Universidad de Valparaíso
Facultad de Ingeniería
Escuela de Ingeniería Civil Industrial

Propuesta De Un Modelo Predictivo Para Analizar El Ausentismo y Rotación Laboral Presente en SGS CHILE

Por

Luis Guillermo Campos Vergara - Luis Andrés Cofré Morales

Trabajo de Título para optar al Grado de
Licenciado en Ciencias de la Ingeniería y título de
Ingeniero Civil Industrial

Prof. Guía Mauricio Valle Barra

Diciembre, 2016

Agradecimientos

Agradezco a Dios por ayudarme a terminar este lindo camino y a perseverar en momentos complicados. También a mi familia que siempre me apoyo y me dio las fuerzas y energías para continuar. Y en definitiva son y fueron el pilar fundamental en mi vida.

Sin dejar de lado agradezco a mis amigos y profesores en la universidad, que me acompañaron y guiaron a lograr este objetivo.

Pero por sobre todas las cosas dedico y agradezco este logro a un Ángel que sigue cada paso que doy y que siempre espero este momento, mi amada y querida Abuela.

Luis Guillermo Campos Vergara

Agradezco a mi familia por el apoyo, ayuda y comprensión en este proceso. Por darme fuerza y enseñarme el mejor camino en la vida. También por ser las personas que siempre me acompañaron en los buenos y malos momentos, además de depositar sus esperanzas en mí.

También agradezco a mis amigos y profesores que fueron un pilar fundamental para mi desarrollo personal y académico dentro de la universidad. Que me enseñaron lo importante para seguir adelante y superar los obstáculos de la vida.

Luis Andrés Cofré Morales

Índice

Lista de ecuaciones **7**
Lista de graficos **8**
Lista de tablas **9**
Lista de Ilustraciones **10**
Glosario **11**
Lista de abreviaturas **12**
Resumen **14**
Introducción **15**
1. Presentación del tema **16**
1.1. Planteamiento del problema **16**
1.2. Objetivos **17**
1.2.1. Objetivo general **17**
1.2.2. Objetivos específicos **17**
1.3. Metodología utilizada **18**
2. Marco de antecedentes **19**
3. Marco Conceptual **25**
3.1. Metodología Crisp **25**
3.2. Técnicas de Minería de Datos **28**
3.2.1. Técnicas algebraicas y estadísticas **29**
3.2.2. Técnicas bayesianas **29**
3.2.3. Arboles de decisión **30**
3.2.4. Redes neuronales **31**
3.3. Justificación de la técnica de minería de datos **33**
3.4. Arboles de Decisión **33**
3.4.1. Arboles de Decisión para Clasificación **34**
4. Desarrollo del Modelo **41**
4.1. Construcción del Modelo **46**
4.1.1. Modelo A: Modelos de clasificación para predecir el estado del trabajador en la empresa **46**
4.1.2. Modelo B: Modelo de Clasificación para predecir los días de ausentismo que presentara un trabajador **46**
5. Resultados **48**

5.1. Resultados del modelo A 48
5.1.1. Matriz de Confusión 52
5.1.2. Validación Modelo A 52
5.2. Resultados del Modelo B 54
5.2.1. Matriz de Confusión Modelo B 57
5.2.2. Validación del modelo B 59
5. Sugerencias 60
7. Evaluación Económica 64
8. Conclusión y Recomendaciones 69
Anexos 72

Lista de ecuaciones

Ecuación 1: Función de entropía.....36
Ecuación 2: Ganancia de información.....37
Ecuación 3: Cálculo de exactitud 40
Ecuación 4: Cálculo de error.....41
Ecuación 5: Cálculo de verdaderos positivos.....41
Ecuación 6: Cálculo de falsos positivos 41
Ecuación 7: Cálculo de verdaderos negativos 41
Ecuación 7: Cálculo de falsos negativos 41
Ecuación 9: Cálculo de precisión 42

Lista de graficos

Grafico 1: Aporte Sectorial de servicios en chile para el año 2013.....19
Grafico 2: Promedio días de ausentismo por año.....20
Grafico 3: Causales de Ausentismo en empresa de servicios. 21

Lista de tablas

Tabla 1: Utilización Técnica de Minería de Datos 32
Tabla 2: Matriz de Confusión 40
Tabla 3: Ejemplo de matriz de confusión.....42
Tabla 4: Variables predictoras para el desarrollo del modelo 45
Tabla 5: Matriz de confusión del modelo A 52
Tabla 6: Calculo de promedio 53
Tabla 7: Tabla de variables para el modelo B 54
Tabla 8: Matriz de confusión del modelo B 57
Tabla 9: Calculo de promedio 59
Tabla 10: Situación actual vs Situación propuesta.....64
Tabla 11: Remuneraciones 64
Tabla 12: Costos Asociados al Trabajador (Situación Actual) 65
Tabla 13: Costo de un día de ausencia.....66
Tabla 14: Costo de la situación actual 66
Tabla 15: Costos Asociados al Trabajador (Situación Propuesta) 67
Tabla 16: Costo total de la situación propuesta.....67

Lista de Ilustraciones

Ilustración 1: Causal de Ausentismo por Mes 22
Ilustración 2: Porcentaje cantidad de trabajadores vs calidad de salud.....23
Ilustración 3: Cantidad de Fallas de Dotación Activa de la Empresa de Servicios en el año 2014 23
Ilustración 4: Cantidad de Personas que Presentan Faltas por Sector de Negocio.....24
Ilustración 5: Fases de Metodología CRISP.....28
Ilustración 6: Construcción del árbol de decisión 35
Ilustración 7: Árbol de decisión finalizado 38
Ilustración 8: Algoritmo de cobertura, en donde (a) es la cobertura de las instancias y (b) es el árbol de decisión para el mismo problema 39
Ilustración 9: Modelo A para clasificación entre contrato vigente (CV) y termino de contrato (RE) 48
Ilustración 10: Modelo B para predecir los días de ausentismo (DA) 55
Ilustración 11: Flujo de proceso de Incorporación 62

Glosario

Algoritmo: es un conjunto de operaciones que permiten mediante el cálculo hallar una solución a un problema.

Árboles de Decisión: son modelos de predicción utilizados para obtener conocimiento de una base de datos con gran cantidad de variables. Sirven para representar y categorizar las restricciones o condiciones presentes en dicha base de datos.

Payrrol (software): es un programa que tiene como objetivo agilizar y automatizar el proceso de pagar a los empleados de una empresa, además de aportar en la re-portabilidad y cálculos automáticos asociados a diferentes proceso de recursos humano.

Full Equivalence: corresponde a la cantidad de días ausentes contabilizados divididos en los días totales efectivos mensuales (30 días).

Metodología Crisp: metodología utilizada para la implementación de proyectos de minería de datos. Consta de seis fases que parte desde la preparación de la base de datos y termina con la implementación de un modelo a partir del conocimiento obtenido de dicha base de datos.

Data Mining (minería de datos): es un proceso que tiene como objetivo descubrir patrones en un conjunto de datos.

Exactitud (Acurracy): es la proporción del número total de predicciones que son correctas.

Error: representa un porcentaje de datos erróneos que arroja el modelo en comparación al total de los datos.

Tasa de falsos negativos: es la proporción de casos positivos que fueron clasificados incorrectamente como negativos.

Tasa de verdaderos Positivos (recall): es la proporción de casos positivos que fueron identificados correctamente como positivos.

Tasa de falsos Positivos**:** es la proporción de casos negativos que fueron clasificados incorrectamente como positivos.

Tasa de verdaderos negativos: es la proporción de casos negativos que fueron clasificados correctamente como negativos.

Precisión: es la proporción de los casos positivos predichos que eran correctos.

Lista de abreviaturas

Tipo de Empresa	TDE
Número de licencias presentadas	NLP
Edad de Trabajador	Edad
Causal de Ausentismo	CA
Días de Ausencia	DA
Mes Presentación Licencia	MPL
Estado Civil	EC
Nivel de Estudios	NE
Hijo Menor	HM
Numero de Parientes	NP
Años de Trabajo	AT
Tipo de Ausentismo	TA
Abogado	A
Administrativo	B
Analista	C
Asesor	D
Asistente	E
Auditor	F
Auxiliar	G
Ayudante	H
Cajero	I
Capataz	J
Chofer	K
Coordinador	M
Ejecutivo	N
Electromecánico	O
Encargado	P
Espectroscopista	Q
Estadístico	R
Fundidor	S
Geólogo	T
Gerente	U
Ingeniero	V
Inspector	W
Jefe	X
Maestro	Y
Mecánico	Z
Muestrero	AB
Operador	AC
Planificador	AD

Preparador de Muestra	AE
Prevencioncita	AF
Químico	AG
Recuperador	AH
Refinador	AI
Soldador	AJ
Sub Gerente	AK
Técnico	AN
Tecnólogo	AO
Zona Norte	ZN
Zona Sur	ZS
Zona Centro	ZC
Regio Metropolitana	RM
Administración	AQ
Agriculture	AR
Analítica	AS
As	AT
Automotriz	AU
Comercial	AV
CTS	CTS
División Club Deportivo	AX
Environmental	AY
Finanzas	AZ
Geo-metalurgia	BA
GIS	GIS
Inspectores	BC
IT	BD
Legal	BE
Management	BF
Metalurgia	BG
Mineralogía	BH
OI	OI
Oil, gas & chemical	OGC
Outsourcing	BK
Proyectos	BI
S&SC	SSC
Servicios Estratégicos	BO
Sin Clasificación	BP

Resumen

Uno de los aspectos más relevantes para una organización, es el capital humano. Ya que es el que permite alcanzar objetivos y guiar a la empresa a un crecimiento y desarrollo constante.

Hoy en día en el mercado nacional, estudios afirman que el ausentismo laboral está aumentando en el país. De esta manera la región metropolitana ha alcanzado un 5 % de ausentismo laboral por empresa. Por otro lado el 2014 se realizó un estudio en 26 empresas nacionales, que abarco a 63.000 empleados. Su conclusión fue que los chilenos se están amparando en las licencias y faltan al año 16,8 días. (Valenzuela O., 2015).

Por otro lado, según el estudio realizado por la consultora Randstad, la rotación laboral en chile alcanzo en el 2014 el 29%. Las personas que más cambiaron de empleo fueron aquellos trabajadores de 18 a 24 años (41%), seguidos por los de 25 a 34 años (28%).

SGS CHILE LTDA es una empresa encargada de ofrecer servicios de inspección, verificación, pruebas, ensayos y certificación a nivel global. Estos servicios permiten operar de forma sostenible, mejorando la calidad y la productividad.

Es necesario mirar las necesidades de los integrantes de la organización y así proponer estrategias para mantener a la gente motivada y hacerlos sentir importantes dentro de la empresa. Para comprender y guiar a estos incentivos, se debe conocer el motivo y comportamiento de los trabajadores, con base en esto se tienen personas productivas y se bajan los índices de ausentismo laboral. (Méndez, Leonett., 2005).

Actualmente el ausentismo y rotación laboral no es analizado por la empresa SGS CHILE, ya que no existe una forma de comprender el comportamiento o patrón a seguir, para tomar decisiones con respecto a las ausencias generadas.

El objetivo de esta tesis es el desarrollo de un modelo predictivo basado en un árbol de decisión que permita llevar a cabo un análisis del comportamiento de ausencias y rotación de los trabajadores de SGS CHILE. Con el propósito de establecer patrones que lleven a la empresa a manejar futuras decisiones.

Introducción

Para toda empresa el recurso humano es uno de los aspectos más relevantes para alcanzar objetivos y cumplir metas, además son las personas las que guían a la empresa a su crecimiento y desarrollo constante.

El ausentismo laboral es un tema relevante para Chile, debido a su importante aumento en los últimos años. INMUNE, como organización que promueve el buen uso de las licencias y enfocada a reducir el fraude en el sistema de salud, ha calculado que el costo de un trabajador con licencia equivale al menos al 50% de su sueldo, incluyendo los costos directos e indirectos, seguros, beneficios, costos de capacitación del reemplazante y baja productividad de éstos, lo que provoca un empeoramiento del ambiente laboral, etc.

La rotación laboral en Chile ha aumentado considerablemente, según la empresa KPI Estudios, estimo que la tasa de desvinculación de personal en el mercado chileno habría crecido un 25%, siendo esta tasa la más alta de la región latinoamericana. Esto implica pérdida de productividad y de información estratégica.

Es importante para la compañía conocer y analizar los factores que provocan el ausentismo y la rotación de sus trabajadores, ya que, son éstos los que realizan las actividades dentro de la empresa. Cuando se conocen dichos factores, la empresa puede brindar a sus empleados las soluciones necesarias para resolver sus problemas y evitar futuras ausencias o desvinculaciones.

De lo anterior se desprende que la investigación tiene como propósito el desarrollo de un modelo predictivo basado en un árbol de decisión que permita llevar a cabo un análisis del comportamiento de ausencias y rotación de los trabajadores de SGS CHILE.

Para desarrollar dicho modelo es necesario utilizar la metodología CRISP-DM, término que hace referencia a la minería de datos y el análisis de bases de datos.

