismo según jornada laboral.

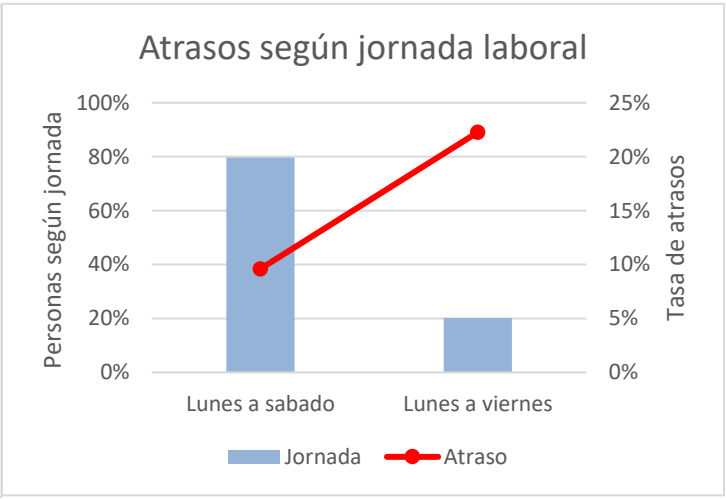


Gráfico 37: Atrasos según jornada laboral.

19. **Distancia:** A través de la *API* de *Google Maps*[18] se calculan las distancias de los trabajadores desde el domicilio hasta el lugar de trabajo y se ve una relación directamente proporcional entre la distancia y las tasas de atrasos y ausentismos.

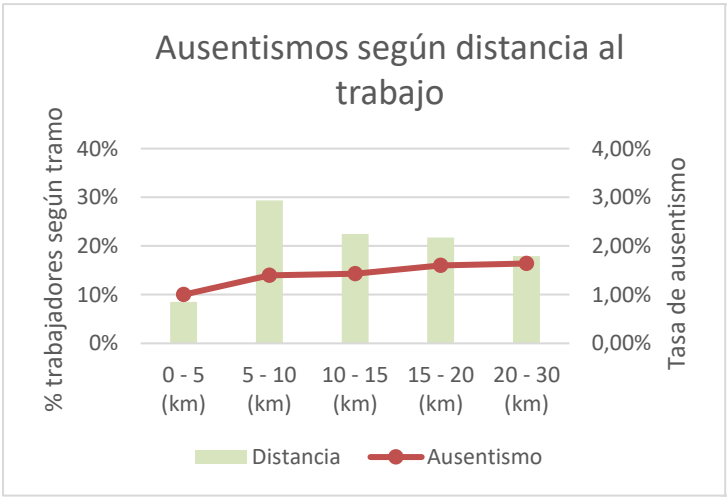


Gráfico 38: Ausentismo según distancia.

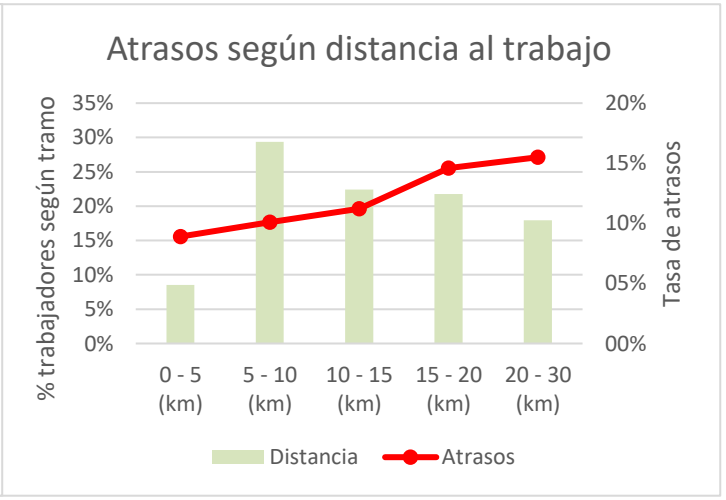


Gráfico 39: Atraso según distancia.

[18] https://console.cloud.google.com/apis/library/directions-backend.googleapis.com?hl=es&project=totemic-chalice-270820

20. Temperatura: A partir de las bases de datos del *Center for Climate and Resilience Research*[19], se extraen las temperaturas diarias mínimas de los sensores más próximos a los terminales de marcaje.

Se observa que, a nivel de ausentismo, no hay una relación significativa. Sin embargo, para los atrasos se ve que en los días más fríos la tasa de atrasos es mayor, lo que se explica por la presencia de mayor tasa de enfermedades, posibilidades de lluvia y el comportamiento de las personas en base a las bajas temperaturas.

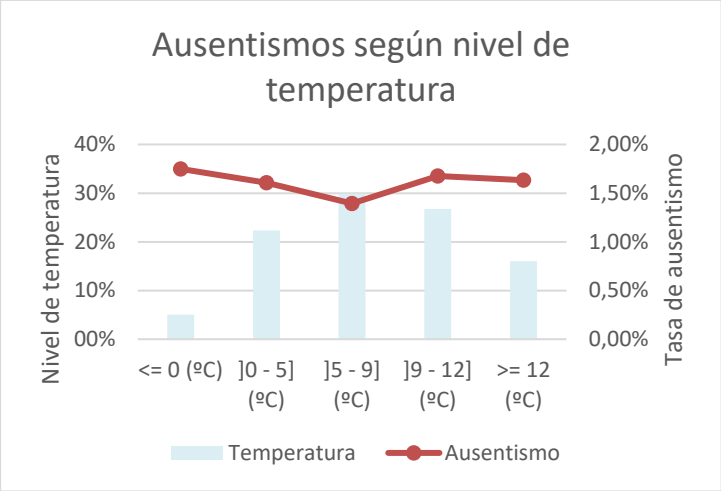


Gráfico 41: Ausentismo según temperatura.

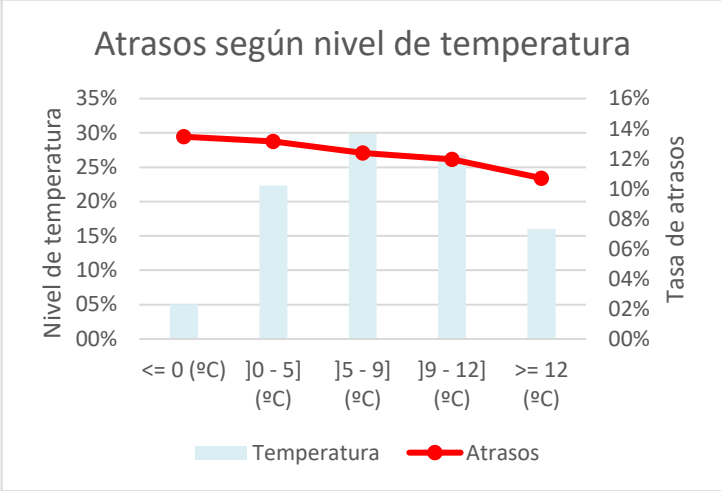


Gráfico 40: Atraso según temperatura.

21. Precipitaciones: A partir de otra base de datos del *Center for Climate and Resilience Research* [20], se extraen las precipitaciones diarias acumuladas de los sensores más cercanos a los terminales de marcaje.

De acuerdo con Monjo (2010), se clasifican las precipitaciones según su intensidad en tres escalas: lluvia débil, lluvia moderada y lluvia fuerte.

Del gráfico 42, se observa que a medida que aumenta la intensidad de las precipitaciones suben los atrasos, lo que puede explicarse por los efectos adversos que tienen las precipitaciones sobre las vías de transporte.

Por otro lado, la tasa de ausentismo sube para las lluvias moderadas, pero contra intuitivamente baja en las lluvias fuertes, esto puede entenderse porque para las fechas con lluvia, la exposición a enfermedades crece, pero como la frecuencia de las lluvias fuertes es menor al 1%, no es posible capturar este efecto.

[19] http://www.cr2.cl/bases-de-datos/#observacionales

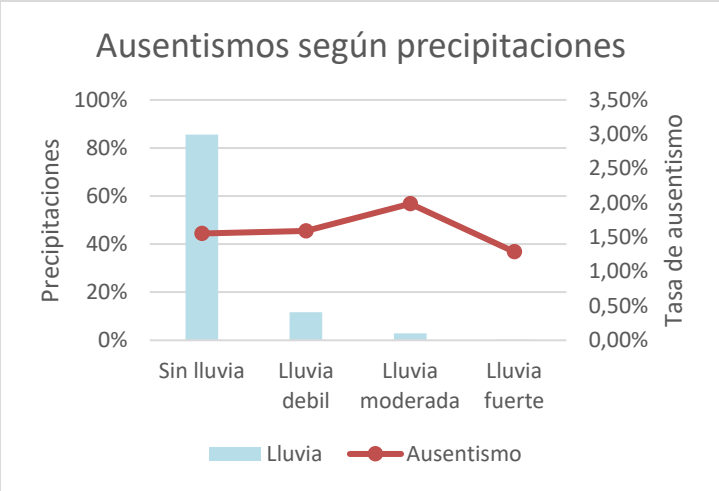


Gráfico 43: Ausentismo según precipitaciones.

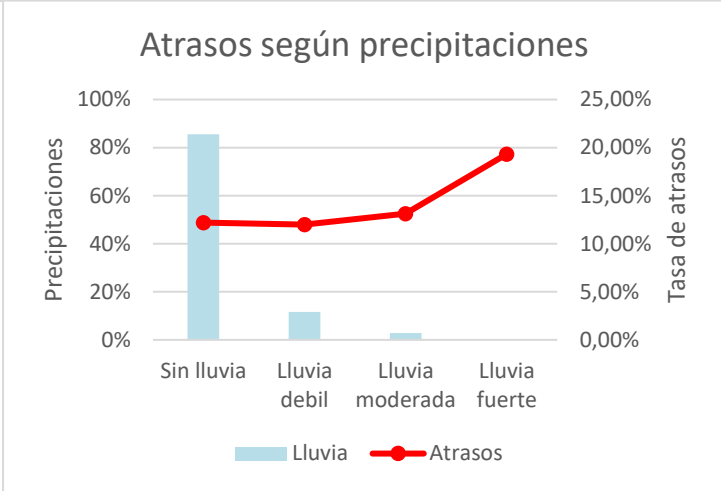


Gráfico 42: Atraso según precipitaciones.

Resumen etapa exploratoria

En esta sección se muestra la relación entre las variables a disposición y los atrasos y ausentismos, detectando que ciertas variables serán de ayuda para estimar estos fenómenos.

Para los atrasos, se observan relaciones importantes en el género, edad, antigüedad, cargo, turno, estado civil, horario laboral, nacionalidad, forma de pago, tipo de contrato, jornada laboral, precipitaciones, distancia y temperaturas.

Asimismo, con el objetivo de mostrar relaciones cuantificables entre las variables y la magnitud de los minutos de atrasos, se corre una regresión lineal utilizando el método de Mínimos Cuadrados Ordinarios (MCO), cuidando que se cumplan los supuestos de correlación de variables y homocedasticidad. De esta forma, se obtienen los siguientes coeficientes:

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.5233	0.139	53.977	0.000	7.250	7.797
SEXO	1.043e+09	6.63e+10	0.016	0.987	-1.29e+11	1.31e+11
NACION	1.839e+05	1.17e+07	0.016	0.987	-2.27e+07	2.31e+07
EDAD	-0.0773	0.003	-23.956	0.000	-0.084	-0.071
FERIADO	6.053e+04	3.85e+06	0.016	0.987	-7.48e+06	7.6e+06
PARTIDO	0.2820	0.134	2.110	0.035	0.020	0.544
MOVILIZACION	0.3115	0.102	3.066	0.002	0.112	0.511
SEXO	-1.043e+09	6.63e+10	-0.016	0.987	-1.31e+11	1.29e+11
NACION	-1.839e+05	1.17e+07	-0.016	0.987	-2.31e+07	2.27e+07
FC	-0.0027	0.521	-0.005	0.996	-1.023	1.018
MESES_CONTRATA	0.0068	0.000	22.255	0.000	0.006	0.007
Horas_extra	2.2750	0.029	77.266	0.000	2.217	2.333
Atraso t-1	2.5494	0.033	78.317	0.000	2.486	2.613
LLUVIA	0.4069	0.090	4.521	0.000	0.230	0.583
15 - 20 km	1.8015	0.099	18.185	0.000	1.607	1.996
10 - 15 km	1.4220	0.102	13.893	0.000	1.221	1.623
20 - 30 km	2.3816	0.074	32.134	0.000	2.236	2.527
5 - 10 km	0.7925	0.088	9.046	0.000	0.621	0.964

El R^2 de este modelo es 0.026, por lo que no se pueden generalizar todas las conclusiones a partir de la interpretación de los coeficientes. No obstante, estos coeficientes ayudan a reforzar las relaciones que se han mostrado hasta ahora. Por ejemplo: las distancias, lluvias y meses de contratación influyen de manera directamente proporcional en los minutos de atrasos, mientras que bajan a medida que el trabajador es mayor. Además, se agregan las horas extra y se evidencia una correlación fuerte entre éstas y los minutos de atraso, de lo que se puede interpretar que los trabajados hacen 2 minutos de horas extra por cada minuto de atraso.

Para los ausentismos, las variables con relaciones más claras son el mes, día de semana, edad, antigüedad, cargo, turno, estado civil, nacionalidad, tipo de contrato y distancia.

9.4 Minería de datos

9.4.1 Entrenamiento y testing

Para entrenar y testear los modelos de minería de datos, se seleccionaron las observaciones de la base de datos a partir del 1 de enero de 2017 hasta el 11 de octubre de 2019 y se separó la base de datos en entrenamiento y testing.

Para el entrenamiento de los modelos de atraso y ausentismo, se seleccionó el 80% de las observaciones y un 20% para testear. Se tuvo en consideración, que las fechas de las observaciones de entrenamiento y testing fueran distintas y no existiera traslape.

Los modelos son usados para estimar los atrasos y ausentismos. Como criterio de selección se optó por los modelos con mayor *AUC.*

9.4.2 Resultados modelos de atraso

Para estimar los atrasos se utilizaron dos modelos *weak learners*[21]: *Regresión Logística* y *Naive Bayes*, y otros modelos de *Machine Learning* más robustos, *Decision Tree, Gradient Boosting, Random Forest y XGBoost*.

A continuación, se muestran los resultados de estos modelos que son usados como línea base, con el objetivo de aplicar técnicas metodológicas para mejorar sus desempeños en las secciones posteriores.

Curva *ROC* de los modelos de atrasos:

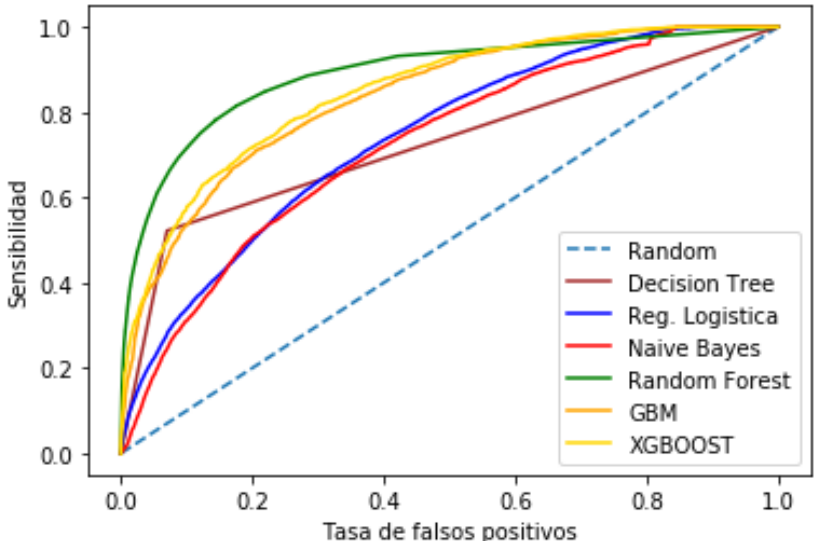


Ilustración 1: Curva ROC modelos de atraso.

[21] Un clasificador *weak learner* es uno que tiene un desempeño bajo, cuyo *accuracy* es siempre mejor que el azar en la base de entrenamiento.

Al observar la curva *ROC* de los modelos, se ve que los modelos que usan árboles en su estructura son mejores clasificadores que la *Regresión Logística* y *Naive Bayes*. Además, se descubre que los mejores modelos son el *XGBOOST* y el *Random Forest*.

Para comparar de mejor manera estos modelos, se expone una tabla resumen con las métricas de desempeño de cada uno[22]:

Modelo	***F_1***	***AUC***	***Accuracy***	***Precision***	***Recall***
Random Forest	0.543	0.891	0.914	0.698	0.445
Decision Tree	0.507	0.721	0.884	0.504	0.521
XGBoost	0.508	0.848	0.917	0.768	0.38
Naive Bayes	0.235	0.723	0.264	0.133	0.986
Gradient Booster	0.15	0.835	0.891	0.678	0.084
Logistic Regression	0.048	0.742	0.886	0.504	0.025

Tabla 7: Métricas de desempeño, modelos de atraso.

El modelo que mejor *f_1-score* presenta es el *Random Forest* y el con mayor precisión es el *XGBoost*. A continuación, se muestra un análisis más detallado de los modelos:

Tabla de comparación de métricas de desempeño para cada clase

	Random Forest		***XGBoost***	
	Puntual	**Atrasado**	**Puntual**	**Atrasado**
Precision	0,93	0,70	0,92	0,77
Recall	0,98	0,45	0,99	0,38
F_1 Score	0,95	0,54	0,95	0,50

Tabla 8: Métricas de desempeño, Random Forest y XGBoost, modelos de atraso.

Si bien el *XGBoost* tiene una precisión más alta para la clase Atrasado, el *recall* del *Random Forest* es mucho mayor, lo que convierte al *Random Forest* en un mejor modelo en cuanto a las métricas de *AUC* y *f_1-score*. En la sección posterior, se describe la mejora de este modelo usando técnicas metodológicas.

Cabe destacar que la desventaja de estos modelos es que funcionan como una caja negra, lo que no permite entender cómo influye cada una de las variables en el fenómeno de estudio.

[22] Se considera un 50% como umbral de corte

Importancia de variables, atraso laboral

Con el objetivo de mitigar el sesgo de la ganancia de información del *Random Forest* y *XGBoost*, se categorizan las variables continuas (distancia, antigüedad, precipitaciones y temperatura) y se reduce la ordinalidad de las variables categóricas con mayor dimensión.

Al computar la ganancia de información de las variables del *Random Forest*, se obtiene la siguiente figura:

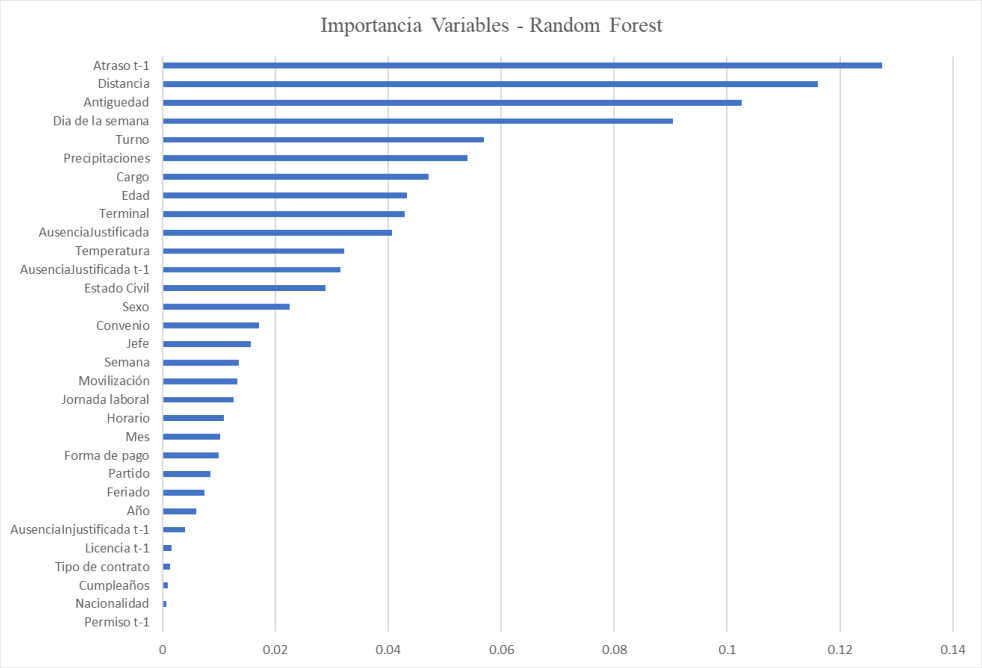


Ilustración 2: Importancia de variables, Random Forest de atrasos.

Se observa que el modelo considera con alta importancia las variables de atraso t-1, distancia, antigüedad, día de la semana, turno, precipitaciones, cargo, la edad y el terminal (ubicación) donde marca. Esto es coherente a los resultados del análisis exploratorio, donde se ven relaciones directamente proporcionales entre los atrasos y la antigüedad de los trabajadores, distancia y precipitaciones, junto con otras relaciones no necesariamente lineales con las demás variables.

De la misma forma, al computar la importancia de las variables para el *XGBoost*¸ se observa la siguiente figura:

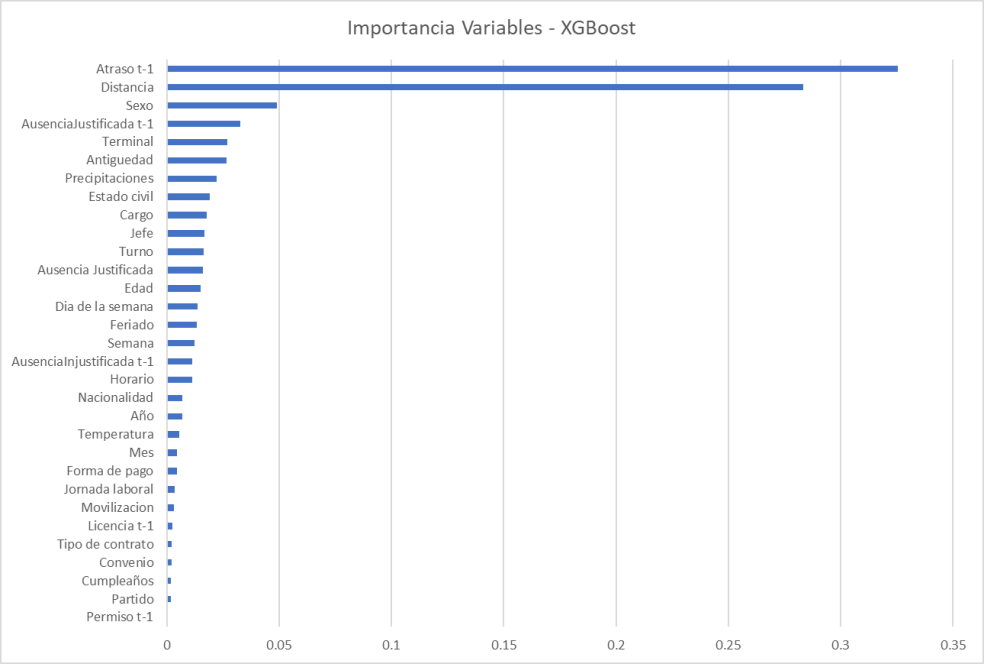


Ilustración 3: Importancia de variables, XGBoost de atrasos.

Asimismo, se observa que ambos modelos ordenan de manera similar las variables según su importancia, lo que refuerza la alta importancia de las variables de distancia, atraso en t-1, antigüedad, turno laboral, sexo, precipitaciones, cargo y edad.

Por otro lado, las variables con menos importancia son los permisos en t-1, partidos de futbol, cumpleaños, licencias en t-1 y tipo de contratos.

9.4.3 Resultados modelos de ausentismo

Para estimar los ausentismos se utilizaron los mismos modelos que para los atrasos: *Regresión Logística, Naive Bayes, Decision Tree, Gradient Boosting, Random Forest y XGBoost*.

Análogo a los modelos de atrasos, se muestran los resultados de estos modelos que son usados como línea base, con el objetivo de aplicar estrategias metodológicas para mejorar sus desempeños en las secciones posteriores.

Curva *ROC* de los modelos de ausentismo:

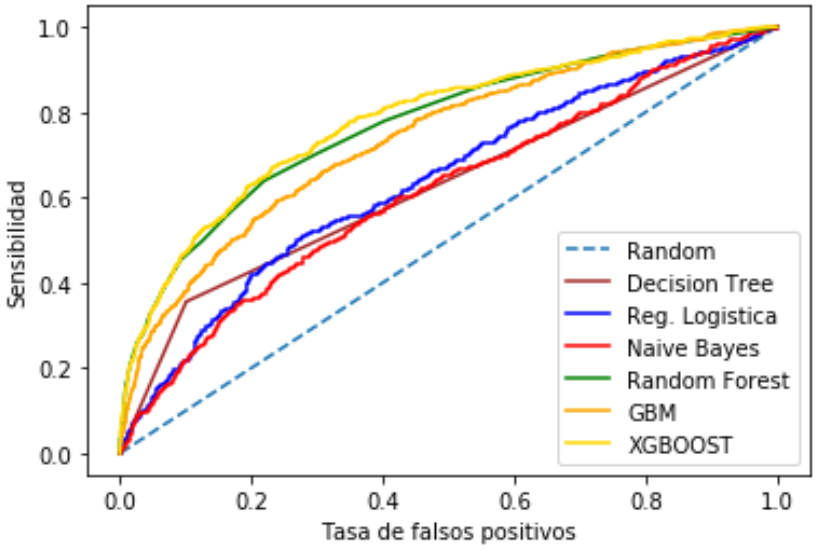


Ilustración 4: Curva ROC, modelos de ausentismo.

Similar a los modelos de atrasos, se ve que los modelos que usan árboles en su estructura son los mejores clasificadores, siendo nuevamente el *XGBoost* y el *Random Forest* los modelos con mejor curva *ROC*.

Modelo[23]	***F_1***	***AUC***	***Accuracy***	***Precision***	***Recall***
Decision Tree	0.142	0.575	0.984	0.129	0.158
Logistic Regression	0.016	0.665	0.992	0.124	0.009
Naive Bayes	0.019	0.669	0.158	0.010	0.988
Random Forest	0.049	0.763	0.992	0.489	0.026
Gradient Booster	0.002	0.747	0.992	0.200	0.001
XGBoost	0.021	0.787	0.992	0.500	0.011

Tabla 9: Métricas de desempeño, modelos de ausentismo.

[23] Se considera un 50% como umbral de corte.

Según las métricas de desempeño, los mejores modelos son el *Random Forest* y el *XGBoost*. No obstante, no dejan de ser deficientes, por lo que en la sección posterior se exhiben mejoras en el desempeño del *Random Forest*.

A continuación, se analizan con mayor detalle estos modelos:

Tabla de comparación de métricas de desempeño para cada clase

	Random Forest		***XGBoost***	
	Puntual	**Atrasado**	**Puntual**	**Atrasado**
Precision	0,99	0,49	0,99	0,50
Recall	0,99	0,03	0,99	0,01
F_1 Score	0,99	0,05	0,99	0,02

Tabla 10: Métricas de desempeño, Random Forest y XGBoost, modelos de ausentismo.

El *XGBoost* tiene una precisión ligeramente mayor para la clase Atrasado, pero el *recall* y *f_1-score* del *Random Forest* son mayores.

Importancia de variables, ausentismo laboral

De manera análoga a la importancia de las variables en los modelos de atraso, se computa la importancia del *Random Forest*, categorizando las variables continuas y reduciendo la ordinalidad de las variables categóricas de alta dimensión.