1. Presentación del tema

1.1. Planteamiento del problema

El ausentismo laboral y la rotación del personal son factores que afectan la productividad y generan costos dentro del proceso productivo de una organización. Estos factores provocan un menor flujo de las actividades y procesos al interior de la empresa. El ausentismo laboral se produce cuando el empleado no asiste a realizar sus labores por la cual fue contratado y se dice que el personal rota cuando los trabajadores se van de la empresa, ya sea, por despidos o renuncian, y son reemplazados por otro en su lugar de trabajo. Según el ranking de la Organización para la Cooperación y el Desarrollo Económico (OCDE), los países tienen un promedio de ausentismo laboral de 11 días al año. Chile aparece con 15,6 días, esto deja al país con un 41,8% más de ausentismo que los demás países de la organización.

El problema radica en que los últimos meses se están presentando altos índices de rotación de empleados al interior de la empresa, lo que supone altos costos de reclutamiento y entrenamiento para la compañía.

Además cabe señalar que SGS Chile no genera un análisis o estudios sobre la cantidad de ausencias de los trabajadores. Dentro de la empresa existen aproximadamente 3.700 datos que describen el ausentismo generado por un empleado dentro del periodo 2013 y 2014. La falta de patrones no le permite tomar decisiones con respecto a este factor, lo que provoca una falta de control para ciertos grupos de personas que poseen elevados y reiterados días de ausentismo, además de desconocer el índice de ausentismo presente en la empresa.

De lo expuesto anteriormente, la empresa solo registra los datos del ausentismo y la rotación en el sistema, desaprovechando la oportunidad de seguir un proceso estandarizado que permita realizar un seguimiento y análisis de sus trabajadores. Cada vez que una persona se ausenta o es desvinculada del lugar de trabajo, se aumenta la carga de trabajo a otro empleado, con esto el empleado baja su rendimiento con respecto a la realización de tareas en horario laboral.

Esta memoria requiere plantear un modelo predictivo de la rotación y ausentismo de empleados con un enfoque de minería de datos a partir de los extensivos registros de ausentismo y desvinculaciones de trabajadores de la empresa. Con el objetivo de comprender el comportamiento y reducir así, el nivel de ausentismo y rotación del personal presente en la organización.

1.2. Objetivos

1.2.1. Objetivo general

Desarrollar un modelo predictivo basado en un árbol de decisiones, para analizar el ausentismo y rotación laboral de SGS CHILE.

1.2.2. Objetivos específicos

Para desarrollar los objetivos específicos, se considera la metodología de minería de datos (Data Mining), por lo tanto los objetivo a continuación mencionados, deben estar alineados con dicha metodología.

Los Objetivos Específicos son:

a) Establecer una base de datos depurada y filtrada del ausentismo laboral y la rotación del personal.
b) Entrenar un modelo predictivo basado en arboles de decisiones.
c) Validar y verificar el modelo predictivo.
d) Establecer reglas de decisión que expliquen el ausentismo laboral y la rotación del personal.

1.3. Metodología utilizada

Para realizar el estudio en la empresa SGS Chile es necesario cumplir con 6 fases. Estas etapas comprenden a la realización de la metodología CRISP-DM (Chapman et al., 2000).

- Fase 1: Conocimiento de la empresa.

En esta etapa se visita la empresa y en particular el área de recursos humanos, con el fin de conocer y comprender la situación actual de la empresa.

- Fase 2: Conocimiento de los datos.

Obtener y conocer la base de datos sobre ausentismo y rotación laboral presente en SGS Chile con el fin de identificar la calidad de la información y dimensionar la cantidad de trabajadores que se ausentan de su lugar de trabajo

- Fase 3: Preparación de los datos.

Identificar los datos necesarios para comprender los factores que causan el ausentismo en los trabajadores, se necesita cruzar la información con los antecedentes personales del trabajador.

Filtrar y limpiar las datos inconsistentes que serán utilizados para generar el modelo predictivo.

- Fase 4: Diseño del modelo de datos.

Después de analizar varias técnicas de modelado, se determina la ideal para trabajar y construir el modelo predictivo.

Crear el modelo predictivo sobre ausentismo y rotación laboral en la empresa.

- Fase 5: Evaluación del modelo.

Evaluar el modelo y revisar su construcción para comprobar si cumple con los objetivos del estudio.

- Fase 6: Generar un informe final.

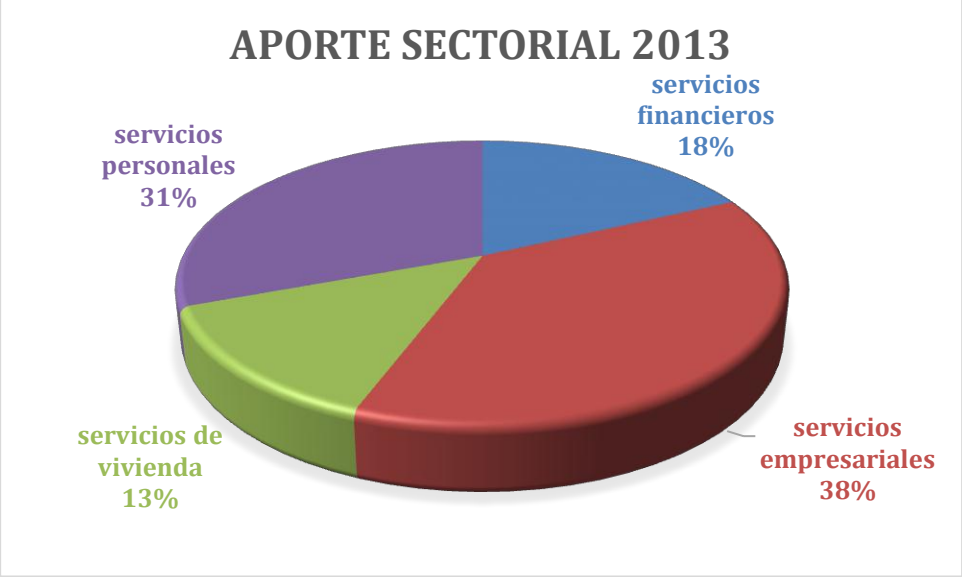
Generar un informe de manera tal que el usuario pueda entender y utilizar la información presentada.

2. Marco de antecedentes

En Chile el sector de servicios viene creciendo de manera dinámica y exponencial. Durante el año 2013 se exporto un total de US $ 12.800 millones en servicios, los que muestran solo un 17% del total de las exportaciones en Chile. Sin embargo en las exportaciones que poseen valor agregado, el sector de servicios crece en un 30%, la participación de los servicios en el PIB de Chile llega a un 67% y alcanza más de un 70% del empleo total del país. (Sandoval F., 2014).

A continuación se muestran los diferentes servicios dentro del aporte sectorial que representan.

Grafico 1: Aporte Sectorial de servicios en chile para el año 2013.

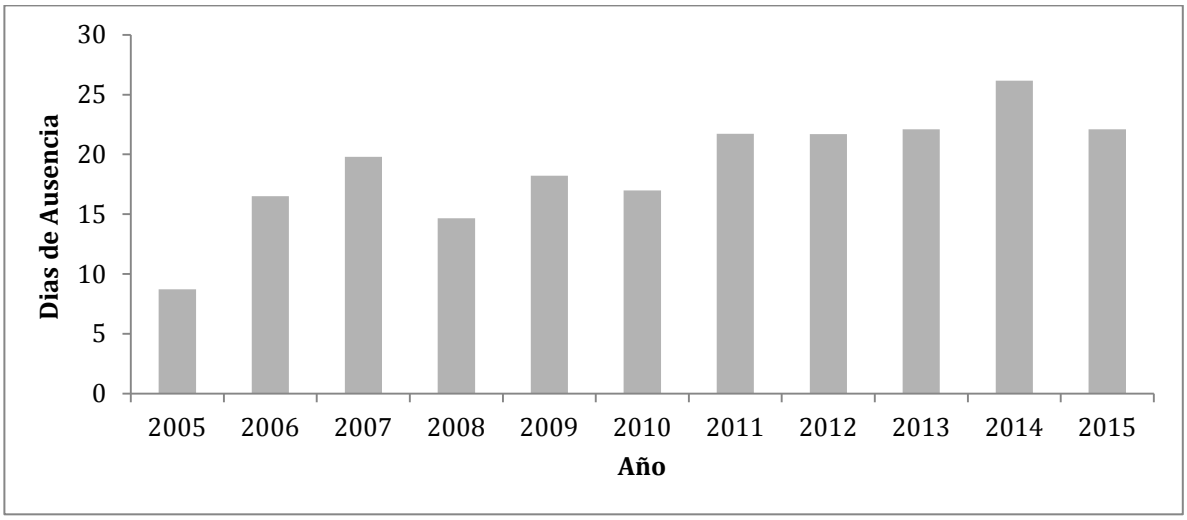


Fuente: Elaboración Propia a partir de datos del banco central de chile.

En el Grafico 1, se observa el aporte sectorial en términos de ocupación para el año 2013. El sector de servicios representa un 33,9 % de participación entre todos los sectores. El grafico muestra los porcentajes por categoría de servicios específicos y su porcentaje de relevancia. Siendo el mayor el sector de servicios empresariales con un 38%.

Según el ranking de la OCDE, los países tienen un promedio de ausentismo laboral de 11 días al año. En este aspecto, Chile presenta 15,6 días, cifra que resulta preocupante (Guihard T., 2012). Esta realidad también la presenta SGS, que ha generado un crecimiento significativo de días de ausentismo los últimos años. Llegando a un promedio de 26,15 días de ausentismo por trabajador para el año 2014.

Grafico 2: Promedio días de ausentismo por año



Fuente: elaboración propia a partir de información entregada por el área de recursos humanos en SGS.

A partir del Grafico 2, se observa el crecimiento en los días de ausentismo desde el año 2005 al año 2015. Llegando a un máximo de 26,15 días para el año 2014.

Debido a lo anterior, es relevante que las empresas generen un mayor control y análisis sobre el comportamiento de sus trabajadores. De manera de predecir un patrón o comportamiento que se relacione con las causales y conductas del trabajador al momento de presentar un ausentismo. Muchas causantes del ausentismo y rotación laboral son por la falta de pertenencia y desmotivación hacia la empresa, ya que la organización no hace parte al trabajador de sus acciones.

Para los trabajadores es fundamental que la compañía los apoye en el ámbito de desarrollo personal y motivación, ya que, mientras menos problemas tenga una persona, su calidad de vida, su satisfacción y rendimiento en el lugar de trabajo mejorarán notablemente (Guihard T., 2012).

Para la elaboración de un indicador de comportamiento o estrategia a implementar es necesario conocer los motivos o causales respecto al porque los trabajadores se van de la empresa o faltan a su lugar de trabajo, con esta información se puede llegar a predecir y tomar decisiones anticipadas para evitar el ausentismo y rotación laboral. Es por este propósito con el que se tomara la decisión de proponer un modelo predictivo basado en un árbol de decisiones para analizar el ausentismo laboral de la empresa SGS chile.

El plan de acción de la empresa cuando un trabajador falta a su lugar de trabajo, es el de verificar mediante el software de control de asistencia asociado a Payroll (software de recursos humanos), quien utiliza una tecnología de lectura de huella dactilar. Arrojando hora de ingreso y hora de salida, así como también días de ausencias. Estos datos son

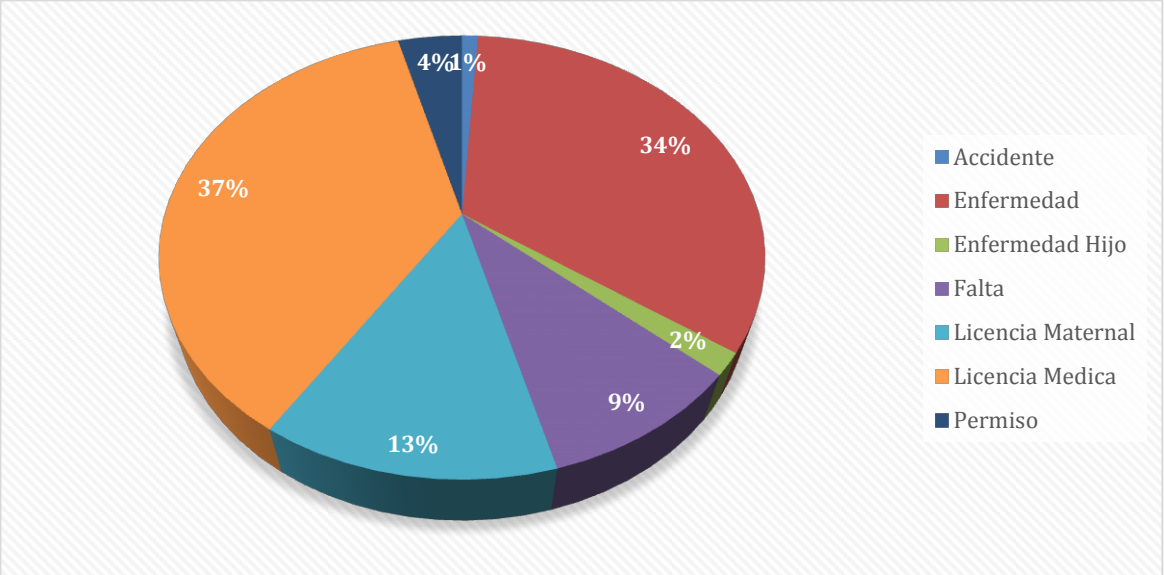
recolectados por el área de control de asistencia quien lo entrega al área de remuneraciones, luego es consolidado en payroll, vinculando cada inasistencia con las características del trabajador, para finalmente asociarlo a una causal de ausentismo. (Analista de Recursos Humanos en SGS, 2015).