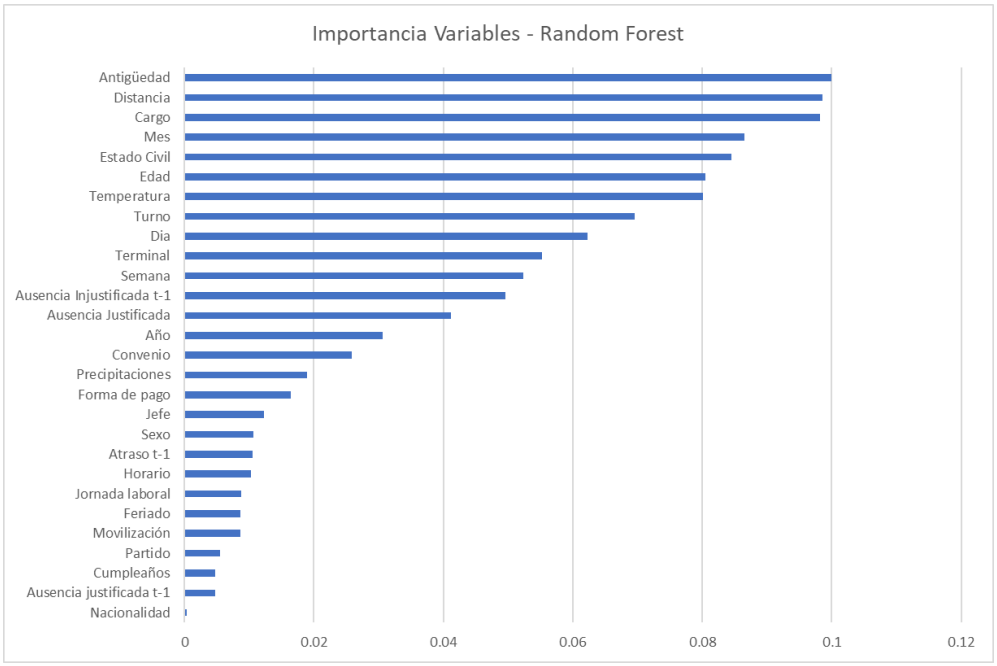


Ilustración 5: Importancia de variables, Random Forest de ausentismos.

El modelo pondera con gran importancia las variables de antigüedad, distancia, cargo, mes, estado civil y edad, lo que es coherente a lo observado en el análisis exploratorio,

Por otro lado, al computar la importancia de variables para el *XGBoost*, se obtiene la siguiente figura:

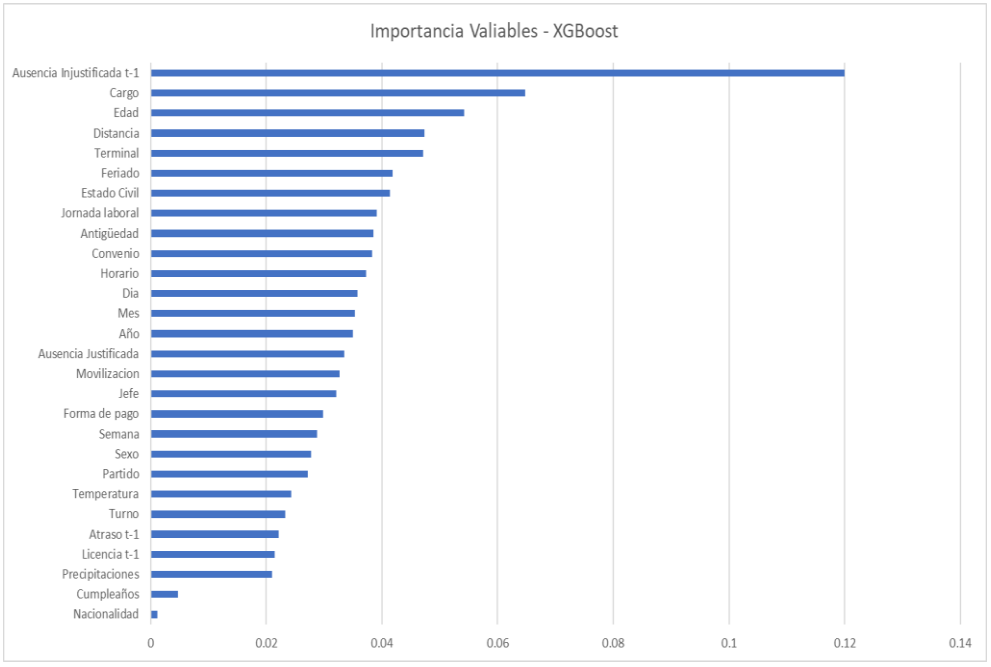


Ilustración 6: Importancia de variables, XGBoost de ausentismo.

Se observa que este modelo le asigna una mayor importancia a si el trabajador se ausenta el día anterior o no (Ausencia Injustificada t-1). Cabe señalar que rara vez los trabajadores se ausentan dos o más veces consecutivas.

Asimismo, se desprende que ambos modelos ordenan las variables de manera similar según su importancia, donde se destacan las siguientes variables: distancia, edad, cargo, terminal, estado civil y meses de contratación. Por otro lado, las variables con menos importancia son: la nacionalidad, licencias en t-1, atrasos en t-1, cumpleaños, partidos de futbol y movilizaciones, lo que es comparable a las relaciones encontradas en el análisis exploratorio de los datos.

9.4.4 *Hyperparameter Tunning*

En esta sección se muestra como el *tunning* de hiperparámetros ayuda a mejorar el desempeño de los modelos. Tanto para los atrasos como para los ausentismos, los modelos que culminaron en mejores resultados fueron *Random Forest.*

9.4.4.1 *Random search with cross validation*

En primer lugar, se crea un conjunto de hiperparámetros, el cual entrega 2*12*2*3*4*10 = 5.760 combinaciones posibles y considerando que cada *Random Forest* se demora aproximadamente tres minutos en ejecutar. Éstas tomarían aproximadamente 12 días en ejecutar, es por esto que se escogen solo 100 combinaciones de manera aleatoria, ejecutando un total de 300 *Random Forest* para estudiar el ausentismo y 300 para el atraso laboral, ya que para cada combinación se ejecuta una validación cruzada de 3 *Folds*.

Se encuentra que la forma que converge a los mejores resultados es ejecutando esta técnica buscando maximizar el f_1*-score*, para posteriormente maximizar la *precision* del modelo aplicando *Grid Search with Cross Validation*. Esta combinación de optimizaciones hace que el modelo converja a una tasa mayor de verdaderos positivos, minimizando los falsos positivos, es decir, un mejor clasificador en términos de *ROC* y *AUC*.

Modelo de atrasos

Se ejecuta esta técnica y se obtienen los siguientes resultados:

	Métricas base	**Métricas *Random Search***	**Mejora**
AUC	0,889	0,910	2,36%
Precision	0,698	0,583	-16,48%
Recall	0,445	0,648	45,62%
F1-Score	0,543	0,613	12,89%
Accuracy	0,914	0,900	-1,53%

Tabla 11: Métricas de confusión, caso base y Random Search, modelo de atrasos.

Como es de esperar, el *AUC y* f_1*-score* aumentan de manera considerable, disminuyendo la *precision* y *accuracy* del modelo.

Modelo de ausentismos

Se ejecuta esta técnica y se obtienen los siguientes resultados:

	Métricas base	**Métricas *Random Search***	**Mejora**
AUC	0,763	0,826	8,26%
Precision	0,489	0,240	-50,92%
Recall	0,026	0,155	496,15%
F1-Score	0,049	0,189	285,71%
Accuracy	0,992	0,989	-0,30%

Tabla 12: Métricas de confusión, caso base y Random Search, modelo de ausentismos.

De manera similar al caso de atrasos, esta técnica mejora considerablemente el *AUC* y el f_1-*score* del modelo, sin embargo, la precisión disminuye a la mitad.

9.4.4.2 *Grid search with cross validation*

El *Random Search* permitió reducir el rango de cada hiperparámetro, por lo que ahora se pueden concentrar los recursos en un espacio más acotado y probar todas las combinaciones de configuraciones posibles. De esta manera, se optimizan los recursos y se obtienen mejores hiperparámetros.

En esta parte, se prueban 72 combinaciones posibles para los ausentismos y los atrasos, por lo que aplicando validación cruzada de 3 *Folds*, se ejecutan 216 *Random Forest* de atrasos y 216 *Random Forest* de ausentismos.

Modelo de atrasos

A continuación, se resume la mejora que se obtiene tras usar las dos técnicas descritas anteriormente:

	Métricas base	**Métricas *Random Search***	**Métricas *Grid Search***	**Mejora respecto *Random Search***	**Mejora respecto base**
AUC	0,893	0,910	0,911	0,11%	2,02%
Precision	0,698	0,583	0,777	33,28%	11,31%
Recall	0,445	0,648	0,445	-31,33%	-0,02%
F1-Score	0,543	0,613	0,561	-8,48%	3,31%
Accuracy	0,914	0,906	0,920	1,55%	0,66%

Tabla 13: Mejora de Grid Search, modelo de atrasos.

Se observa que el *recall* y el f_1-*score* disminuyen en comparación al *Random Search.* Sin embargo, el *AUC* y la *Precision* mejoran considerablemente y todas las métricas, excepto el *recall*, aumentan en relación con el caso base, por lo que el enfoque utilizado fue bastante provechoso.

Modelo de ausentismos

De manera análoga a los atrasos, se resume la mejora obtenida tras usar las técnicas descritas:

	Métricas base	**Métricas *Random Search***	**Métricas *Grid Search***	**Mejora respecto *Random Search***	**Mejora respecto base**
AUC	0,763	0,826	0,836	1,21%	9,57%
Precision	0,489	0,240	0,600	150,00%	22,70%
Recall	0,026	0,155	0,028	-81,94%	7,69%
F1-Score	0,049	0,189	0,053	-71,96%	8,16%
Accuracy	0,992	0,989	0,991	0,20%	-0,10%

Tabla 14: Mejora de Grid Search, modelo de ausentismos.

El *recall* y el f_1*-score* disminuyen considerablemente en relación con el *Random Search*, no obstante, la *precision* aumenta cuantiosamente y todas las métricas de desempeño, a excepción del *accuracy*, mejoran en comparación al caso base, por lo que esta técnica resultó de gran utilidad.

En conclusión, el enfoque utilizado en esta sección resultó ser muy beneficioso, pues la primera maximización del *f1-score* hizo que el *recall* de los modelos aumentara, disminuyendo la *precision*, donde esta última funcionó como una cota, pues el cálculo del *f1-score*[24] incluye a la *precision.* Posteriormente, en la segunda maximización se vuelve a maximizar la *precision,* lo que incrementa la tasa de verdaderos positivos y minimiza la tasa de falsos positivos, culminando en un mejor modelos en términos de *AUC.*

[24] $\boldsymbol{F_1 = 2 * \frac{precision * recall}{precision + recall}}$

9.4.5 Análisis de sensibilidad

9.4.5.1 Modelo de atrasos

Con el objetivo de contextualizar esta sección, se recuerda que en la empresa analizada un atraso se define como aquellas instancias donde un trabajador marca su entrada al menos cinco minutos después que su turno asignado.

Para hacer el análisis de sensibilidad, se definen los atrasos de las siguientes formas:

1. El caso base se define como aquellos marcajes que ocurren 5 minutos después de la entrada del turno.
2. Aquellos marcajes que ocurren 10 minutos después de la entrada del turno.
3. Aquellos marcajes que ocurren 15 minutos después de la entrada del turno.
4. Aquellos marcajes que ocurren 20 minutos después de la entrada del turno.
5. Aquellos marcajes que ocurren 30 minutos después de la entrada del turno.
6. Aquellos marcajes que ocurren 40 minutos después de la entrada del turno.
7. Por último, se muestra la curva *ROC* del *Grid Search* con respecto al caso base.

Bajo estos escenarios se construye la curva *ROC* del modelo y se obtiene la siguiente figura:

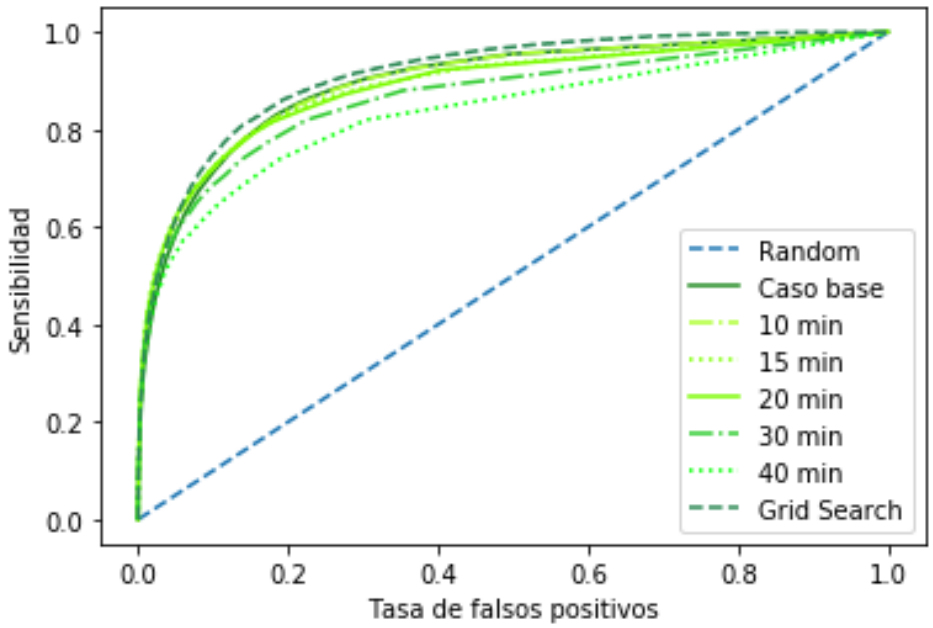


Ilustración 7: Curva ROC, análisis de sensibilidad de distintos cortes de atrasos.

Al observar la curva *ROC* de los modelos, se ve que la del *Grid Search* es superior a las demás curvas, lo que es coherente a la sección anterior, pues esta técnica mejoró en un 2% el *AUC* del modelo base.

De esta manera, se ve que a medida que se aumenta el corte de minutos de atrasos, se contrae la curva *ROC*, donde para los casos: base, 10 min, 15 min y 20 min la curva es similar y no varía a simple vista. Sin embargo, para el corte de 30 minutos la curva disminuye.

Al extraer el *AUC*, los verdaderos positivos y falsos negativos para cada escenario, se obtiene la siguiente tabla:

	AUC	Verdaderos Positivos	Falsos Negativos	**Atrasos**
Grid Search	0,914	5245	6240	11787 (11,5%)
Caso base	0,897	5245	6240	11787 (11,5%)
10 min	0,896	4016	4883	8899 (8,7%)
15 min	0,892	2750	3852	6602 (6,5%)
20 min	0,886	1859	3245	5104 (5%)
30 min	0,864	944	2406	3350 (3,28%)
40 min	0,837	528	1896	2424 (2,38%)

Tabla 15: Análisis de sensibilidad del modelo de atrasos con distintos cortes.

De la tabla se puede cuantificar de mejor manera que, a medida que descienden los casos de atrasos, también lo hace el *AUC*, donde la mayor sensibilidad se ve entre los 20 y 30 minutos, pues la cantidad de atrasados se reduce en un 34%, lo que hace que el modelo pierda poder estadístico. Esto se explica porque el modelo aprende de mejor manera cuando se le suministra una mayor cantidad de observaciones de la clase positiva (atrasado), sobre todo cuando estas observaciones representan una mayor porción de la muestra total.

Por otro lado, el predictor estima la probabilidad de atraso a nivel diario, por lo que es interesante analizar el comportamiento del modelo en horizontes de tiempo más extensos, por ejemplo, a nivel semanal y mensual.

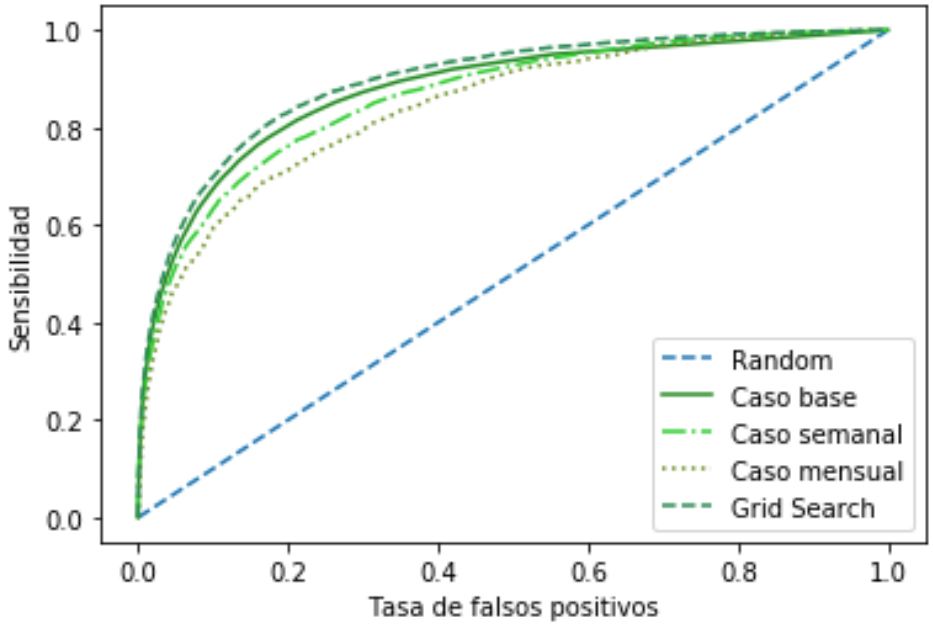


Ilustración 8: Curva ROC, análisis de sensibilidad del modelo de atrasos con distintos horizontes de tiempo.

De la figura se observa que, a pesar de que se agregan las observaciones por semana y por mes, el desempeño del modelo empeora, lo que se visualiza de mejor manera en la siguiente tabla:

	AUC	Verdaderos Positivos	Falsos Negativos	**Atrasos**	**Obs. Totales**
Grid Search	0,914	5245	6240	11787 (11,55%)	102051
Caso base	0,897	5245	6240	11787 (11,55%)	102051
Caso semanal	0,861	3457	2282	5739 (28,63%)	20045
Caso mensual	0,842	1805	608	2413 (52,65%)	4583

Tabla 16: Análisis de sensibilidad del modelo de atrasos con distintos horizontes de tiempo.

En los escenarios semanal y mensual, si bien la cantidad de observaciones de atrasos aumentan con relación a las observaciones totales, estas últimas disminuyen en un 80% y 96% respectivamente. Esto provoca que el modelo no pueda aprender de manera correcta la diferencia entre una observación de atraso y una puntual, lo que se proyecta en su curva *ROC*, donde la tasa de falsos positivos aumenta rápidamente. Una manera de mejorar esto sería establecer un corte más estricto para asignar una observación a la clase atrasado, por ejemplo: que en el escenario semanal se defina como atrasado a aquellas observaciones que tienen más de dos atrasos por semana; mientras en el escenario mensual ese criterio sea de cuatro atrasos.

9.4.5.2 Modelo de ausentismos

Para el modelo de ausentismo, se analiza la sensibilidad del modelo agregando los ausentismos injustificados de forma semanal y mensual.

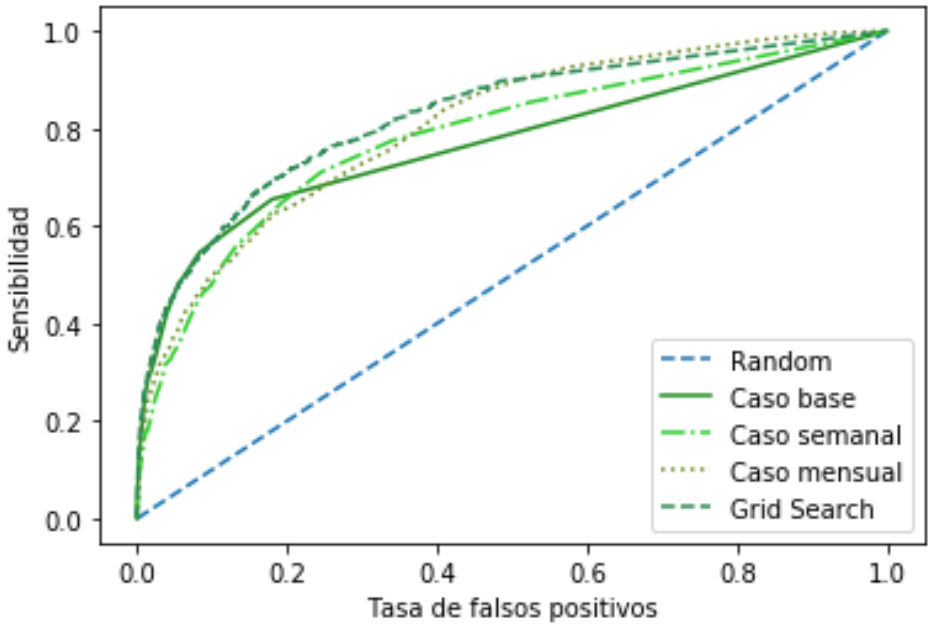


Ilustración 9: Curva ROC, análisis de sensibilidad del modelo de atrasos con distintos horizontes de tiempo.

De la ilustración, se observa que la curva *ROC* es mayor en el caso semanal y mensual con respecto al caso base. No obstante, esta mejora se concentra a medida que la tasa de falsos positivos se acerca a uno, lo que significa que el modelo aumenta su tasa de verdaderos positivos a costa de una mayor tasa de falsos positivos.

	AUC	Verdaderos Positivos	Falsos Negativos	**Ausencias Injustificadas**	**Obs. Totales**
Grid Search	0,836	47	1552	1599 (1,58%)	101200
Caso base	0,763	42	1557	1599 (1,58%)	101200
Caso semanal	0,790	36	625	661 (3,41%)	19384
Caso mensual	0,804	104	412	516 (12,69%)	4066

Tabla 17: Análisis de sensibilidad del modelo de ausentismos con distintos horizontes de tiempo.

Al comparar los casos semanales y mensuales, se observa que el *AUC* es mayor que en el caso base, ya que los ausentismos son tan atípicos que para el modelo es más fácil predecir este fenómeno de forma agregada que a nivel individual. Además, para el caso mensual, la frecuencia de ausencias injustificadas aumenta considerablemente.

Por otro lado, es interesante ver que en ningún caso el *AUC* es mayor al del *Grid Search*, pero es probable que, aplicando esta técnica metodológica para el caso semanal y mensual, se pueda mejorar el desempeño del modelo.

Resumen etapa minería de datos

En esta sección se obtuvo que el *Random Forest* es el modelo que mejor estima los atrasos y ausentismos. Este modelo presenta gran eficiencia en el manejo de grandes bases de datos; variables de distinto tipo; robustez ante valores perdidos y atípicos gracias al ensamble de árboles de decisión; y flexibilidad en la configuración de hiperparámetros. De esta manera, permitió optimizar sus métricas de desempeño, llegando a métricas buenas en el caso de los atrasos (*AUC* de 0,91). Sin embargo, a pesar de que este también fue el mejor modelo para estimar ausentismos, la singularidad de estos provocó que los modelos estudiados no pudiesen ajustarse de buena forma. Por tanto, este modelo podría haber obtenido mejores resultados agregando otras variables características más correlacionadas al ausentismo y un tamaño de muestra superior.

Entre las variables que mayor importancia le asignan los modelos de ausentismo y atrasos, destacan la distancia, antigüedad, turno, terminal, estado civil, cargo y precipitaciones. Esto permite suponer que los modelos mejoren su desempeño agregando variables correlacionadas a las anteriormente mencionadas como: estudios, salario, número de hijos, edad de los hijos, forma de transporte, tiempo de traslado, amonestaciones y enfermedades crónicas.

El enfoque usado en la sección de *Hyperparameter tunning* resultó ser muy beneficioso, pues al maximizar el f_1*-score* en primera instancia y maximizar la *precisión* en segundo lugar, ayudó a mejorar el *ROC* y *AUC* del modelo en la dirección correcta. Esto se debe a que la primera maximización con *Random Search* aumenta el *recall* del modelo en desmedro de la *precision*, por lo que incrementó la tasa de falsos positivos. No obstante, al maximizar la *precision* con *Grid Search*, se maximiza la tasa de verdaderos positivos, minimizando la tasa de falsos positivos, lo que culmina en una curva *ROC* más cercana al punto (0,1) y un *AUC* superior.

En el análisis de sensibilidad, se estudia la robustez que tiene el modelo de atrasos, donde se observa que es bastante robusto para los cortes de 10, 15 y 20 minutos, pero pierde poder para los cortes de 30 y 40 minutos. Esto se explica porque los casos de atrasos para estos intervalos disminuyen en gran medida con respecto al caso base. Por otro lado, al analizar la sensibilidad del modelo a nivel agregado semanal y mensual, el modelo pierde rápidamente su poder, pues la similitud de las observaciones de atraso y no atraso, no permiten al modelo diferenciar cada caso. Al estudiar la sensibilidad del modelo de ausentismos, el modelo gana poder estadístico para los escenarios semanal y mensual con respecto al caso base, ya que resulta muy complejo calcular la irregularidad del ausentismo a nivel individual. Además, el análisis permite vislumbrar que es posible crear un modelo de predicción de tendencias de atraso y ausentismo de trabajadores, para que sea usado en períodos de contratación, con el fin de filtrar a aquellos candidatos con mayores probabilidades a atrasarse y/o a ausentarse.

Por último, cabe destacar que, para la correcta implementación del modelo de atrasos en una herramienta de apoyo a los jefes de tienda, este se deberá validar con data actualizada y calibrar de manera semanal, de esta manera el modelo se ajustará de mejor manera a los datos nuevos, sin sobreajustarse, y no significará un costo muy grande debido a que la calibración se realizaría de manera automática.

10 Productos complementarios afines

10.1 Optimización de distancia de los trabajadores a la sucursal de destino

A partir de la sección de análisis descriptivo y la importancia de variables indicada en la sección de minería de datos, se desprende que una de las variables que más se relaciona a los ausentismos injustificados y atrasos, es la distancia del trabajador a la sucursal de destino. En efecto, en conversaciones con jefes de tienda de la empresa analizada, reportan que uno de los factores por los que más se excusan los empleados por sus atrasos y ausentismos es por el traslado hacia el trabajo.

Bajo lo expuesto, se propone relocalizar de manera óptima a los trabajadores, minimizando las distancias de éstos con la sucursal de destino.

Además, al hacer un conteo simple de los trabajadores que tienen otra sucursal que esté ubicada entre 0 – 15 km de su domicilio y que en ésta exista el cargo del trabajador, se obtiene lo siguiente:

Distancia de domicilio al trabajo[25]	***Porcentaje de trabajadores a los que es posible reubicar en otra sucursal ubicada entre 0 -15 km***
20 - 30 km	41%
15 - 20 km	62%
10 - 15 km	53%
5 - 10 km	39%
0 - 5 km	35%

Tabla 18: Porcentaje de trabajadores reubicables a otras sucursales.

Esta aplicación es muy interesante de analizar, pues no solo contribuiría a disminuir las tasas de atraso y ausentismo, sino que, de acuerdo con Chatterjee et al. (2017), también contribuiría en mejoras de la satisfacción laboral, calidad de vida y retención de los trabajadores.

A continuación, se deja plantea el Problema de Programación Lineal recomendado que podría ayudar a optimizar las distancias:

Planteamiento matemático:

Conjuntos:

- *Personas:* $\boldsymbol{i} = \{i_1, i_2 \dots i_n\}$*, son todos los trabajadores de la empresa.*
- *Sucursales:* $\boldsymbol{j} = \{Santiago_1, Santiago_2, \dots, Antofagasta, Temuco\}$*, son todas las sucursales de la empresa.*

[25] Corresponde a la distancia actual del domicilio de los trabajadores al trabajo.