Por otro lado si se produce una renuncia o retiro del trabajador, se actualiza la planilla del trabajador dentro de Payroll, quedando con una de las condiciones de desvinculación.

Sin embargo con la información almacenada y recolectada del sistema (Payroll), no se establecen parámetros o metodologías a seguir para ver que se hace con las ausencias, ni mucho menos se establecen políticas de amonestación o decisiones a tomar en caso de reiteradas faltas. Para el caso de rotación solo se generan reportes de altas y bajas (aumento o disminución de trabajadores dentro de una división de la empresa), con el fin de saber sí se debe contratar otro trabajador para el puesto de trabajo con una baja (Gerente de Recursos Humanos en SGS, 2015).

Una de las variables que permiten construir el modelo, serán las causales de ausentismo que se presentan a continuación.

Grafico 3: Causales de Ausentismo en SGS CHILE.



Fuente: elaboración propia a partir de empresa SGS CHILE.

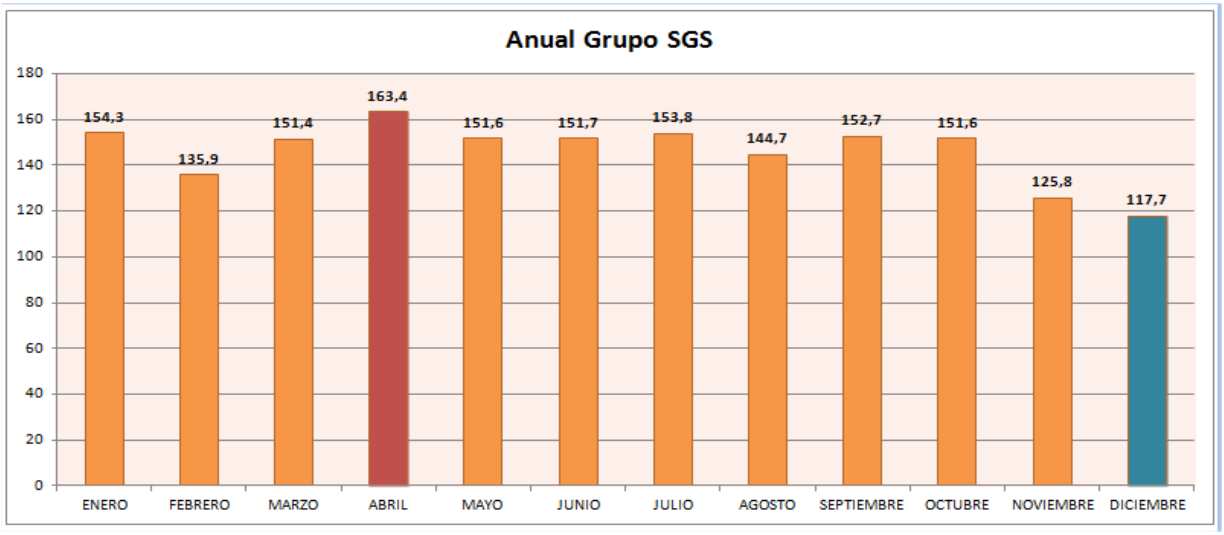
El grafico 3 se construyó a partir de una base de datos que comprende a todos los trabajadores que presentan una ausencia a su lugar de trabajo durante el año 2013 y 2014.

En el grafico 3 se observan las 4 primeras causales de ausentismo que se repiten con mayor frecuencia, siendo estas la licencias médica (37%), enfermedad (34%), licencia maternal (13%) y falta (9%).

En la ilustración 1 se muestra la cantidad de trabajadores en Full Equivalence (Correspondiente a la cantidad de días ausentes contabilizados divididos en los totales de días efectivos mensuales (30 días)) del año 2014 que presentaron ausencias en las siguientes causales (analista de mejora continua, 2015).

- Licencia Médica
- Licencia Maternal
- Licencia por enfermedad del hijo
- Falta
- Accidente

Ilustración 1: Causal de Ausentismo por Mes



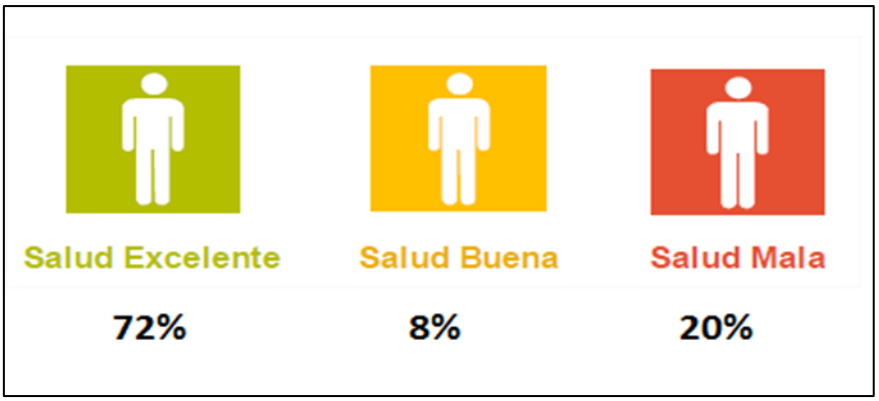
Fuente: elaborado por el analista de recursos humanos de la empresa de servicios, 2015.

Como se aprecia en la ilustración 1 el máximo valor se registra en el mes de Abril y en el mes de Diciembre la mínima, es decir, para el mes de abril existieron 164 trabajadores que presentaron una falta a su lugar de trabajo. Mientras que para el mes de diciembre, solo 117 trabajadores presentaron una ausencia laboral por algunas de las causales de ausentismo.

Para comprender la salud laboral presente en la empresa de servicios, es necesario identificar la dotación activa presente al mes de diciembre del año 2014, considerando solo las personas que presentan licencias médicas. Además se establecen ciertos rangos de aceptación para clasificar a los trabajadores según sus ausencias anuales:

- Una salud excelente, la que comprende valores menores a 3 días de ausencias.
- Una salud buena, la que comprende valores mayores a 3, pero menores a 7 días de ausencias.
- Una salud mala, la que comprende valores iguales o superiores a 7 días de ausencias.

Ilustración 2: Porcentaje de cantidad de trabajadores con su respectiva calidad de salud

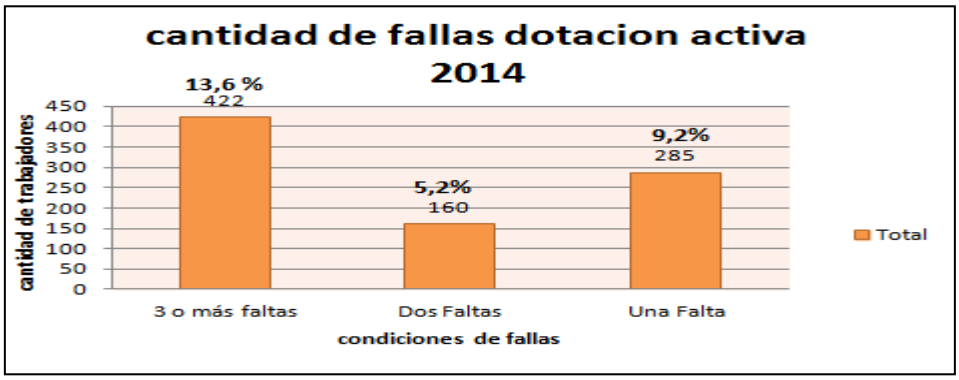


Fuente: elaboración a partir de Towers Watson, 2014

En la ilustración 2 se observa que del total de personas que presentaron una licencia médica, el 72% de ellos faltaron entre 1 y 3 días, mientras que el 20% presento una ausencia mayor a 7 días.

Para poder interpretar de mejor manera las causales de ausencias y el comportamiento de los trabajadores al interior de la empresa, se observan en la figura 3 a los trabajadores que han presentado desde 0 faltas, hasta los que han registrado más de 3. Según ciertos estándares de la empresa se sostiene que más de tres faltas, es un punto crítico para comenzar a tomar la decisión de desvincular a la persona de su puesto de trabajo, sin embargo antes de eso este debe ser advertido mediante una amonestación. (Analista de recursos humanos, 2015).

Ilustración 3: Cantidad de Faltas de Dotación Activa de la Empresa de Servicios en el año 2014

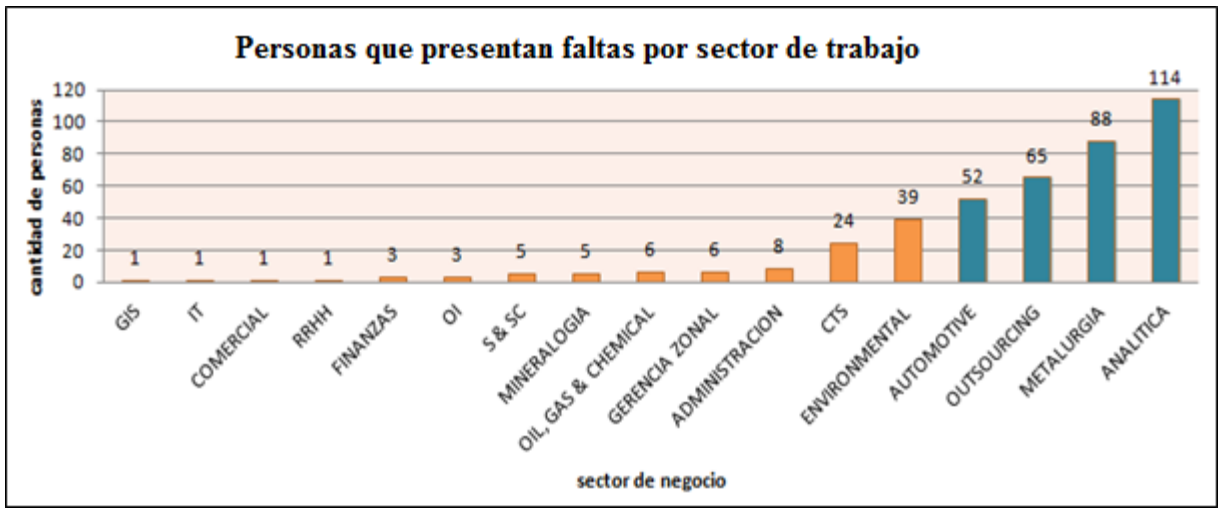


Fuente: Grafica elaborada por analista de recursos humanos de empresa de servicios, 2015.

La Ilustración 3 representa la cantidad y el porcentaje de la dotación activa de los trabajadores en SGS, siendo el mayor los trabajadores que presentan 3 o más faltas durante el año (de enero a diciembre del 2014). Sin embargo no hay que dejar de analizar y controlar a tiempo todos aquellos trabajadores que poseen 1 y 2 faltas respectivamente. Esta información se calculó en base a la cantidad total de trabajadores activos (3093 trabajadores). (Analista de mejora continua, 2015)

Para analizar con más detalle a los trabajadores con más de tres faltas presentes en SGS, es que se presenta a continuación la ilustración 4 que representa las cantidades de faltas por sector de negocio. Además se analizan las faltas por dotación respecto a cada sector, para obtener conclusiones más generales respecto a las faltas presentes en SGS. (Analista de recursos humanos, 2015).

Ilustración 4: Cantidad de Personas que Presentan Faltas por Sector de Negocio



Fuente: Grafica elaborada por analista de recursos humanos de la empresa de servicios, 2015

La ilustración 4 muestra al total de personas que presentan faltas por sector de negocio dentro de la empresa. Donde se puede observar que el mayor número de personas lo presenta el sector de Analítica y tanto Gis, IT, Comercial y RR.HH. presentan el menor número de trabajadores con faltas a su lugar de trabajo durante el año de medición 2014.

Es así como se obtiene que el sector de analítica es quien posee mayor cantidad de faltas, sin embargo estos también suele explicarse por la gran dotación existente en este sector en comparación con otros.

3. Marco Conceptual

3.1. Metodología Crisp

Esta metodología posee un conjunto de actividades seleccionadas en base a la experiencia de prueba y error recolectada a través de variados proyectos. Estas actividades están ordenadas en seis fases sucesivas que se generan durante el proceso de data mining, desde la definición de los objetivos del negocio que se pretende obtener hasta la vigilancia y el mantenimiento del modelo que se proponga. Cada una de estas fases se subdivide en tareas jerarquizadas y ordenadas, desde un mayor a un menor nivel de detalle. (Chapman et al., 2000).

Las tareas se componen a su vez de actividades específicas, y de un conjunto de resultados concretos. La metodología CRISP-DM constituye un mapa de ruta que permite determinar qué actividades a desarrollar y en qué etapa, de manera de alcanzar los objetivos finales del proyecto. (Chapman et al., 2000).

Los objetivos de la metodología Crisp son: (Goicochea A., 2009).

- Aprender nuevas técnicas para comprender y aplicar la minería de datos
- Desarrollar proyectos de minería de datos, mediante un proceso estandarizado.

A continuación se presentan las 6 fases dentro de la metodología Crisp: (Chapman et al., 2000).