Parámetros:

- $\boldsymbol{P0_{i \in Personas, j \in Sucursales}}$, *este parámetro representa la sucursal inicial j a la que está asignada cada persona i.*
- $\boldsymbol{Dist_{i \in Personas, j \in Sucursales}}$, *este parámetro representa la distancia de la persona i con cada sucursal j.*
- $\boldsymbol{Disponibilidad_{i \in Personas, j \in Sucursales}}$, *este parámetro indica si la persona i está disponible a ser asignada a la sucursal j. Es decir, si el cargo de la persona i existe en la sucursal j.*
- $\boldsymbol{Dda_{j \in Sucursales}}$, *este parámetro indica la demanda de trabajadores en cada sucursal.*
- $\boldsymbol{FR}$, *este parámetro indica un factor de relocalización, que funciona como ponderador para no relocalizar demasiado lejos a trabajadores que ya viven cerca de su sucursal asignada.*

Variable:

- $\boldsymbol{X_{i \in Personas, j \in Sucursales}}$, *variable binaria que es 1 si la persona i es asignada a la sucursal j y 0 si no. Indica la posición final de la persona i.*

Función objetivo:

$$\boldsymbol{min} \sum_{i \in Personas, j \in Sucursales} \boldsymbol{X}_{i \in Personas, j \in Sucursales} * \boldsymbol{Dist}_{i \in Personas, j \in Sucursales}$$

Restricciones:

- $\boldsymbol{X_{i \in Personas, j \in Sucursales}} \in \{\boldsymbol{0, 1}\}$
- $\boldsymbol{Disponibilidad_{i \in Personas, j \in Sucursales}} \geq \boldsymbol{X_{i \in Personas, j \in Sucursales}}, \forall\, \boldsymbol{i, j} \in \boldsymbol{Personas, Sucursales}$
- $\sum_{\boldsymbol{j \in Sucursales}} \boldsymbol{X_{i \in Personas, j \in Sucursales}} = \boldsymbol{1},\ \forall\, \boldsymbol{i} \in \boldsymbol{Personas}$
- $\sum_{\boldsymbol{j \in Sucursales}} (\boldsymbol{X_{i \in Personas, j \in Sucursales}} * \boldsymbol{Dist_{i \in Personas, j \in Sucursales}}) \leq \sum_{\boldsymbol{j \in Sucursales}} (\boldsymbol{P0} * \boldsymbol{Dist_{i \in Personas, j \in Sucursales}} * \boldsymbol{FR}), \forall\, \boldsymbol{i} \in \boldsymbol{Personas}$
- $\sum_{\boldsymbol{i \in Personas}} (\boldsymbol{X_{i \in Personas, j \in Sucursales}} - \boldsymbol{P0_{i \in Personas, j \in Sucursales}}) = \boldsymbol{Dda_{j \in Sucursales}}, \forall\, \boldsymbol{j} \in \boldsymbol{Sucursales}$

El conjunto solución de este problema se puede calcular utilizando la librería *Gurobi*[26] en *Python*. Este trabajo no contempla la resolución de éste, por lo que se instala como una propuesta para un proyecto futuro.

[26] https://www.gurobi.com/documentation/9.0/quickstart_mac/py_python_interface.html

10.2 Predictor de tendencia de ausentismo y atraso

De la base de datos creada, se pueden extraer variadas aplicaciones. Una de estas es un predictor de tendencias de ausentismos y atrasos laborales. De esta manera, se puede crear un modelo de predicción de tendencias de atraso y ausentismo de potenciales trabajadores, al llevar los atrasos y ausencias injustificadas hacia un horizonte de evaluación de un mes y silenciando ciertas variables, como: antigüedad, atraso en t-1, licencia en t-1, ausencias en t-1 y permisos en t-1 (todas las variables que dependan de la historia del trabajador en la empresa). Estos modelos se podrían utilizar en períodos de contratación, con el objetivo de tener una variable más para agregar al proceso de selección.

Respecto a la construcción de estos modelos, se entrena con el 80% de los trabajadores y se testea con el 20%, cuidando que las observaciones de entrenamiento y testing estén balanceadas en la misma proporción.

Para estimar la tendencia de ausentismo, se considera que una observación tiene tendencia de ausentismo si se ausenta al menos una vez de manera injustificada en un periodo de un mes.

El modelo construido corresponde a un *Random Forest*, que entrega un *AUC* de 0,75 y, considerando un umbral de corte de 0,5, resulta en las siguientes métricas:

Clase	***Precision***	***Recall***	***F1-Score***
Tendencia no faltar	0,91	0,99	0,95
Tendencia ausentismo	0,64	0,21	0,31

Tabla 19: Métricas de desempeño, Random Forest de tendencia de ausentismo, umbral de corte de 0,5.

Por otra parte, para estimar la tendencia de atraso, se considera que una observación tiene tendencia de atraso si registra al menos cuatro atrasos en un periodo de un mes.

El modelo construido corresponde a un *Random Forest*, que entrega un *AUC* de 0,90 y, considerando un umbral de corte de 0,5, resulta en las siguientes métricas:

Clase	***Precision***	***Recall***	***F1-Score***
Tendencia puntualidad	0,88	0,96	0,92
Tendencia atraso	0,78	0,56	0,66

Tabla 20: Métricas de desempeño, Random Forest de tendencia de atraso, umbral de corte de 0,5.

Las métricas señaladas indican que estos modelos podrían servir como apoyo a la empresa en etapas de contratación.

11 Discusión

Con el objetivo de contextualizar el problema en la empresa analizada, se calcula que los costos asociados a la mano de obra son, en promedio, el 30% de las utilidades relacionadas a la actividad directa de esta. Por tanto, la rentabilidad sobre la hora trabajada es de un 233%.

En promedio ocurren 75 atrasos al día (12,2% del total de marcajes diarios), donde el atraso promedio es de 32 minutos[27], lo que se traduce en 21,3 FTE[28] perdidos en un mes, es decir, cerca de un 1% de la planificación total del mes[29] y 256 FTE perdidos al año. Cabe destacar que, si bien los atrasos se descuentan del salario de los trabajadores, las rentabilidades que se podrían haber generado con el trabajador en su puesto de trabajo son difíciles de recuperar.

Por otro lado, diariamente se ausentan nueve personas de manera injustificada en promedio (1,5% del total de marcajes), lo que se traduce en aproximadamente 33,66 FTE[30] perdidos en un mes, es decir, cerca de un 1,5% de la planificación total del mes y 403,2 FTE al año.

El *output* del modelo de atrasos puede servirles a los jefes de cada sucursal para planificar de mejor manera las tareas a realizar. En consecuencia, cuando se detecta alta probabilidad de atraso de un grupo de trabajadores, podrá reorganizar a los trabajadores, otorgar mayor prioridad a aquellas tareas más urgentes o bien citar a una persona extra para cubrir eventuales vacíos en la planificación. Para ilustrar este punto, el promedio de atrasos diarios es de 12,2%, lo que se traduce en 40 horas en atrasos al día. Bajo estas condiciones, el modelo de atrasos calcula que en promedio se atrasan 43 personas diarias, de las cuales el 78% son atrasos reales (34 personas). La empresa puede reorganizar los turnos del 5% del total de los trabajadores, con esto podría reorganizar a 31 trabajadores que se detecten con alta probabilidad de atraso y, bajo el supuesto que no se generan atrasos adicionales, esta reorganización implicaría un ahorro del 32% de las horas pérdidas en atrasos[31].

En la misma línea, el *output* de los modelos, es decir, la probabilidad de atraso y/o ausentismo se puede utilizar como una restricción blanda que penalice en un *Scheduler*[32], de manera que el planificador reubique a los trabajadores en horarios y días donde la probabilidad de atraso y ausentismo disminuya. Esto reduciría considerablemente la carga de trabajo de los jefes de tienda y el único costo adicional en el que incurriría la empresa si adopta esta medida, sería el del desarrollo e integración de esta restricción en un planificador de horarios.

---

[27] Recordar que se define como atraso a aquellos marcajes de entrada que ocurren al menos 5 minutos después del turno asignado, es decir, las demoras entre 0 y 5 min no se consideran como atrasos.

[28] $\frac{75pp*32min*24días}{60min} = 986\frac{horas}{mes} = 22\ FTE/mes$

[29] $planificación\ mensual = (610pp * 25\% * 4horas * 24días + 610pp * 75\% * 8horas * 24días)/45 = 2.277\ FTE/mes$

[30] Al mes, de los que se ausentan 25% son part time y 75% son full time => $9pp * 25\% * 4horas * 24días + 9pp * 75\% * 8horas * 24días = \ 33{,}6\ FTE$

[31] $\frac{31*78\%}{75} = 32\%$

[32] Modelo de optimización que planifica el horario de los trabajadores.

Por otra parte, el *output* de este modelo se puede conectar a un módulo que envíe mensajes a los trabajadores cada vez que se detecte que llegarán atrasados. De esta manera, los trabajadores podrán cambiar su comportamiento a tiempo para no registrar un atraso más en su historial. El efecto que podrían tener estos mensajes sobre el comportamiento de los trabajadores se deja propuesto para ser analizado en un experimento futuro.

Finalmente, en el caso del modelo de ausentismos, las métricas de desempeño sugieren que este modelo es de menor utilidad en cuanto al apoyo para los jefes de tienda en la planificación de personal diario. No obstante, como se observa en la sección anterior, las métricas mejoran considerablemente para un horizonte de tiempo de un mes y silenciando ciertas variables. Por tanto, el área de recursos humanos, durante el proceso de contratación, podría utilizar este modelo para identificar si un trabajador será propenso a faltar o no. Este mismo uso se le puede dar al modelo de atrasos, silenciando las variables que estén relacionadas con la historia pasada del trabajador[33].

[33] Antigüedad en la empresa, atraso en t-1, licencia en t-1, ausencias en t-1 y permisos en t-1.

12 Conclusiones

Las conclusiones derivadas de los objetivos y resultados de este trabajo son las siguientes:

En primer lugar, mediante revisión bibliográfica, uso de estadígrafos básicos y extracción de importancia de variables por los modelos de *Random Forest* y *XGBoost* usados para estimar los ausentismos y atrasos, se desprende que las variables más relevantes se relacionan de manera similar para ambos casos. Estas variables son: la distancia al trabajo, edad, antigüedad, cargo, estado civil y precipitaciones, lo que permite entender estos fenómenos para identificar acciones que pueden mitigar el efecto negativo de estos. Lo anterior, junto con la literatura revisada, permite suponer que los modelos pueden mejorar su desempeño agregando otras variables características como: los estudios, salario, número de hijos, forma de transporte, tiempo de traslado y enfermedades crónicas, ya que se relacionan con las variables con mayor importancia de los modelos.

En segundo lugar, se determinan distintos modelos predictivos basados en *Machine Learning* para estimar ausentismos y atrasos laborales. Estos dos fenómenos de comportamiento laboral se modelan utilizando: *Logistic Regression*, *Naive Bayes*, *Gradient Boosting*, *Decision Tree*, *Random Forest* y *XGBoost*, donde el mejor modelo para ambos casos resulta un *Random Forest*, el cual es mejorado con técnicas metodológicas de *tunning* de hiperparámetros, *Random Search with Cross Validation* y *Grid Search with Cross Validation* que resultaron ser muy valiosas, pues se obtuvieron grandes mejoras en el *AUC*, *recall* y *precision* de los modelos.

El *Random Forest* de atrasos entrega un *AUC* de 0,91 y bajo un umbral de corte de 0,5 alcanza un 78% de *precision* y 45% de *recall*. Esto lo convierte en un modelo robusto y con alto poder estadístico, por lo que el *output* de este modelo se puede integrar en una herramienta de apoyo a los jefes de tienda, que entregue la probabilidad de atraso de los trabajadores.

El *Random Forest* de ausentismos entrega un *AUC* de 0,84 y bajo un umbral de corte de 0,5 alcanza un 60% de *precision* y 3% de *recall*. A partir de los resultados, no se puede afirmar que es un modelo con alto poder estadístico capaz de servirle a los jefes de tienda en la planificación diaria, no obstante, si se desprenden otras aplicaciones valiosas para este modelo.

Los *outputs* de ambos modelos pueden ser usados en herramientas de visualización para los jefes de tienda, para que estos puedan tomar medidas correctivas si se detecta una alta probabilidad de atraso y/o ausentismo de un grupo de colaboradores, donde pueden reorganizarlos, citar personal extra o replanificar las tareas con mayor urgencia, de manera de aumentar la resiliencia del sistema de planificación. Bajo las condiciones expuestas en la sección de discusión, la reorganización de turnos podría implicar un ahorro del 32% de las horas pérdidas en atrasos, por otro lado, para el caso de los ausentismos, el bajo desempeño del modelo no permite calcular un ahorro directo a partir de la estimación de ausentismos diarios. Asimismo, los *outputs* de los modelos se pueden utilizar como una restricción blanda que penalice en un planificador de horario, de manera que este asigne a los trabajadores el horario y día tal que minimice las probabilidades de atraso y ausentismo, además de reducir la carga de trabajo de los jefes de tienda, esto sería un proceso automático que no requeriría de ningún costo adicional.

Frente a lo expuesto, este estudio ayuda a identificar las variables más relevantes que influyen tanto en los ausentismos como en los atrasos laborales, lo que permite determinar acciones para disminuir los efectos negativos de estas malas prácticas en el ambiente laboral. Por ejemplo, implementar, estratégicamente, buses de acercamiento para los trabajadores que vivan lejos de las sucursales y/o relocalizarlos a sucursales más cercanas de sus domicilios, mediante el problema de programación lineal que se plantea en la sección de productos complementarios afines.

Por otro lado, a partir de los estudios realizados, se determinan modelos de estimación de tendencia de atraso y tendencia de ausentismo que pueden ser usados en épocas de contratación con el objetivo de generar más información de los empleados a contratar.

Finalmente, se declara el cumplimiento del objetivo general, tras el desarrollo de un modelo de predicción que estima inasistencias y un modelo que estima atrasos laborales, que, junto con la aplicación propuesta y acciones correctivas planteadas en las recomendaciones, ayudan a disminuir los efectos negativos de estos fenómenos y aumentan la potencia y resiliencia del sistema de planificación.