1. Fase de compresión del negocio: se centra en la compresión de los objetivos del proyecto de data mining desde el factor del negocio. También se debe definir el problema y el plan de acción para realizar los objetivos planteados. Las tareas a realizar en esta fase consisten en:
    - Determinar los objetivos de negocios: entender lo que el negocio quiere y descubrir los factores importantes que pueden influir en los resultados del proyecto.
    - Evaluar la situación del proyecto (recursos, restricciones y requerimientos): la idea es profundizar en los detalles a considerar para la creación del análisis de datos y el plan del proyecto.
    - Determinar objetivos de minería de datos: definir los objetivos en términos técnicos para lograr los objetivos del negocio.
    - Crear un plan de acción: describir los pasos esperados del proyecto para cumplir con los objetivos de minería de datos y del negocio.

2. Fase de comprensión de los datos: consiste en la recolección inicial de datos, identificando la calidad y relaciones de ellos. En esta fase de deben describir los datos en términos de número de registros, número de campos por registro y significado de cada campo. Además comprende todas aquellas actividades que permitan entender y familiarizarse con la base de datos, para identificar problemas, detectar subconjuntos de datos y formular las hipótesis necesarias. Las tareas de esta fase son:

- Recolección de la data inicial: obtener acceso a la base de datos necesaria para resolver el problema y realizar los objetivos del proyecto.
- Describir los datos: describir y ordenar de mejor manera la base de datos para su fácil entendimiento.
- Explorar los datos: en esta tarea se ven los temas de análisis estadísticos, las relaciones entre los datos, agrupación y transformación de los datos. Comprende las preparaciones necesarias de la base de datos para su posterior análisis.
- Determinar la calidad de los datos: se determinan y buscan los datos incompletos, los errores en la base de datos, valores perdidos o fuera de rango.

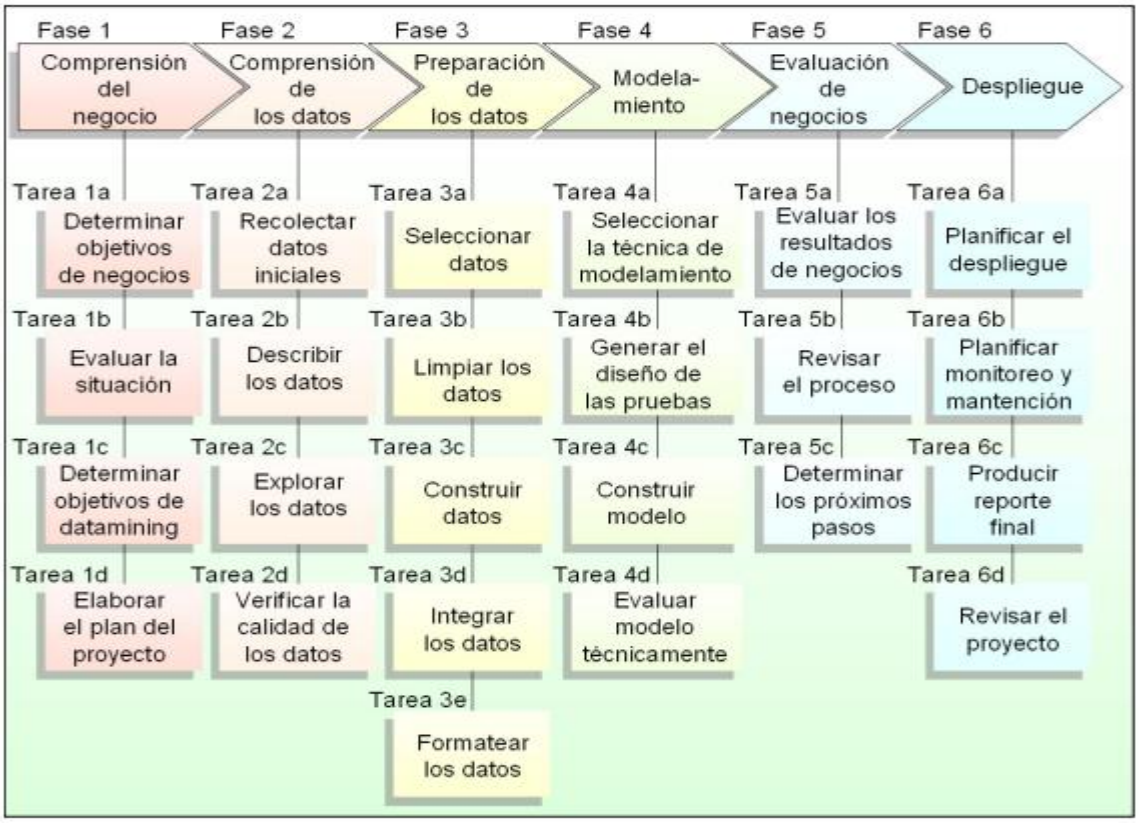
3. Fase de preparación de los datos: en esta fase se debe constituir la última base de datos (datos que se introducen en la herramienta para ser modelados), la cual debe contener todas las características para determinar el valor de la variable que se pretende predecir. En esta fase se desarrolla la selección de los datos, la tarea de limpieza y la tarea de construir e integrar datos adicionales. Esta fase debe entregar datos que estén en un formato adecuado para la técnica de modelamiento que se empleara en la fase siguiente. Las tareas son:
    - Selección de datos: decidir los datos que se utilizaran para el análisis, determinar las columnas y filas necesarias.
    - Limpieza de los datos: elevar la calidad de los datos requeridos para el análisis, seleccionando subconjuntos, insertando valores predeterminados o modelando datos faltantes.
    - Construcción de datos: la idea es producir atributos nuevos y transformar los valores de los atributos existentes.

4. Fase de modelamiento: consiste en descubrir una relación entre un conjunto de variables y una variable que se espera predecir. Contempla la selección de una técnica de modelamiento, entre las cuales pueden mencionarse arboles de decisión para segmentación, redes neuronales o regresión logística para predicción, inducción de reglas generalizada para descubrimiento de patrones y análisis de factores para disminuir la dificultad de los datos. Debido a que existen varias técnicas de modelamiento para un mismo problema, es necesario encontrar la que mejor se adapte a la base de datos y los objetivos del proyecto. Las tareas necesarias de esta fase son:
    - Seleccionar la técnica de modelado: seleccionar la herramienta de modelado específica que se utilizara en la extracción de conocimiento de la base de datos.
    - Generar un diseño de pruebas: preparar la base de datos para crear pruebas que determinen la calidad y validez del modelo.
    - Construir el modelo: ejecutar la herramienta de modelado sobre la base de datos.
    - Evaluar el modelo: interpretar el modelo generado de acuerdo al conocimiento extraído, los criterios de éxito y el diseño de pruebas esperado.

5. Fase de evaluación: Durante esta fase se debe evaluar el nivel de satisfacción durante el proyecto. Incluye la tarea de evaluar los resultados, la tarea de revisar el proceso de data mining y determinar los pasos a seguir. El objetivo de esta fase es determinar que se pueda asegurar el cumplimiento de los objetivos del negocio y encontrar algún problema importante. Las tareas de esta fase son:
    - Evaluar los resultados: evaluar el grado en que el modelo cumple con los objetivos del negocio. También se debe poner a prueba el modelo con datos reales.
    - Revisión del proceso: se realiza una revisión completa del proceso de minería de datos con el fin de determinar algún factor o problema relevante dentro del proceso que no fue considerado.

6. Fase de despliegue del modelo: en esta fase se define la estrategia para implementar los resultados obtenidos en la minería de datos. Contiene tareas como la planificación del modelo, planificación del monitoreo y la mantención de los modelos, que permiten generar un reporte al final del proyecto. El objetivo es presentar el conocimiento obtenido de tal manera que el cliente pueda ocuparlo. En esta fase se puede realizar un simple informe para implementar un proceso de minería de datos repetible en la empresa. Las tareas son:
    - Planeación de la implementación: el fin es tomar los resultados de la minería de datos y generar una estrategia para su implementación.
    - Monitoreo y mantenimiento: el fin es evitar el uso incorrecto de los resultados de la minería de datos.
    - Informe final: escribir un reporte final, puede ser un resumen del proyecto o una presentación completa del proceso de minería de datos.

Ilustración 5: Fases de Metodología CRISP



Fuente: Obtenida de Luca M. (2006).

La ilustración 5 muestras las 6 fases de la metodología CRISP necesarias para desarrollar e implementar un proyecto de minería de datos

3.2. Técnicas de Minería de Datos

Existen diferentes técnicas para abarcar un problema mediante la minería de datos, dichas técnicas tienen diferentes paradigmas y estos a su vez incluyen diferentes algoritmos y variaciones de los mismos. Por lo tanto la efectividad de un algoritmo depende del dominio de aplicación, es decir, no existe un método universal aplicable a todo tipo de problema. Las técnicas intentan obtener patrones o modelos a partir de los datos recopilados. (Hernández et al., 2004).

Los algoritmos de minería de datos tienen dos categorías, los supervisados o predictivos y los no supervisados o de descubrimiento del conocimiento según Weiss, Indurkhya, citados por (Moreno M. et al., 2006).

- Los algoritmos supervisados o predictivos predicen el valor de un atributo a partir de un conjunto de datos y sus relaciones. Estos tipos de algoritmos se construyen en dos

fases; la fase de entrenamiento, en donde se toman un porcentaje de la base de datos para construir el modelo, y en la fase de prueba se utiliza el resto de los datos.

- Los métodos no supervisados o de descubrimiento del conocimiento se utilizan para descubrir patrones y tendencias en los datos actuales y no se utilizan datos históricos.

Para poder realizar esta investigación es necesario ocupar los algoritmos predictivos, ya que, se utilizara una base de datos históricos de la empresa de servicios, de la cual se buscara extraer conocimiento y poder predecir algún comportamiento o dato en particular. Dentro de los algoritmos supervisados o de predicción encontramos las diferentes técnicas para realizar el modelo, que se nombran a continuación.

3.2.1. Técnicas algebraicas y estadísticas

Son útiles para las tareas regresión y discriminación (clasificación o agrupamiento). Esta técnica intenta determinar los valores de una o varias variables a partir de un conjunto de datos. Dentro de los conceptos estadísticos encontramos la regresión lineal, múltiple y no lineal, que tratan de expresar modelos y patrones mediante fórmulas algebraicas. Los resultados de esta técnica son utilizados para la predicción de valores continuos (Hernández et al., 2004).

Las aplicaciones son muy diversas a la hora de hablar de este tipo de técnica, ya que cada modelo que se realice para predecir un determinado fenómeno, deberá pasar por un proceso algebraico y estadístico para conseguir la predicción (Hernández et al., 2004).

3.2.2. Técnicas bayesianas

Se utilizan como clasificadores estadísticos que pueden predecir las probabilidades de un número de miembros de una clase y la probabilidad de que una muestra determinada pertenezca a una clase particular. Uno de los métodos más utilizados es el *Naive Bayes*. Este método combinado con otros procedimientos de selección de atributos sirve para eliminar la redundancia. Las redes bayesianas generalizan las topologías de las interacciones probabilísticas entre variables y permiten representar gráficamente dichas interacciones (Hernández et al., 2004). A continuación se presenta una tesis, donde se muestra una aplicación de dicha práctica.

"Técnicas bayesianas de apoyo a la toma de decisiones y sus aplicaciones". Para el desarrollo de esta tesis se utilizaron métodos estadísticos para resolver la toma de decisiones. Donde se utilizó la red bayesiana para proponer un nuevo método de agregación de preferencias que le permite a un grupo poder determinar una decisión mediante la ordenación de diferentes alternativas a un problema. La eficacia del método propuesto se ha comprobado como sistema de soporte y apoyo al diagnóstico médico, proporcionando buenos resultados (Calle F., 2014).

Las redes bayesianas se utilizaron para proporcionar información importante respecto de problemas de decisiones en el ámbito veterinario y en el análisis de riesgos en aviación. El

objetivo de esta tesis es la aplicación de nuevas técnicas estadísticas que ayuden al proceso de la toma de decisiones. Este trabajo se basa como objetivo en el estudio y en la propuesta de un método de clasificación para dicha tarea. Además se plantea una propuesta de un método hibrido que incluye tres fases principales, entre ellas están: la comparación por pares, regresión bayesiana y algoritmo del vecino más cercano (Calle F., 2014).

La metodología bayesiana genera un modelo de aprendizaje, con ello mejora los resultados de clasificación por sobre otros métodos, sobre todo en el ámbito de la medicina (Calle F., 2014).

En definitiva el desarrollo de esta tesis busca mostrar las diferentes aplicaciones donde es necesario utilizar las técnicas bayesianas. Las que permiten tomar decisiones correctas, algunos de sus utilizaciones son experimentos relacionados con el cáncer de mama y medicina veterinaria, entre otros (Calle F., 2014).

3.2.3. Arboles de decisión

Son una serie de decisiones o condiciones organizadas en forma jerárquica, a modo de árbol. Son muy útiles en problemas que mezclan datos categóricos y numéricos. Los arboles de decisión se clasifican en dos tipos (Hernández et al., 2004).

- Los arboles de clasificación: es cuando el árbol de decisión es utilizado para predecir variables categóricas y se distribuyen las instancias en clases.
- Los arboles de regresión: es cuando el árbol es utilizado parar predecir variables continuas.

Esta técnica puede considerarse como una serie de reglas que dan forma al árbol. Cada eje está etiquetado con un atributo-valor y las hojas con una clase. Esta técnica se basa en dos tipos de algoritmos: los algoritmos “divide y vencerás”, como el ID3, C4.5 o el CART, y los algoritmos “separa y vencerás”, como el CN2 (Hernández et al., 2004).

A continuación se presentan dos aplicaciones de los árboles de decisión:

“La generación de un modelo credit scoring para una entidad financiera chilena usando árbol de decisión” cuyo desarrollo consistió en presentar dentro de las herramientas de minería de datos, al Credit Scoring, como una de las técnicas para evaluar una solicitud de crédito y poder determinar si la operación es viable o no (Lepin, Ponce, 2012).

Un credit scoring es un sistema que permite la calificación de créditos, para la toma de decisiones con la finalidad de generar una operación de riesgo. Dentro de las ventajas de este sistema se encuentra la reducción en el tiempo de análisis y el aumento en el servicio prestado a los clientes (Domínguez I., 2015).

Esta tesis surge de la necesidad de mejorar la evaluación de riesgo, debido a que la falta de cobro de créditos no permite generar un crecimiento y genera elevados costos en la institución bancaria. Dentro de las ventajas que podemos encontrar están: el Procedimiento

estandarizado para la obtención de un crédito y el contener variables claves para el éxito en la aprobación de un crédito (Lepin, Ponce, 2012).

Otra tesis donde se utilizó el árbol de decisión como herramienta fue en El "diseño de un modelo predictivo de abandono de clientes para una empresa de telecomunicaciones utilizando arboles de decisión" cuyo desarrollo se centró en la calidad de atención que deben poner las empresas con el objetivo de no arriesgar la cartera de clientes existentes, reduciendo para ello el abandono de clientes. (Contreras, Ferreira, 2014).

Por lo tanto la problemática surge por la falta de conocimiento sobre los motivos o causas de abandono de los clientes de servicios de post pago, ya que no existe un recurso o entidad que permita predecirlo. Producto de esto la empresa no genera una acción proactiva de retención de clientes, muy por el contrario establece una acción reactiva, al no contar con estándares o patrones a seguir cuando se produce esta situación (Contreras, Ferreira, 2014).

Para desarrollar este modelo predictivo se utiliza el árbol de decisión de tipo clasificación, debido a que este tipo de técnica permite generar un conjunto de condiciones organizadas y darle un sentido de forma jerárquica que permita predecir mediante reglas un algoritmo determinado. (Contreras, Ferreira, 2014).

Durante esta tesis se desarrollan dos modelos para predecir el abandono de clientes. Un modelo A que permite predecir si un cliente va a hacer abandono voluntario de los servicios prestados. Y un modelo B donde se eliminó la variable CP (Cancelación Permanente) para poder comparar así resultados respectivos. (Contreras, Ferreira, 2014).

En conclusión el modelo B da a conocer las condiciones que se deben dar para el abandono voluntario de clientes, lo que permite poder actuar de forma proactiva ante esta situación y así poder evitar la disminución de la cartera de clientes. El modelo B, en este sentido tiene un mayor campo de acción, ya que el modelo A predice cuando el cliente cancela los servicios prestados, permitiendo acotar la posibilidad de que el cliente será retenido.

El modelo B otorga certeza de que los clientes van a abandonar voluntariamente los servicios, permitiendo un mayor control sobre el abandono de clientes. (Contreras, Ferreira, 2014).

3.2.4. Redes neuronales

Son paradigmas de computación que permiten modelar problemas complejos, donde existen interacciones no lineales entre variables. Las redes neuronales pueden utilizarse en problemas de clasificación, de regresión y de agrupamiento. Se construyen estructurando en serie de capas denominadas entrada, procesamiento o capa oculta y salida, compuestas por nodos o neuronas (Hernández et al., 2004).

Para comprender de mejor manera esta técnica se da a conocer una tesis que utilizo este tipo de herramienta basada en "Una aplicación de redes neuronales artificiales a la predicción y control de la demanda de energía eléctrica en empresas industriales" Esta tesis se generó debido a que muchas empresas están sobrepasando costos de energía eléctrica por

sobre la demanda máxima contratada, lo que se ve acentuado en los elevados montos de facturación mensual. Esto es producto de la muy mala gestión en cuanto a la demanda de energía eléctrica, debido al alto desconocimiento de patrones de consumo, la pérdida de control de los procesos de planificación y mal utilización de herramientas (Ojeda, 2009).

Este problema se pretende resolver mediante un modelo predictivo basado en redes neuronales artificiales, utilizando para ello la implementación del programa de control, donde se generó un modelo matemático de predicción horaria, evaluándose los modelos de regresión lineal, el uso de ARIMA y las redes neuronales (Ojeda, 2009).

Para el desarrollo del modelo predictivo se utilizó una red neuronal del tipo Narx dinámica modificada, ya que permite adecuar correctamente el comportamiento predictivo a los patrones de producción y demanda. Las redes NARX son redes dinámicas con conexiones de retroalimentación. Se usan fundamentalmente para modelar series temporales, que no pueden desarrollarse con redes estáticas o redes dinámicas simples (Ojeda, 2009).

Este trabajo permitió mejorar la gestión de la demanda, al impedir que se produjeran costos de energía no deseados. Además se logró reducir la perdida de producción de plantas productivas (Ojeda, 2009).

Tabla 1: Utilización Técnica de Minería de Datos

Nombre	Predictivo		Descriptivo		
	Clasificación	Regresión	Agrupación	Reglas de asociación	Correlaciones / Factorizaciones
Redes neuronales	●	●	●		
Arboles de decisión ID3, C4.5	●				
Arboles de decisión CART	●	●			
Regresión lineal y logarítmica		●			●
Regresión logística	●			●	
Técnicas bayesianas	●				

Fuente: elaboración propia a partir de Hernández et al., 2004

En la tabla 1 se presentan las características que hacen distintivas a cada técnica de minería de datos.

3.3. Justificación de la técnica de minería de datos

A partir de estos ejemplos de temas de tesis relacionados cada uno de ellos con alguna técnica de minería de datos distinta, permiten generar un resultado común, que es el análisis y la toma de decisiones a partir de una base de datos.

Como se puede observar tanto las técnicas de redes neuronales, técnicas bayesianas, técnicas algebraicas y estadísticas y árboles de decisión permiten identificar, clasificar y predecir conductas o patrones de diversa índole. Sin embargo para el desarrollo de esta tesis se utilizara el árbol de decisión, ya que es una herramienta que presenta una mejor representación gráfica del modelo y que permitirá ver un orden jerárquico de las ideas y condiciones necesarias para agrupar y clasificar a los trabajadores en los modelos respectivos.

Ventajas de Arboles de Decisión según Hernández et al. (2004).

- Aplicables a varias tareas de minería de datos para la clasificación y regresión.
- Tratan con atributos numéricos (continuos) y nominales (discretos).
- Son fáciles de usar.
- Son tolerantes a atributos no significativos de valores faltantes.
- Tienen una representación gráfica que se despliega de mayor a menor detalle, siendo fácil de entender.
- Existen sistemas gratuitos para la construcción de árboles de decisión.

3.4. Arboles de Decisión

Según la revista electrónica ALTONIVEL (2014), en las empresas siempre se están tomando decisiones bajo una estrategia establecida que analice todas las variables. Para facilitarlo existen metodologías que permiten mostrar de manera gráfica el camino más adecuado, en beneficio de la organización y el personal. Estas herramientas son los arboles de decisión que permiten estimar alternativas ante la solución de problemas.

Esta técnica se basa en la construcción de un modelo, hipótesis o representación gráfica, con el fin de ser comprensibles y de representar de manera simbólica el conjunto de condiciones presentes en la base de datos. Además tiene como ventaja que las opciones posibles a partir de una determinada condición son excluyentes, es decir, que permite analizar una situación, y siguiendo las ramas del árbol de decisión apropiadamente, se puede llegar a una acción o decisión a tomar (Hernández et al., 2004).

Para el caso de esta investigación se utilizan dos modelos de predicción. Por lo tanto, los modelos A y B, en donde la predicción es un atributo discreto, se utilizara un árbol de decisión para clasificación.

3.4.1. Arboles de Decisión para Clasificación

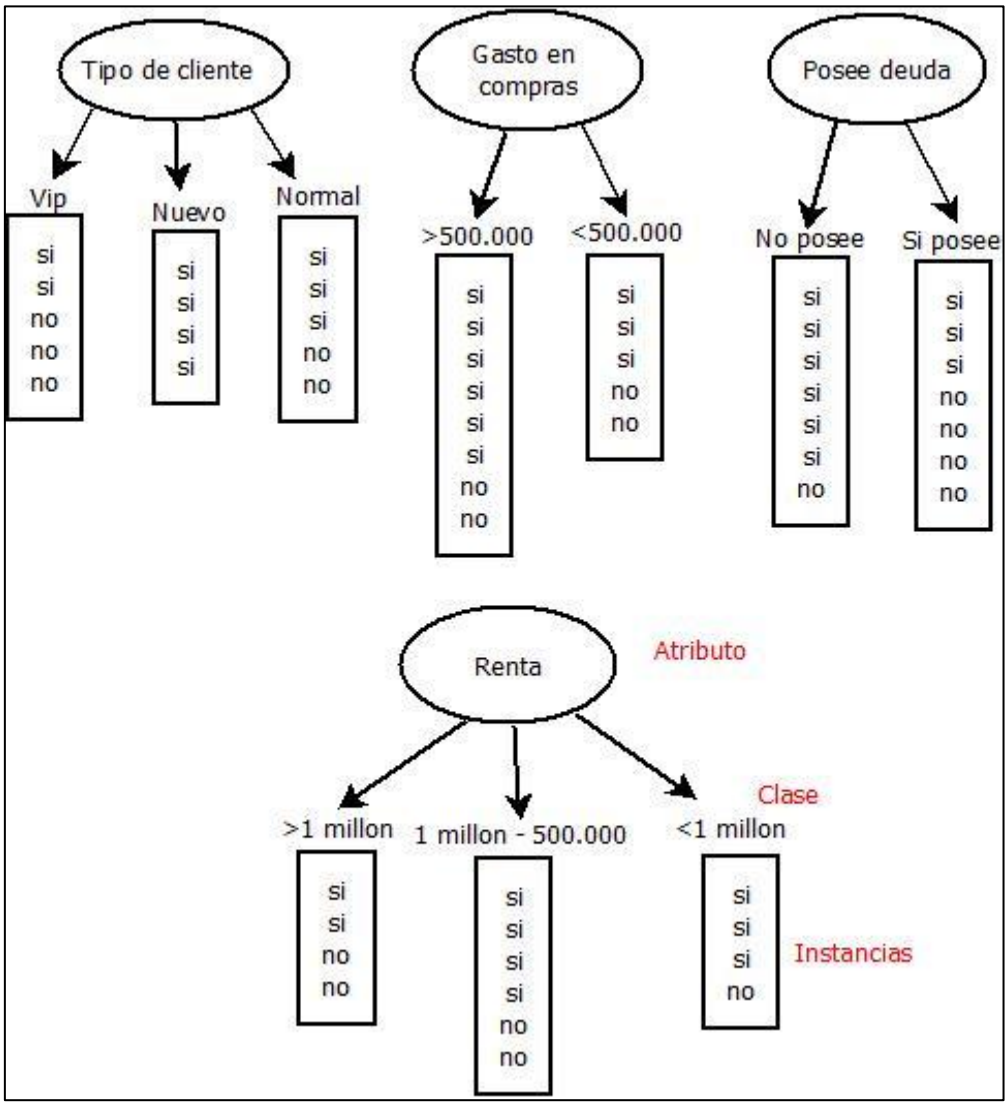
Los algoritmos de árboles de decisión, tales como ID3, C4.5 y CART, están destinados originalmente para los arboles de clasificación. La construcción de un árbol de decisión para clasificación es un proceso en donde se genera una estructura en forma de diagrama de flujo. Cada nodo interno se crea a partir de la clasificación de los atributos, cada rama corresponde a un resultado de la prueba que clasifica a los atributos y cada nodo externo (hoja) es el resultado de la clasificación y genera la variable clase para la predicción. En cada nodo, el algoritmo elige el mejor atributo para dividir los datos y generar clases individuales (Soumen Chakrabarti et al. 2008).

Cuando se usa la técnica de árbol de decisión para la clasificación de los atributos de la base de datos dada, todos los atributos o variables que no aparezcan en el árbol se supone que son irrelevantes para la predicción. Por lo tanto el conjunto de atributos que aparece en el árbol es un subconjunto reducido de la base de datos total (Soumen Chakrabarti et al. 2008).

Para construir el árbol de decisión de esta investigación, se utilizara el algoritmo divide y vencerás. En primer lugar el algoritmo selecciona a un atributo para colocar en el nodo raíz y hace una rama para cada posible valor, esto divide al conjunto de ejemplos en subconjuntos, uno para cada valor del atributo. Luego el proceso se repite de forma recursiva para cada rama del árbol, usando solo aquellos casos que llegan a la rama. Si en un momento todas las instancias en un nodo tienen la misma clasificación, se detiene el desarrollo de esa parte del árbol. Por último, para complementar la generación del árbol es necesario crear un conjunto de reglas, con el fin de abarcar todas las instancias que sean iguales de una clase (Soumen Chakrabarti et al. 2008).

A continuación se muestra un ejemplo para calcular la medida de pureza de los nodos y así determinar que atributo es la mejor opción para ser el nodo raíz del árbol. En la ilustración 7 podemos observar un ejemplo de un árbol de decisión para determinar si una empresa de retail debe ofrecer un producto nuevo a un determinado cliente (Soumen Chakrabarti et al. 2008).