13 Recomendaciones comerciales

En conversaciones con jefes de tienda, reportan que el traslado hacia el trabajo es uno de los factores por los que más se excusan los empleados por sus atrasos y ausentismos. De acuerdo con estudios internos realizados en la empresa, esto explica gran parte del descontento de los trabajadores. Asimismo, en este trabajo se detecta que la distancia es una de las variables con mayor importancia en los ausentismos y atrasos laborales, por lo que sería interesante analizar distintas alternativas para mitigar el efecto negativo de la distancia sobre los trabajadores. Por ejemplo, integrar buses de acercamiento o relocalizar de manera óptima a los trabajadores, minimizando las distancias de éstos con la sucursal de destino.

Si la empresa optara por buses de acercamiento, entonces incurrirían en costos fijos y probablemente en costos de logística, que podrían terminar perjudicando aún más el problema. Debido a esto, una solución más viable podría ser una relocalización de los trabajadores en las sucursales, pues en períodos de contratación, los trabajadores son asignados a las sucursales con vacantes disponibles y no necesariamente a la más cercana a su domicilio. Esto provoca que una gran cantidad de trabajadores estén localizados en sucursales que les quedan considerablemente más lejos que otras.

Como se expuso en la sección de productos complementarios, de los trabajadores que se encuentran entre 15 – 30 km de sus sucursales se puede relocalizar a aproximadamente un 49% de ellos en sucursales localizadas entre 0 – 15 km de sus domicilios.

Bajo el supuesto que se puedan relocalizar entre 0 – 15 km a un 30% de los trabajadores que vivan entre 15 – 30 km, se obtendrían los siguientes cambios en las tasas de atraso y ausentismo:

1. La tasa de atraso del 30% de los que viven entre 15 – 20 km, descendería un 30%[34] y la tasa de ausentismo descendería un 19%[35].
2. La tasa de atraso del 30% de los que viven entre 20 – 30 km descendería un 34%[36] y la tasa de ausentismo descendería un 21%[37].

Es importante mencionar que el promedio de atraso de los trabajadores que viven entre 15 – 20 km es de 35 minutos, mientras que para los trabajadores que viven entre 0 – 15 km es de 29 minutos. Este escenario, además de reducir los atrasos en cuatro personas diarias, disminuiría la media de los minutos de atraso totales de 32 minutos a 31 minutos. Esta reducción en los atrasos significaría 56 horas menos de atrasos al mes, es decir, 15 FTE al año. Asimismo, se ausentarían aproximadamente dos personas menos por semana[38], 10 personas menos al mes y 115 personas menos al año (20 FTE al año).

[34] La tasa de atrasos descendería de 14,6% a 10,3%.

[35] La tasa de ausentismo descendería de un 1,6% a 1,3%.

[36] La tasa de atrasos descendería de un 15,5% a un 10,3%.

[37] La tasa de ausentismo descendería de un 1,64% a 1,3%.

[38] La reducción es de 0.4 personas diarias.

De forma agregada, esta reducción de atraso y ausentismo traería una reducción del 5%[39] en atrasos y 2,7% en ausentismos, lo que se traduciría en un aumento de 33 FTE al año. También se recuperarían las rentabilidades asociadas directamente a estas horas, lo que es equivalente a 110 FTE en rentabilidades. Esto ayudaría a reducir las horas extra[40], donde si se toma el coeficiente de correlación del MCO planteado en la sección de análisis descriptivo, se ahorrarían 152 FTE en horas extra, junto con contribuir con mejoras en la satisfacción laboral, calidad de vida y retención de los trabajadores (Chatterjee et al., 2017).

Otra variable importante que explica los atrasos es la antigüedad de la persona en la empresa, pues aquellos con más trayectoria laboral tienden a atrasarse con mayor frecuencia, pues con la experiencia los trabajadores aprenden que los atrasos rara vez culminan en causal de despido. Para solucionar esto, se recomienda la integración de incentivos a los trabajadores más experimentados, para que no lleguen tarde ni incurran en otras malas prácticas. Por ejemplo, otorgar un bono de productividad individual que se le asigne a todos los trabajadores que lleven más de 7 años en la empresa y que no tengan salidas tempranas, licencias, atrasos y ausentismos en el período de pago.

Otras de las variables importantes que se analizaron, son las precipitaciones, pues se observa una gran relación con los atrasos y ausentismos. Si bien corresponden a un fenómeno natural inevitable, se pueden tomar medidas para mitigar los efectos negativos en la empresa. Por ejemplo, permitir a los trabajadores que lleguen 30 minutos más tarde los días que se detecte una alta probabilidad de lluvia, pues de esta manera los trabajadores no incurrirían en tantos descuentos de salario y se generarían menos horas extra.

En la misma línea, cabe destacar que hay una gran diferencia entre los atrasos y ausentismos según el cargo y el terminal de marcaje, por lo que las nuevas medidas para mitigar estos fenómenos deberían concentrarse primordialmente en los cargos y sucursales con mayores tasas de atrasos y ausentismos.

Actualmente, la hora extra se define como los minutos de diferencia entre las marcas de salida que ocurren 15 minutos después del turno de salida, independiente si el trabajador llegó atrasado o no. Como se menciona anteriormente, los trabajadores que se atrasan tienden a generar más horas extra, pues en algunos casos esto es un incentivo, pues a pesar de que les descuentan los atrasos, los bonifican entre un 25% y 30% por cada hora extra. Por tanto, una medida que mitigaría este problema sería redefinir las horas extra como los minutos por sobre el contrato que se generan si el trabajador marca su salida al menos 30 minutos después de su salida de turno. Esta redefinición habría disminuido las horas extra pagadas en 2019 en un 66%[41].

Finalmente, en la sección de productos comerciales afines, se muestran dos predictores de tendencias de ausentismo y atraso laboral. Se recomienda a la empresa que estos modelos sean utilizados para obtener una variable más a considerar en épocas de contratación, pues

[39] La tasa de atrasos agregada reduciría de 12.2% a 11.6%.

[40] Se considera como hora extra a todas aquellas instancias donde la marca de salida ocurre al menos 30 minutos después que la salida de turno.

[41] El año 2019, la empresa pagó 30.305 horas extra a 650 trabajadores y bajo el concepto propuesto solo se habrían pagado 10.303 horas extra.

uno de los problemas de la empresa analizada es que los trabajadores más nuevos tienden a ausentarse y atrasarse en mayor proporción que el resto[42]. Si bien estos modelos pueden ser de bastante utilidad, no se recomienda que el *output* de estos modelos se utilice como único motivo de exclusión o inclusión de una persona a la fuerza de trabajo, sino que como una variable más a considerar en la toma de decisión.

[42] Los trabajadores entre 0 – 1 año tienden a ausentarse y atrasarse un 3% y 14% de los días respectivamente.

14 Trabajos futuros

Debido a las protestas sociales desarrolladas en el marco de la crisis social desatada en Chile en octubre de 2019 y los desafíos que ha traído la expansión de la pandemia del COVID-19 a nivel mundial, los modelos construidos no se probaron con data nueva a partir del 11 de octubre de 2019. Por tanto, en trabajos futuros se proyecta entrenar y calibrar los modelos con la data ya existente y testearlos con datos nuevos, para evaluar el cambio de las métricas de desempeño de dichos modelos.

Adicionalmente, sería interesante probar los modelos con empresas de distintas industrias. Si bien las tasas de atraso y ausentismo laboral en la empresa elegida son altas, existen otras compañías con tasas mucho mayores, lo que probablemente ayudaría a estimar el fenómeno de mejor manera.

A pesar de que los modelos construidos en este trabajo se mejoran usando técnicas metodológicas, éstos pueden usarse como línea de base para seguir mejorándolos. En este contexto, en la literatura existen problemas similares abordados con modelos de aprendizaje profundo (*Deep Learning*) que servirían como punto de comparación y/o mejora.

Por otro lado, es interesante analizar el efecto que podrían tener mensajes direccionados a los individuos con mayor probabilidad de atrasarse/faltar. Para esto, se deja propuesto diseñar y ejecutar un experimento que permita evaluar diferencias significativas en el comportamiento de los trabajadores sometidos al estímulo (mensaje) contra un grupo control. También, se podría evaluar un experimento aleatorio para modelar el *uplift* de los trabajadores.

Entre las causas más frecuentes de ausentismo justificado son las licencias médicas, que generan horas pérdidas y espacios vacíos en la planificación que se terminan rellenando con horas extra. Cabe destacar que la frecuencia de estas excepciones es mucho mayor a la de los ausentismos injustificados, por lo que probablemente sea menos complejo estimarlas en un trabajo futuro.

Este estudio se enfoca en modelos de predicción de ausentismo y atraso a nivel individual, pero para trabajos futuros se plantea desarrollar otros modelos que consideren la predicción de ausentismos y atrasos a nivel agregado. De esta forma, el *output* podría usarse para saber cuántos trabajadores extra planificar y así construir un sistema de trabajo más resiliente a este tipo de malas prácticas.

Bibliografía

ACEC; Accenture; AED. (2019). *Machine Learning, Inteligencia Artificial y Big Data.*

Badubi, R. M. (2017). A Critical Risk Analysis of Absenteeism in the Work Place. *Journal of International Business Research and Marketing, 2(6),* 32 - 36.

Bell, D. (2018). How can Absentee Leaders ruin your Organisation? *Linkedin*.

Bellazini, R., & Zupan, B. (2008). *Predictive data mining in clinical medicine: current issues and guidelines.* International Journal of Medical Informatics.

Bengio, Y., & Bergstra, J. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*.

Breiman. (2001). Random Forests. En *Machine Learning* (págs. 5-32).

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.*

Chatterjee, K., Clark, B., Martin, A., & Davis, A. (2017). *The Commuting and Wellbeing Study: Understanding the Impact of Commuting on People's Lives. UWE, UK*. Bristol, UK.

Chen, T., & Carlos , G. (2016). *XGBoost: A Scalable Tree Boosting System.*

Codigo del trabajo. (Edicion 2020). *Articulo 160, inciso 3.* Dirección del trabajo, Gobierno de Chile.

Cucchiella, F., Gastaldi, M., & Ranieri, L. (2014). Managing absenteeism in the workplace: the case of an Italian Multiutility Company. *Procedia - Social and Behavioral Sciences, 150,* 1157 - 1166.

Elith, J., Leathwick, J., & Hastie, T. (2008). A working guide to boosted regression trees. *British Ecological Society*.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1991). From Data Mining to Knowledge Discovery in Databases. *AI Magazine,* 37-54.

Fernandez, P. (1996). Investigación: Determinación del tamaño muestral. *Cad Aten Primaria, 3,* 138-142.

Fernández, S. d. (2011). Regresión Logística. Madrid.

Ferreira, R. (2018). Artificial Neural Network And Their Application In The Prediction of Absenteeism At Work.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. En *Annal of Statistics, Volumen 29, Número 5* (págs. 1189-1232).

Gayathri, T. (2018). Data mining of absentee data to increase productivity. *International Journal of Engineering and Techniques*.

Géron, A. (2019). Better Evaluation Using Cross-Validation. En *Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow* (págs. 76-79). Second Edition.

Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow. Segunda edición.

Hastie, T., Tibshirani , R., & Friedman, J. (2009). *Elements of Statistical Learning.* Ed. 2, Springer.

Investopedia. (2013). *Forbes*.

Jensen, T., & Sun, Q. (2013). *Absenteeism Prediction and Labor Force Optimization in Rail Dispatcher Scheduling. Tesis de Magister*.

Korkki, P. (2007). For the Chronically Late, It's Not a Power Trip. *The New York Times*.

Linkedin. (2016). *Millennials job-hop more than previous generations, but they'll slow down eventually*.

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice.*

Martiniano, A., Ferreira, R., Sassi, P., & Affonso, C. (2012). *Application of a neuro fuzzy network in prediction of absenteeism at work.*

Monjo, R. (2010). El índice n de la precipitación intensa. *Fundación para la investigación del Clima*.

Monjo, R. (2010). El índice n de la precipitación intensa. *Divulga Meteo*.

Natter, E. (2018). How Employees With Poor Attandance Affect the Workplace. *Chron*.

Nunung, N., Qomariyah, Y., & Sucahyo, G. (2014). *Employees attendance patterns prediction using classification algortihm case study: a private company in Indonesia.*

Va, A. (2014). Employee Absences Have Consequences for Productivity and Revenue, SHRM Research Shows. *shrm.org*.

Zhang, H. (2004). *The optimality of Naive Bayes.* New Brunswick.

Apéndice A

En esta sección se muestra una ficha ejemplo de cómo se visualiza la base de datos, Empleados,

A.1 Motivo de Ausencia

Código	Paycode
1	Acc, Trabajo
2	Vacaciones
3	Maternal
4	Licencia
5	Comisión de servicio
6	Compensación día feriado
7	Permiso
8	Permiso nacimiento
9	Compensación dia normal
10	Conciliación familiar
11	Dia administrativo
12	Permiso fallecimiento
13	Permiso recuperado
14	Compensación dia domingo
15	Permiso matrimonio
16	Dia festivo
17	Capacitacion SC
18	Paseo de fin de año

Fuente: Elaboración propia

A.2 Ficha de persona ficticia de la base *Empleados*:

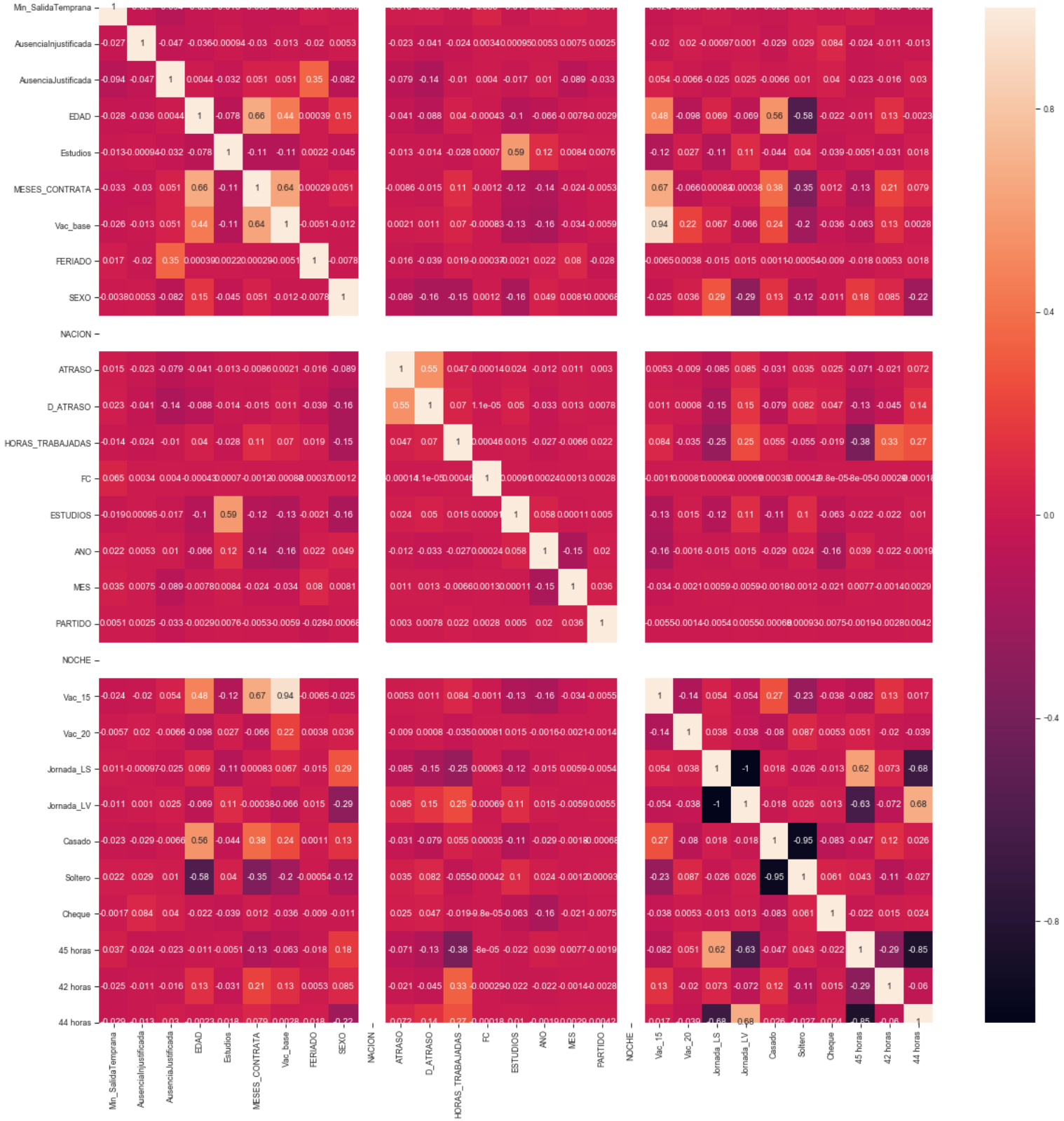
Atributo	Descripcion
Rut	12,345,678-9
ID	12345678
APATERNO	Perez
AMATERNO	Perez
Nombres	Juan Franco
Descripcion Cargo	Ayudante Granel San Fdo,
Descripcion Centro de Costo	Transporte Granel OSF
Sexo	Masculino
Direccion	Montevideo 143 0161 73Antofagasta 5 3
Fecha de Retiro	Fecha de abandono de la empresa
Fecha nacimiento	18-10-1989
Nacion	4
Est_civil	Soltero/a
Estudios	5
Titulo	9999
Clase	1
Idiomas	161
Renmax	0
Fecha de ingreso	22-10-1979
Tipo de documento	9999
Jefe	23,567,890-1
Fecha retiro	30-05-2014
Descripcion Horario	Articulo 22
Descripcion lPago_Descrip	Maipu - Planta Maipu

Descripcion Unidad	Gerencia de administración
Descripcion AFP	Cuprum
Descripcion Isapre	Fonasa
Descripcion Sindicato	No sindicalizado
Descripcion Division	Proyectos Tecnologias de Informacion
Descripcion Clasificacion	Subgerencia de Tecnologias de Informacion
Descripcion Convenio	Contrato Colectivo N 1
Descripcion Seccion	Oficina Central
Descripcion Banco	Banco del estado
Cuenta Corriente	123456789
Descripcion Forma de Pago	Cheque Finiquito
Descripcion Motivo de Retiro	CONTRATO VIGENTE
Descripcion Tipo de Contrato	Indefinido
fecha de Nacimiento	12-06-1990
Horas	176
Jornada	Lunes a viernes

Fuente: Elaboración propia

Apéndice B

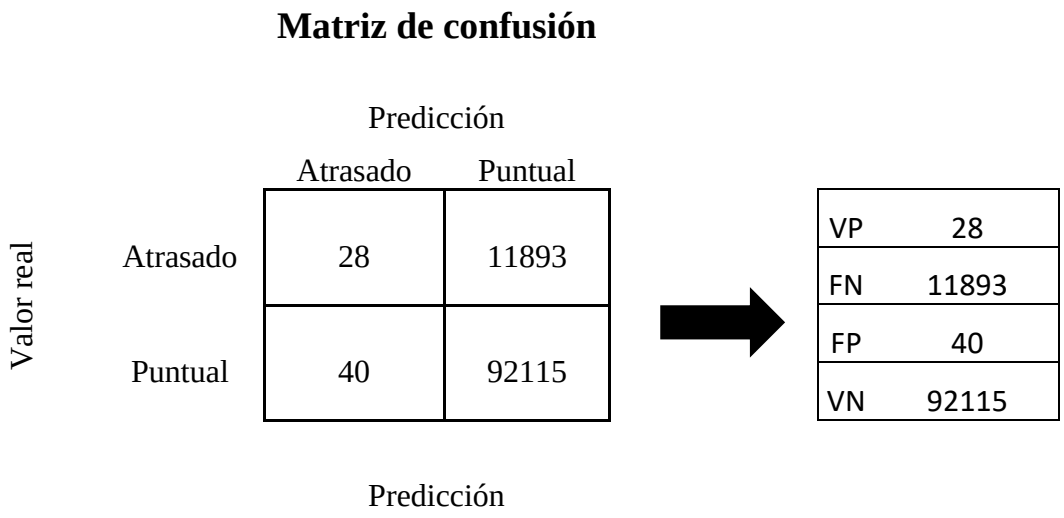
B.1 Tabla de correlación de variables:



Apéndice C

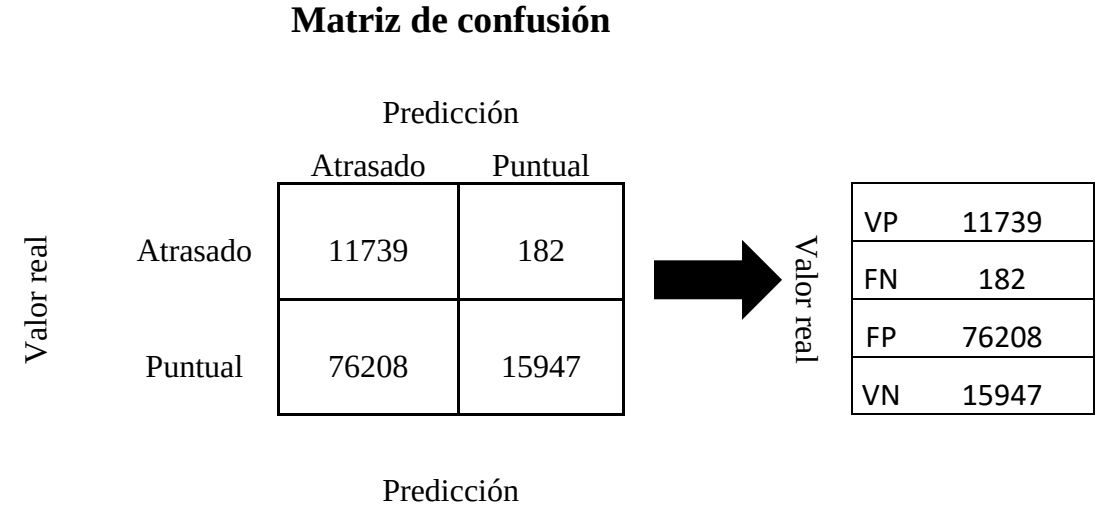
En esta sección se muestran las matrices de confusión y métricas de desempeño de los modelos usados para predecir **atrasos**:

C.1 Matriz de confusión Regresión Logística:



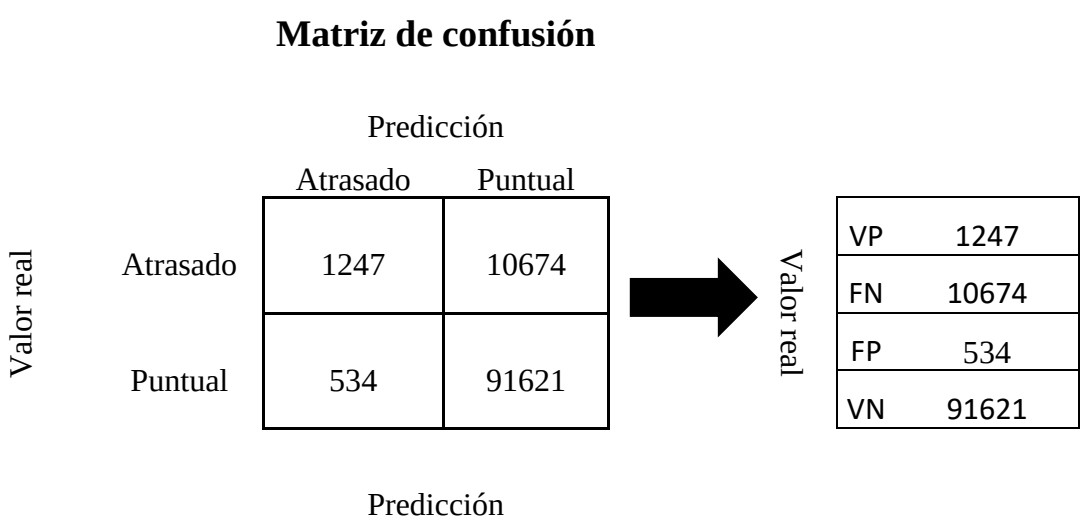
Matriz de confusión 1 Regresión Logística, atrasos.

C.2 Matriz de confusión *Naive Bayes*:



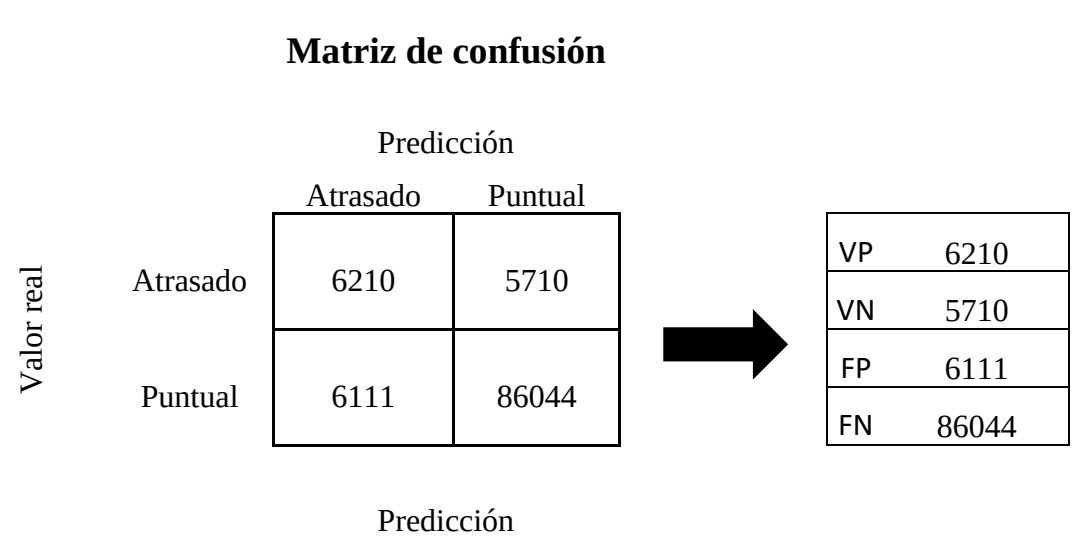
Matriz de confusión 2 Naive Bayes, atrasos.

C.3 Matriz de confusión *Gradient Boosting*:



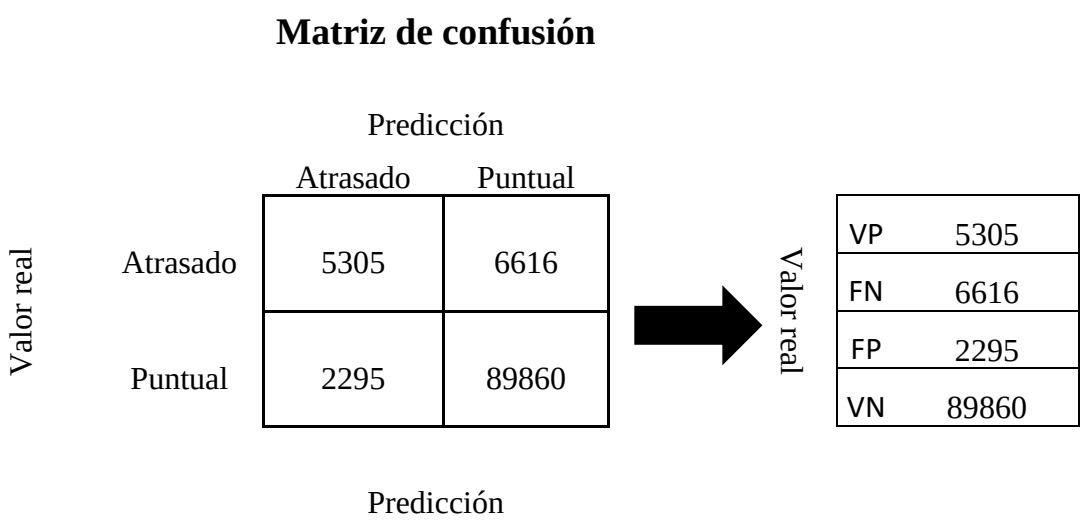
Matriz de confusión 3 Gradient Boosting, atrasos.