Ilustración 6: Construcción del árbol de decisión



Fuente: elaboración propia a partir de (Soumen Chakrabarti et al. 2008).

En primer lugar el algoritmo decide que nodo formara parte del inicio del árbol, para esto es necesario calcular la pureza de cada nodo. La medida de pureza calculada se llama información y se mide en bits. Asociada con un nodo del árbol, representa la cantidad de información esperada que sería necesaria para especificar si debe seguir dividiendo. Este cálculo tiene las siguientes propiedades y se aplica sobre los nodos de un árbol, utilizando la cantidad de "si" y "no" de una clase como se observa en la ilustración 6 (Soumen Chakrabarti et al. 2008).

- Cuando el número de si o no de una clase es cero, la información obtenida es cero. Por lo tanto, dicha clase no se sigue dividiendo. Como se observa en la ilustración 6, la clase "nuevo" del atributo "tipo de cliente" obtiene una valor de información igual a

cero, porque la instancia "no" es igual a cero. Podemos decir que dicha clase no se debe seguir dividiendo.

- Cuando el número de si y no es igual en una clase, la información alcanza un máximo y dicha clase debe dividirse. Ejemplo info[3,3] = 1 bits.

La función que sirve para calcular la información de los nodos de un árbol de decisión, es el cálculo de la entropía (Soumen Chakrabarti et al. 2008).

Ecuación 1: Función de entropía

$$entropia\ (p_1, p_2, \dots, p_n) = -p_1 \log(p_1) - p_2 \log(p_2) \dots - p_n \log(p_n)$$

Como se puede ver en la ecuación 1, los signos negativos en la función se deben a que los logaritmos de las fracciones $p_1, p_2 \dots, p_n$ son negativos, por lo que la entropia da como resultado un valor positivo (Soumen Chakrabarti et al. 2008).

Los términos p_1, p_2 ... de la fórmula de entropía se deben expresar en fracciones que sumadas den como resultado uno, de manera que al igualarla con la función del cálculo de la información sea igual a: (Soumen Chakrabarti et al. 2008).

$$info\ ([1,2]) = entropia\ (\frac{1}{3}, \frac{2}{3})$$

De esta forma podemos comenzar a calcular la información de cada nodo expuesto en la ilustración 6. Por lo tanto, para la clase "vip" del atributo "tipo de cliente", la información es la siguiente:

$$info[2,3] = -\frac{2}{5} * \log\left(\frac{2}{5}\right) - \frac{3}{5} * \log\left(\frac{3}{5}\right) = 0{,}971$$

Con el valor obtenido se puede decir que dicha clase debe seguir dividiéndose. Es necesario mencionar que los logaritmos utilizados en las fórmulas son de base 2.

Siguiendo la fórmula de entropía, se puede calcular la medida de información de cada clase en los diferentes atributos del ejemplo expuesto en la ilustración 6. Los valores de información para cada atributo son:

- Tipo de cliente:

$$\text{Vip: } info\ ([2,3]) = 0{,}971\ bits$$
$$Nuevo: info\ ([4,0]) = 0{,}0\ bits$$
$$Normal: info\ ([3,2]) = 0{,}971\ bits$$

- Gastos en compras:

$$> 500.000: info[6,2] = 0{,}811$$
$$< 500.000: info[3,2] = 0{,}971$$

Así sucesivamente hasta calcular la medida de información de cada clase en los diferentes atributos. Además podemos observar que mientras más uniformes las instancias de una clase, mayor será su entropía.

Ahora para evaluar y determinar que nodo formara parte de la raíz del árbol, es decir, para determinar al atributo que mejor clasifica a los datos, se debe analizar la ganancia de información. La ganancia de información se produce cuando la división de un nodo envía instancias distintas al siguiente nodo. Entonces para seleccionar dicho nodo y sus divisiones, se debe elegir al nodo que produzca la mayor ganancia de información lo que es equivalente a seleccionar al nodo con la menor cantidad de entropía ponderada.

La fórmula para calcular la ganancia de información de cada nodo presentado en la ilustración 6 es:

Ecuación 2: Ganancia de información

$$ganancia(nodo) = info(nodo) - info(ponderada\ de\ las\ clases)$$

Primero se calcular la información del nodo, se debe tomar el total de si y no que existen en cada clase. De la ilustración 6 podemos observar que el nodo "tipo de cliente" posee 9 si y 5 no. Por lo tanto la $info(tipo\ de\ cliente)$ es:

$$info[9,5] = 0{,}94\ bits$$

En segundo lugar es necesario calcular el valor ponderado de información de las clases presentes en el nodo. Se debe tener en cuenta el número total de instancias que descienden de cada clase, siguiendo con el nodo "tipo de cliente", podemos observar que descienden 5 instancias de la primera y tercera clase, mientras que en la segunda clase descienden solo 4 instancias. Además la suma de todas las instancias da como resultado 14 instancias en total. Por lo tanto el cálculo de la información ponderada de las clases es de la siguiente manera:

$$info[2,3],[4,0],[3,2] = \frac{5}{14} * 0{,}971 + \frac{4}{14} * 0 + \frac{5}{14} * 0{,}971 = 0{,}693\ bits$$

Ahora la ganancia para el nodo "Tipo de Cliente" queda de la siguiente forma:

$$ganancia(Tipo\ de\ Cliente) = info([9,5]) - info([2,3],[4,0],[3,2]) = 0{,}940 - 0{,}693 = 0{,}247\ bits$$

Luego el camino a seguir para la construcción del árbol se realiza calculando la ganancia de información para cada atributo y para posteriormente elegir el atributo que gana la mayoría de la información para dividirse sucesivamente (Soumen Chakrabarti et al., 2008).

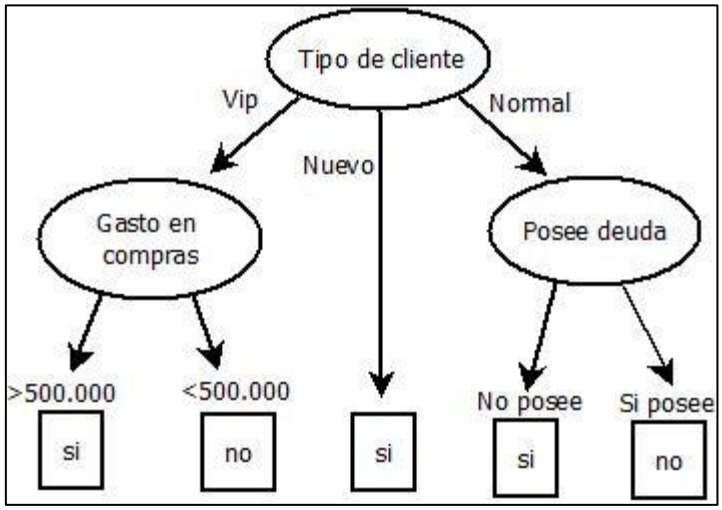
A continuación se presentan los valores de ganancia de información para cada atributo expuesto en la ilustración 6.

$$Ganancia(tipo\ de\ cliente) = 0{,}247\ bits$$
$$Ganancia(gastos\ en\ compras) = 0{,}152\ bits$$
$$Ganancia(posee\ deudas) = 0{,}048 bits$$
$$Ganancia(renta) = 0{,}029\ bits$$

De los valores de ganancia de cada atributo, el algoritmo seleccionara al atributo tipo cliente como nodo raíz, ya que, tiene la mayor ganancia de información. Luego el algoritmo seleccionara al atributo "gastos en compras" como la siguiente mejor opción para seguir construyendo el árbol debido a que produce un nodo casi completamente puro o uniforme y es el segundo atributo con la mayor ganancia de información. Como no hay necesidad de seguir dividiendo este nodo, se da por terminada la rama.

Continuado con la aplicación de los pasos anteriormente mencionados, se obtiene el árbol de la ilustración 7. Idealmente el proceso termina cuando todos los nodos hojas son puros, es decir, cuando las instancias que contienen son de la misma clasificación. (Soumen Chakrabarti et al., 2008).

Ilustración 7: Árbol de decisión finalizado



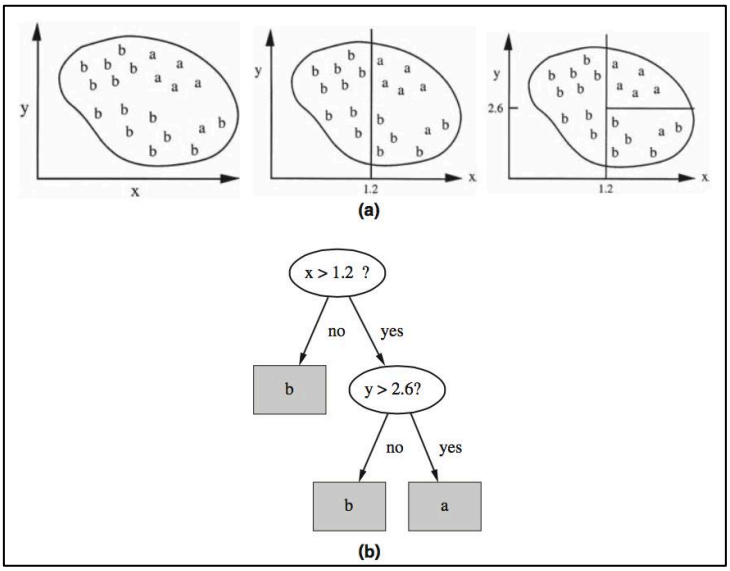
Fuente: elaboración propia a partir de Soumen Chakrabarti et al., (2008).

Como hemos visto el algoritmo divide y vencerás va construyendo el árbol de arriba hacia abajo, buscando en cada etapa un atributo para dividir y que mejor separe a las clases. También es necesario crear un conjunto de reglas que complemente la generación del árbol.

Una forma de hacer estas reglas es tomar cada clase y buscar una forma de cobertura que abarque todas las instancias de una misma clasificación. Esto se llama enfoque de cobertura porque en cada etapa se identifica una regla que cubra a algunos casos. Este enfoque conduce a un conjunto de reglas dentro del árbol de decisión y que abarca a cada nodo (Soumen Chakrabarti et al., 2008).

El método de cobertura se puede visualizar fácilmente observándolo en un espacio de dos dimensiones como se muestra en la ilustración 8. Primero se hace una regla que cubra a todas las instancias "a". Para la primera prueba en la regla, se debe dividir el espacio verticalmente como se muestra en grafico superior (a) de la ilustración 8. Es así como comienzan a construirse las reglas del árbol de decisión (Soumen Chakrabarti et al., 2008).

Ilustración 8: Algoritmo de cobertura, en donde (a) es la cobertura de las instancias y (b) es el árbol de decisión para el mismo problema.



Fuente: obtenida de Soumen Chakrabarti et al., 2008.

Por lo tanto la regla generada de la ilustración 8 para clasificar la instancia a del árbol de decisión es (Soumen Chakrabarti et al., 2008).

$$Si\ x > 1{,}2\ \wedge y > 2{,}6 => clase = a$$

Mientras que para la clasificación de b se generan dos reglas (Soumen Chakrabarti et al., 2008).

$$Si\ x \leq 1{,}2 => clase = b$$
$$Si\ x > 1{,}2\ \wedge y \leq 2{,}6 => clase = b$$

El algoritmo divide y vencerás opera mediante la sumatoria de pruebas al árbol que se está construyendo, siempre con el objetivo de maximizar la separación entre las clases. Al aplicar estos algoritmos la idea es poder juntar la mayor cantidad de instancias de la clase deseada como sea posible y excluir la mayor cantidad de instancias de la otra clase. Es así como se va construyendo el árbol de decisión mediante estas diferentes reglas para poder realizar una mejor clasificación y posterior predicción (Soumen Chakrabarti et al., 2008).

Por ultimo para la validación de dicho árbol de clasificación creado por el algoritmo, es necesario analizar la matriz de confusión creada a partir de las predicciones del árbol.

Una matriz de confusión contiene información sobre las clasificaciones actuales y previstas realizadas por un sistema de clasificación. El rendimiento de estos sistemas se puede evaluar usando los datos en la matriz. La Tabla 2 muestra la matriz de confusión para un clasificador de dos clases. (Kohavi y Provost, 1998).

Las entradas de la matriz de confusión tienen el siguiente significado en el contexto de nuestro estudio (Kohavi y Provost, 1998):

- a es el número de predicciones correctas donde la instancia es negativa.
- b es el número de predicciones incorrectas donde una instancia es positiva.
- c es el número de predicciones incorrectas donde una instancia es negativa
- d es el número de predicciones correctas donde una instancia es positiva.

Tabla 2: Matriz de Confusión

		Predicción	
		negativo	positivo
Actual	negativo	a	b
	positivo	c	d

Fuente: Elaboración propia a partir de Kohavi y Provost.