C.4 Matriz de confusión *Decision Tree*:



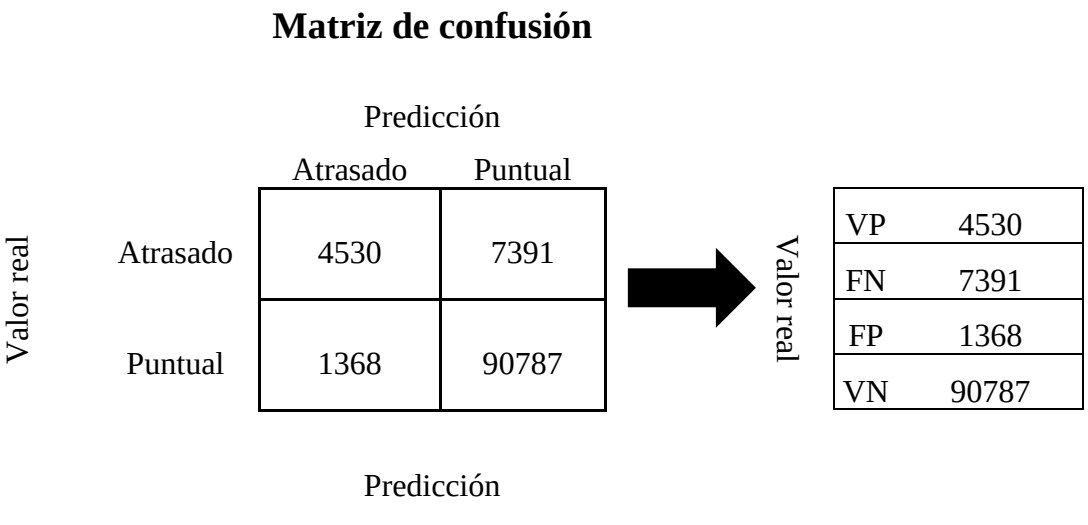
Matriz de confusión 4 Decision Tree, atrasos.

C.5 Matriz de confusión *Random Forest*:



Matriz de confusión 5 Random Forest, atrasos.

C.6 Matriz de confusión *XGBoost*:

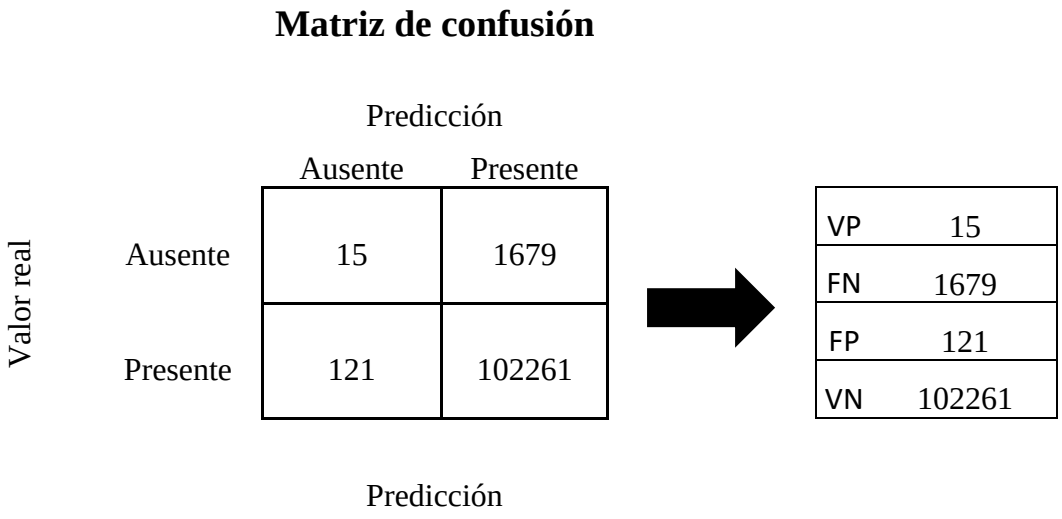


Matriz de confusión 6 XGBoost, atrasos.

Apéndice D

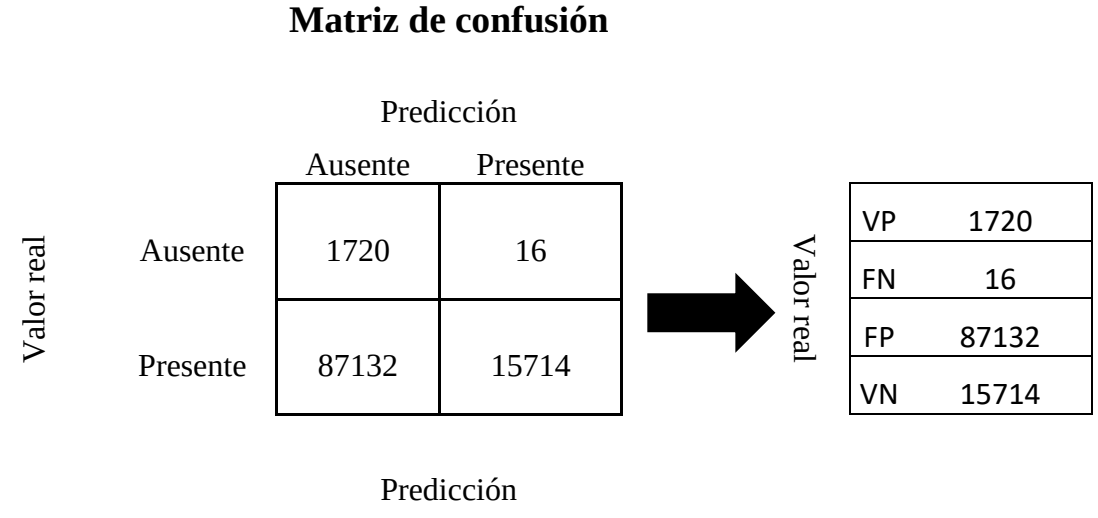
En esta sección se muestran las matrices de confusión y métricas de desempeño de los modelos usados para predecir ausentismo:

D.1 Matriz de confusión Regresión Logística:



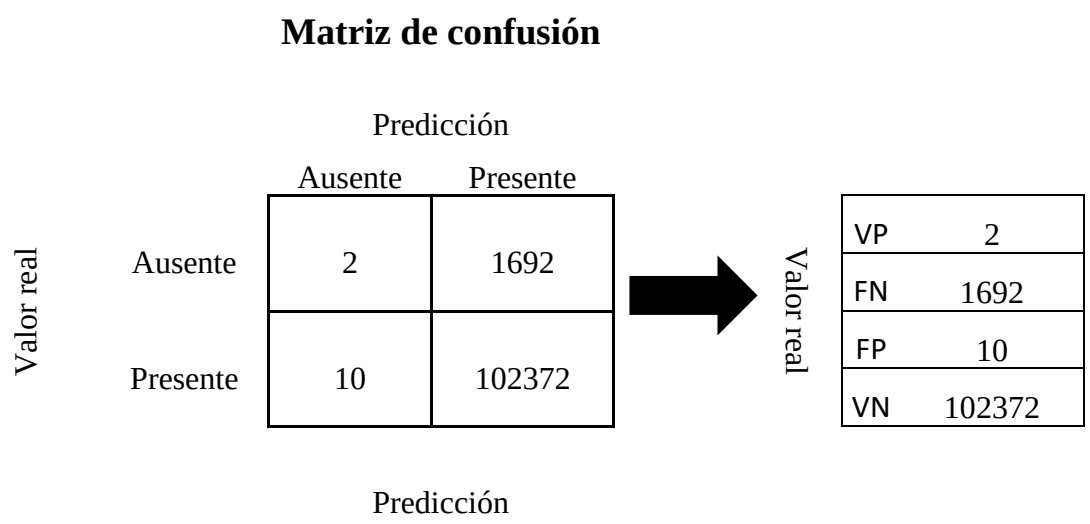
Matriz de confusión 7 Regresión Logística, ausentismos.

D.2 Matriz de confusión *Naive Bayes*:



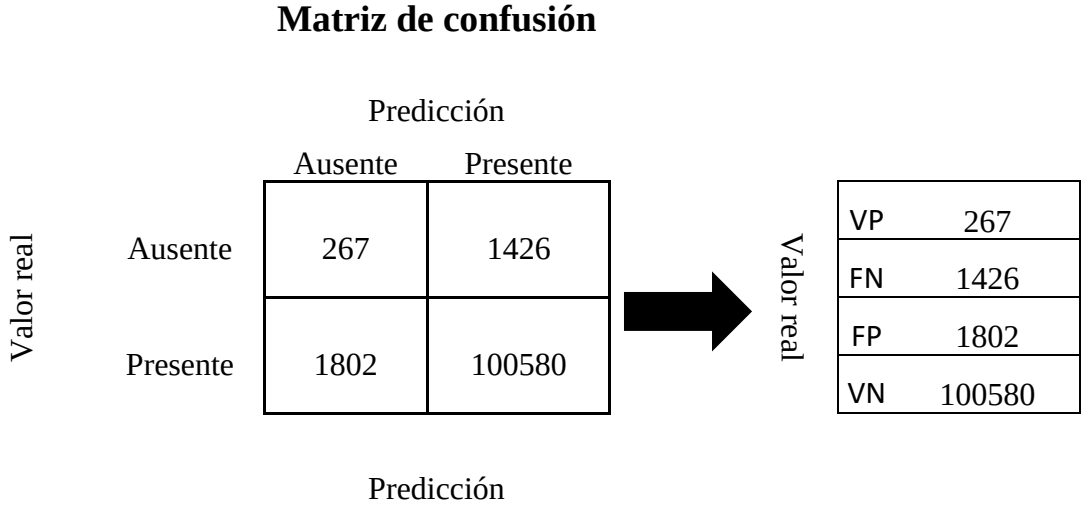
Matriz de confusión 8 Naive Bayes, ausentismos.

D.3 Matriz de confusión *Gradient Boosting*:



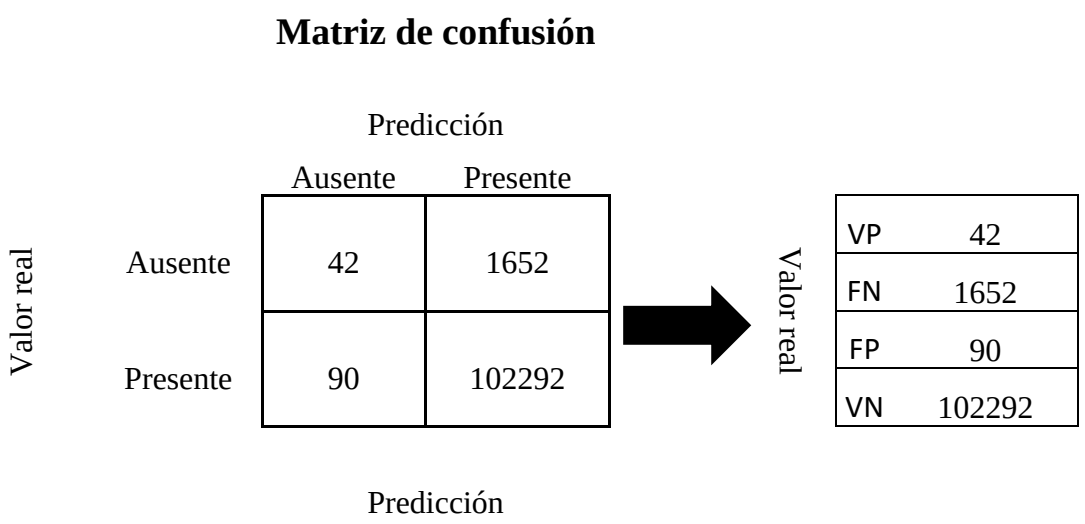
Matriz de confusión 9 Gradient Boosting, ausentismos.

D.4 Matriz de confusión *Decision Tree*:



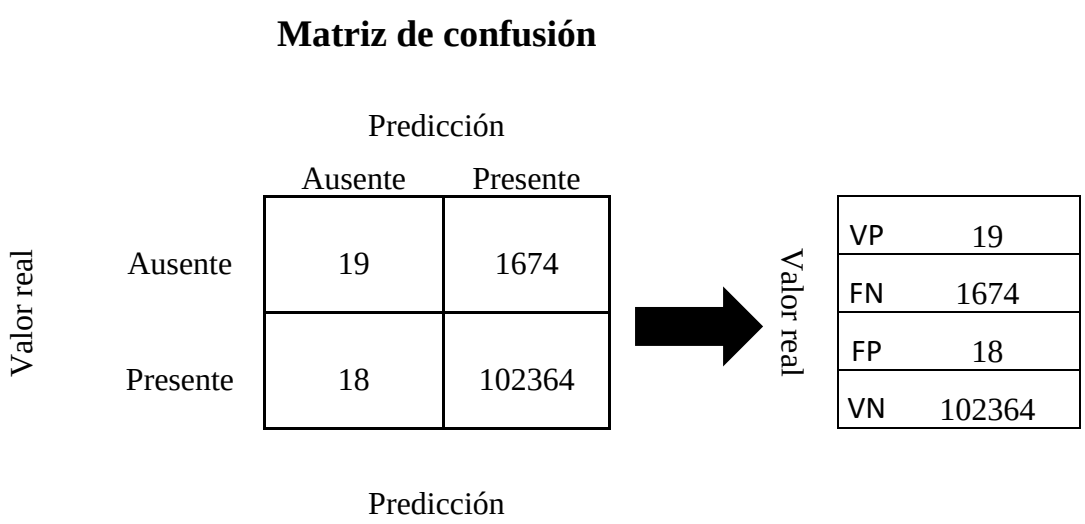
Matriz de confusión 10 Decision Tree, ausentismos.

D.5 Matriz de confusión *Random Forest*:



Matriz de confusión 11 Random Forest, ausentismos.

D.6 Matriz de confusión *XGBoost*:



Matriz de confusión 12 XGBoost, ausentismos.