La exactitud (accuracy), denominada por las letras AC es la proporción del número total de predicciones que eran correctas. Se determina usando la ecuación:

Ecuación 3: Cálculo de exactitud

$$AC = \frac{\mathrm{a} + \mathrm{d}}{\mathrm{a} + \mathrm{b} + \mathrm{c} + \mathrm{d}}$$

El error representa un porcentaje de datos erróneos que arroja el modelo en comparación al total de los datos y se puede calcular mediante la fórmula:

Ecuación 4: Cálculo de error

$$\text{Error} = \frac{c + b}{a + b + c + d}$$

La tasa de verdaderos positivos (recall), se le asignan las letras TP y es la proporción de casos positivos que fueron identificados correctamente, se calcula mediante la siguiente ecuación:

Ecuación 5: Cálculo de verdaderos positivos

$$TP = \frac{d}{c + d}$$

La tasa de falsos positivos (FP) es la proporción de casos negativos que fueron clasificados incorrectamente como positivos, la ecuación es la siguiente:

Ecuación 6: Cálculo de falsos positivos

$$FP = \frac{b}{a + b}$$

La tasa de verdaderos negativos (TN) se define como la proporción de casos negativos que fueron clasificados correctamente, se calcula mediante la siguiente ecuación:

Ecuación 7: Cálculo de verdaderos negativos

$$TN = \frac{a}{a + b}$$

La tasa de falsos negativos (FN) es la proporción de casos positivos que fueron clasificados incorrectamente como negativos, se calcula mediante la siguiente ecuación:

Ecuación 8: Cálculo de falsos negativos

$$FN = \frac{c}{c + d}$$

Por último, la precisión (P) es la proporción de los casos positivos predichos que eran correctos, se calcula mediante la siguiente ecuación:

Ecuación 9: Cálculo de precisión

$$P = \frac{\mathrm{d}}{d + b}$$

Resulta muy necesario aprender de los errores originados por un modelo de clasificación, ya que estos representan la diferencia entre lo que el modelo puede llegar a predecir y lo que el resultado real resulta ser. (Soumen Chakrabarti et al., 2008).

Supongamos que un modelo mostro un resultado binario, tomando valores de 0 y 1. Para este caso el valor absoluto del error solo puede ser 0 o 1 y para este efecto la mejor manera de ver el rendimiento del modelo será mediante la matriz de confusión, ya que esta permite conocer el comportamiento de un modelo de clasificación (Soumen Chakrabarti et al., 2008).

La matriz de confusión permite ordenar los casos del modelo en diferentes categorías. Además es una herramienta importante para poder evaluar e interpretar los resultados de la predicción, ya que hace que resulte fácil entender y aprender de los errores. (SQL server, 2014).

Para entender mejor la interpretación de cada celda de la matriz de confusión, a continuación se explica mediante un ejemplo en que consiste cada celda. Este ejemplo consiste en como predecir que clientes tiene mayor probabilidad de comprar una bicicleta. Para esto se considera un conjunto de datos para los que ya se conocen los valores de los atributos. Solo hay dos resultados posibles, la probabilidad de que una persona compre una bicicleta y por otro lado la probabilidad de que una persona no compre una (SQL server, 2014).

Tabla 3: Ejemplo de matriz de confusión

Predicción	0 (Real)	1 (Real)
0	362	144
1	121	373

Fuente: Elaboración propias a partir de SQL server 2014.

En la tabla 3 se observa que la primera celda de resultados, que contiene el valor 362, representa el número de verdaderos positivos para el valor 0. Ya que el 0 indica que el cliente no compro la bicicleta, por lo tanto este valor estadístico indica que el modelo predijo correctamente las personas que no compraron bicicletas (SQL server, 2014).

Por otro lado la celda que se encuentra por debajo y que contiene un valor asignado de 121, indica el número de falsos positivos o representa la cantidad de veces que el modelo predijo que un persona compraría una bicicleta cuando en realidad no lo hizo (SQL server, 2014).

La celda que posee el valor 144 representa a los falsos positivos para el valor 1, ya que 1 significa que el cliente efectivamente compro la bicicleta. Por lo tanto este valor quiere decir que en 144 ocasiones el modelo predijo que alguien no compraría una bicicleta, cuando en verdad si lo hizo (SQL server, 2014).

Por último la celda que contiene el valor 373, representa el número de verdaderos positivos para el valor 1, por lo tanto el modelo predice que en 373 ocasiones alguien compraría una bicicleta (SQL server, 2014).

Al sumar de forma diagonal los valores podremos determinar el valor de predicciones exactas y por el contrario de la diagonal Obtendremos el valor de predicciones incorrectas. Para este caso serían 735 las predicciones correctas y 265 las incorrectas (SQL server, 2014).

4. Desarrollo del Modelo

Es durante esta etapa donde podemos ver de mejor manera el carácter iterativo del proceso de minería de datos, ya que será necesario probar modelos alternativos hasta llegar al ideal para resolver el problema. El desarrollo de los modelos predictivos debe contener bien definidos los planes de acción para las etapas de entrenamiento y validación, ya que depende de estas la consistencia y mayor predicción del modelo (Hernández et al., 2004).

Los modelos predictivos permiten estimar cual será el comportamiento esperado de los trabajadores utilizando los datos disponibles. Donde resulta muy importante definir correctamente las variables predictoras, que son las que permitirán definir la variable objetivo (variable de clase).

Para el desarrollo de este modelo se siguieron las etapas de la metodología CRISP explicada con anterioridad en el marco conceptual.

Fase de comprensión de los datos: durante esta fase se recopilo la mayor cantidad de información referente a los campos o columnas a utilizar, por ejemplo la edad, nombre completo, Rut. Toda esta información se agrupo y se consolido con el objetivo de dar forma a una base de datos general que agrupara las características específicas de distintos trabajadores de la empresa. Para llevar esta etapa a cabo, fue necesario extraer la información del sistema de información Payroll y adecuar la base de datos a nuestras necesidades. Una vez extraída la base de datos, se procedió a integrar la información obtenida de distintas partes del sistema de información. Con el objetivo de desarrollar datos en base a una estructura, donde cada dimensión de datos corresponde a un atributo.

Fase de preparación de los datos: de la base de datos total obtenida de la empresa, es necesario eliminar los datos irrelevante, faltantes y en blanco, es decir, eliminar aquellos datos que se repiten con menor frecuencia o que en alguno de sus atributos no contengan información. También es necesario agrupar los datos para una mejor interpretación, como por ejemplo, la variable clase utilizada en el modelo de clasificación denominada “termino de contrato” y que significa que un trabajador es desvinculado de la empresa. Esta variable se construyó a partir de la agrupación de las instancias renuncia voluntaria, mutuo acuerdo, necesidades de la empresa y no concurrencia al lugar de trabajo entre otras.

Una vez lista la tarea de limpieza, fue necesario crear nuevos atributos con la finalidad de describir al trabajador de manera completa dentro de la empresa, uno de los atributos creados fue el de años de trabajo y si posee o no hijo menor. Los años de trabajo se crearon a partir de las fechas de ingreso y retiro de un trabajador, mientras que el atributo hijo menor se construyó a partir de una base de datos diferente y que hace referencia al núcleo familiar del colaborador.

Las variables relevantes que se seleccionaron para construir el modelo se pueden observar en la Tabla 1. El objetivo de esta fase fue crear una base de datos nueva para hacer

más fácil la construcción del modelo y que los resultados del árbol de decisión sean interpretativos.

Fase de modelamiento: una vez lista la base de datos depurada, fue necesario definir las variables para construir el árbol de decisión. Dentro de estas variables, se encuentran las variables predictoras. Este tipo de variables permiten predecir la variable objetivo del modelo, que es el propósito al cual se quiere llegar. En nuestro caso es el ausentismo laboral y la rotación de trabajadores.

Durante la fase de preparación de datos, se les asignaron nomenclaturas a las variables predictoras, como por ejemplo: tipo de empresa (TDE), edad de trabajadores (ED), entre otras con el objetivo de disminuir los tiempos de iteración del software R, como se observa en la tabla 4.

Tabla 4: Variables predictoras para el desarrollo del modelo

	Variables Predictoras	**Simbología**	**Tipo de Variable**
1.	Tipo de Empresa	TE	Discreta
2.	Edad de Trabajador	ET	Discreta
3.	Tipo de Cargo	TC	Discreta
4.	Localidad	Localidad	Discreta
5.	División	División	Discreta
6.	Sexo	Sexo	Discreta
7.	Causal de Ausentismo	CA	Discreta
8.	Días de Ausencia	DA	Nominal
9.	Mes Presentación Licencia	MPL	Nominal
10.	Estado Civil	EC	Nominal
11.	Nivel de Estudios	NE	Discreta
12.	Hijo Menor	HM	Discreta
13.	Número de licencias presentadas	LP	Discreta
14.	Años de Trabajo	AT	Discreta

Fuente: Elaboración propia a partir de información entregada por empresa de servicios

También es necesario definir la variable clase que consiste en el objetivo que se quiere predecir, y que en definitiva permite resolver el problema. A continuación se describen las variables clases para los dos modelos a desarrollar (Modelo A y B). La variable clase para el modelo A consiste en el estado de los trabajadores dentro de la empresa y se define como MR. A esta variable se le asignan dos valores, el CV (Contrato Vigente) y el RE (Termino de

Contrato). La variable para el modelo B consiste en los días de ausencia de los trabajadores y se define como DA. Al igual que para el primer modelo la variable a predecir consiste en una variable discreta y se clasifican por tramos los días de ausencia de los trabajadores. Estos tramos se representan por A, B y C, los cuales comprenden entre 0 a 1 días (Tramo A), 2 a 7 días (Tramo B) y de 8 o más (Tramo C) respectivamente. Este modelo representa un árbol de clasificación con el objetivo de predecir los días (entre tramos) en que los trabajadores faltaran a su lugar de trabajo.

Fase de despliegue del modelo: durante esta fase se pretende que a partir de la construcción y validación del modelo se logren analizar las acciones a partir de los resultados obtenidos. Para esta fase es fundamental el uso y control que se le dé al modelo construido. De manera de actualizar y mantener vigente la estructura y aplicación a nuevas bases de datos. Ya que los patrones pueden cambiar a diversas condiciones en que se mueva una empresa. Resulta clave el saber manejar y explicar los resultados que el modelo indica, ya que este punto será clave para generar mejoras al interior de una organización.

4.1. Construcción del Modelo

4.1.1. Modelo A: Modelos de clasificación para predecir el estado del trabajador en la empresa

Para el desarrollo de este modelo se emplearon las variables predictoras antes mencionadas en la tabla 4, menos las variables: Causal de Ausentismo y Mes Presentación Licencia. La primera no se utilizó para el desarrollo del modelo, sin embargo se utilizó para desarrollar otra variable, correspondiente a la cantidad de licencias presentadas por el trabajador. Por otro lado la variable mes de presentación de licencia no se utilizó dada la dificultad para poder clasificar y agrupar esa condición con cada trabajador.

Para construir el árbol de decisión se utilizaron dos conjuntos de datos, el 75% para el entrenamiento y el otro 25% para la prueba del modelo. Estos porcentajes son utilizados de manera estándar para la construcción de este tipo de modelo. Donde se consideró un total de 12 variables predictoras para una base de datos de 4455 trabajadores.

Para el modelo A se ocupa la variable clase MR y se clasifica en contrato vigente (CV) y termino de contrato (RE).

4.1.2. Modelo B: Modelo de Clasificación para predecir los días de ausentismo que presentara un trabajador

Para la construcción de este modelo se utilizaron las variables de la tabla 4, además se agregaron las variables FPI que describe la diferencia entre los días en que el trabajador presento su licencia y el comienzo de dicha licencia. También se agregó la variable MR para clasificar si el trabajador esta con contrato vigente o fue desvinculado de la empresa.

Para el desarrollo del árbol de clasificación se utilizaron un total de 5.164 datos y 15 variables predictoras. Para la construcción del árbol se utilizó el 75% de los datos para el entrenamiento del modelo y el otro 25% para generar la prueba. La base de datos total

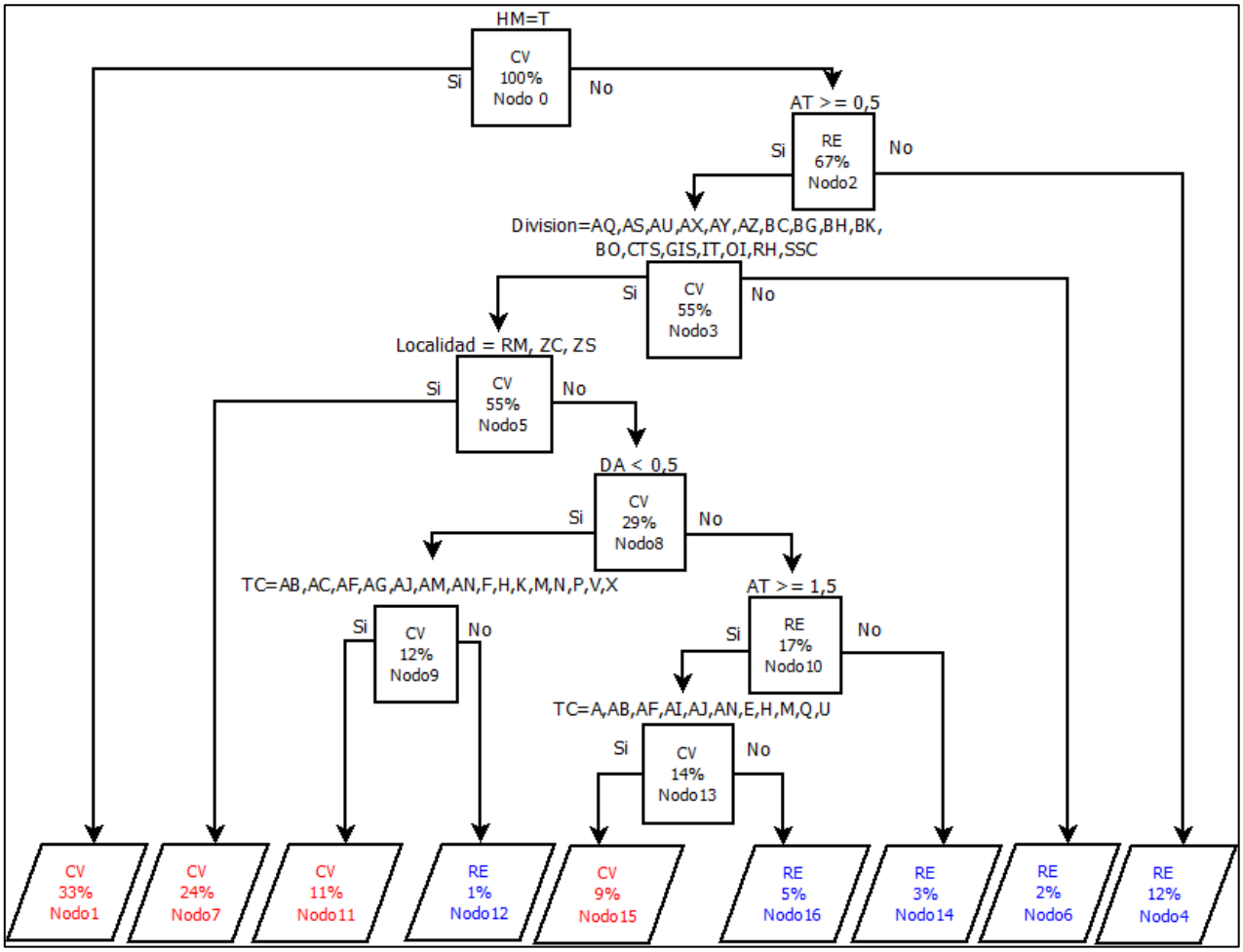
representa las causales de ausentismo de los trabajadores y sus días de ausencia, sin embargo un mismo trabajador puede presentar más de una causal de ausentismo en el transcurso del periodo en el que se extrajo la información. Por lo tanto la dotación total de trabajadores que presentan ausentismo es 1024.

En este caso, el modelo utiliza la variable clase DA (días de ausentismo) y da como resultado un numero comprendido dentro de uno de los tres tramos de clasificación.

5. Resultados

5.1. Resultados del modelo A

Ilustración 9: Modelo A para clasificación entre contrato vigente (CV) y termino de contrato (RE)



Fuente: elaboración propia

La ilustración 9 representa un árbol de decisión que permite predecir si el trabajador quedara con contrato vigente o terminara su contrato. Para ello lo acompañan variables predictoras que permiten clasificar mediante características, patrones o conductas hasta llegar a variadas decisiones. Las variables usadas para la construcción del modelo son; años de trabajo, días de ausencia, hijo menor, localidad y el tipo de cargo.

En definitiva el modelo comprende 9 divisiones de decisión, de estas 5 llegan a un término de contrato y 4 a un contrato vigente.

Las reglas de decisión que resultan de la interpretación del árbol se presentan de la forma si (condición) – entonces (decisión). Debido a que la rotación de personal se produce cuando un trabajador deja su lugar de trabajo (termina su contrato con la empresa), las reglas de decisión para entender la rotación de personal de la empresa, se deben centrar en todos aquellos casos en donde el trabajador llega a la clasificación termino de contrato (RE). Las reglas que se obtienen del árbol de clasificación presentado anteriormente son las siguientes:

- Si $(HM \neq t) \wedge (AT < 0{,}5) \Rightarrow RE$ (Termino de Contrato). Comprenden los nodos 0, 2 y 4.
  Es decir, si un trabajador no tiene un hijo menor a 18 años (HM) y sus años de trabajo (AT) no superar los 0,5 años (5 meses), entonces el trabajador va a renunciar a la empresa.

Para hacer más fácil la interpretación de la regla de clasificación se dejara a las divisiones $AQ, AS, AU, AX, AY, AZ, BC, BG, BH, BK, BO, CTS, GIS, IT, OGC, OI, RH, SS$ como una sola variable y se llamara "X". Estas siglas corresponden a las divisiones administración (AQ), analítica (AS), automotriz (AU), división club deportivo (AX), environmental (AY), finanzas (AZ), inspectores (BC), metalurgia (BG), mineralogía (BH), outsourcing (BK), servicios estratégicos (BO), sistemas de pruebas de consumo o comsumer testing system (CTS), gestión integral de sistemas (GIS), innovación tecnológica (IT), oil gas y chimical (OGC), OI, recursos humanos (RH) y S&SC (SSC).

- Si $HM \neq t \wedge AT \geq 0{,}5 \wedge division \neq X \Rightarrow RE$ (Termino de contrato). Pasa por los nodos 0, 2, 3 y 6.
  Es decir, si el trabajador no tiene hijo menor (HM), sus años de trabajo (AT) superan los 0,5 años y no pertenece en las divisiones de X antes mencionadas, entonces el trabajador va a terminar su contrato. Si no pertenece a las divisiones de X (División ≠X), quiere decir que pertenece a las divisiones que faltan por mencionar, las cuales son: agricultura (AR) as (AT), comercial (AV), geometalurgia (BA), legal (BE), Management (BF) y proyectos (BI).

- Si $HM \neq t \wedge AT \geq 0{,}5 \wedge division = X \wedge localidad \neq RM, ZC, ZS \wedge DA < 0{,}5 \wedge TC \neq AB, AC, AF, AG, AJ, AM, AN, F, H, K, M, N, P, V, X \Rightarrow RE$ (Termino de contrato).
  Comprende a los nodos 0, 2, 3, 5, 8, 9, 12.
  Es decir, si el trabajador no tiene hijo menor (HM), sus años de trabajo (AT) superan los 0,5 años, pertenece a una de las divisiones de X, su localidad de trabajo es en la zona norte (Localidad ≠ RM, ZC, ZS), presenta días de ausencia (DA) menor a 0,5 y tiene un cargo distinto a (Muestrero, Operador, Prevencionista, Químico, Soldador, Supervisor, Técnico, Auditor, Ayudante, Chofer, Coordinador, Ejecutivo, Encargado, Ingeniero, Jefe), es decir, el trabajador pertenece a uno de los cargos que faltan por mencionar, entonces el trabajador terminara su contrato.

- Si $HM \neq t \wedge AT \geq 0{,}5 \wedge division = X \wedge localidad \neq RM, ZC, ZS \wedge DA > 0{,}5 \wedge AT < 1{,}5 \Rightarrow RE$ (Termino de contrato)
  Es decir, si un trabajador no tiene hijo menor (HM), sus años de trabajo (AT) superan los 0,5 años, pertenece a una de las divisiones de X, su localidad de trabajo es en la

zona norte (Localidad ≠ RM, ZC, ZS), presenta días de ausencia (DA) mayor a 0,5 y sus años de trabajo (AT) no superan los 1,5 años, entonces se generara un término de contrato.

- Si $HM \neq t \wedge AT \geq 0{,}5 \wedge division = X \wedge localidad \neq RM, ZC, ZS \wedge DA > 0{,}5 \wedge AT \geq 1{,}5 \wedge ET > 30 \wedge TC \neq A, AB, AF, AI, AJ, AN, E, H, M, Q, U => RE$ (Termino de contrato).
  Es decir, si un trabajador no tiene hijo menor (HM), sus años de trabajo (AT) superan los 0,5 años, pertenece a una de las divisiones de X, su localidad de trabajo es en la zona norte (Localidad ≠ RM, ZC, ZS), presenta días de ausencia (DA) mayor a 0,5, sus años de trabajo (AT) superan los 1,5 años, tiene una edad (ET) superior a 30 años y no pertenece a uno de estos cargos (Abogado, Muestrero, Prevencionista, Refinador, Soldador, Técnico, Asistente, Ayudante, Coordinador, Espectroscopista, Gerente), quiere decir que pertenece a los cargos que faltan por mencionar, entonces el trabajador terminara su contrato.

Para explicar de mejor manera la distribución de los datos desde el inicio del árbol hasta cada uno de los nodos terminales y además entender el porqué de las clasificaciones en los nodos al final del árbol (CV y RE), se presenta a continuación la descripción de cada nodo que se observa en el árbol de clasificación de la ilustración 9.

Nodo 1: Contiene 1094 datos que cumplen con la condición del nodo 0, donde HM (hijo menor a 18 años) es igual a T (True = Verdadero). El 98,72% de los datos son de contrato vigente (CV), mientras que el 1,27% son de término de contrato (RE), por lo tanto, el nodo no se sigue dividiendo y tiene la clasificación de contrato vigente (CV).

Nodo 2: Tiene un total de 2247 datos que cumplen con la condición de HM distinto de T (no tiene hijo menor de 18 años). De estos datos el 48,24% corresponden a contrato vigente (CV), mientras que el 51,75% son de término de contrato (RE).

Nodo 3: Tiene 1824 datos en total, que cumplen con la condición del nodo años de trabajo mayor a 0,5 años (AT>=0,5). Se dividen en 58,33% de datos pertenecientes a contrato vigente CV y un 41,66% de los datos restantes correspondientes a término de contrato (RE).

Nodo 4: Contiene 423 datos que cumplen con la condición años de trabajo menor a 0,5 años (AT<0,5), de los cuales un 4,72% corresponden a contrato vigente CV y el 94,27% son de término de contrato (RE), por lo tanto, este nodo no se sigue dividiendo y tiene la clasificación término de contrato (RE).

Nodo 5: Tiene 1772 datos que cumplen con la condición del nodo 3 división= AQ, AS, AU, AX, AY, AZ, BC, BG, BH, BK, BO, CTS, GIS, IT, OGC, OI, RH, SSC. Un 59,98% de los datos corresponden a contrato vigente (CV), mientras que el 40,01% son término de contrato (RE).

Nodo 6: A este nodo llegan 52 datos que cumplen con la condición división ≠ AQ, AS, AU, AX, AY, AZ, BC, BG, BH, BK, BO, CTS, GIS, IT, OGC, OI, RH, SSC, de los cuales un

1,92% son contrato vigente (CV) y el 98,07% son de término de contrato (RE), por lo tanto este nodo pasa a ser un nodo terminal con la clasificación de término de contrato (RE) y no se sigue dividiendo.

Nodo 7: Contiene 795 datos que cumplen con la condición del nodo 5, es decir, que el trabajado pertenece a una de estas localidades: región metropolitana, zona centro, zona sur (localidad=RM, ZC, ZS). Además el 70,31% de los datos pertenece a contrato vigente (CV) y 29,68% restante es del conjunto de término de contrato (RE). Por lo tanto es un nodo terminal con la clasificación contrato vigente (CV) y no se sigue dividiendo.

Nodo 8: Tiene un total de 977 datos que cumplen con la condición (localidad ≠ RM, ZC, ZS), es decir, el trabajador pertenece a la zona norte. Además el 51,58% de los datos son contrato vigente (CV) y el otro 48,41% es de término de contrato (RE). Por lo tanto el nodo debe seguir dividiéndose.

Nodo 9: Contiene 420 datos en total que cumplen con la condición del nodo 8, días de ausencia menor a 0,5 días (DA<0,5). El 60,71% de los datos pertenece a la clasificación contrato vigente (CV), mientras que el 39,28% es de término de contrato (RE).

Nodo 10: Este nodo tiene 557 datos que cumplen con la condición, días de ausencia mayo o igual a 0,5 (DA>=0,5) presente en el nodo 8. El 44,7% de los datos es contrato vigente y el otro 55,29% pertenece a término de contrato.

Nodo 11: A este nodo llegan 291 datos que cumplen con la condición del nodo 9 (TC= AB, AC, AF, AG, AN, F, H, J, K, M, P, X, Z), es decir el trabajador pertenece a uno de estos cargos; muestrero (AN), operador (AC), prevencionista (AF), químico (AG), técnico (AN), auditor (F), ayudante (H), capataz (J), chofer (K), coordinador (M), encargado (P) . De estos datos el 69,07% corresponde a contrato vigente (CV), mientras que el 30,92% restante es término de contrato (RE). Este es un nodo terminal con la clasificación contrato vigente (CV).

Nodo 12: A este nodo llegan 129 datos que cumplen con la condición tipo de cargo (TC= A, AI, AM, B, D, E, G, Ins, N, R, V, W, Y). De estos datos el 41,86% corresponde a contrato vigente (CV) y el otro 58,13% es término de contrato (RE). Este nodo pasa a ser un nodo terminal con la clasificación término de contrato (RE).

Nodo 13: Tiene 451 datos que cumplen con la condición del nodo 10, años de trabajo mayor a 1,5 años (AT>=1,5). El 52,54% de los datos pertenecen a contrato vigente (CV), mientras que el 47,45% restante es término de contrato (RE).

Nodo 14: A este nodo llegan 106 datos que cumplen con la condición años de trabajo menor a 1,5 años (AT<1,5). De estos datos el 11,32% es contrato vigente (CV) y el 88,67% pertenece a término de contrato (RE), por lo tanto, este es un nodo terminal con la clasificación término de contrato (RE).

Nodo 15: A este nodo llegan 399 datos que cumplen con la condición del nodo 13 (TC=A, AB, AF, AI, AJ, AN, E, H, M, Q, U), es decir, el trabajador pertenece a uno de estos

cargos; abogado (A), muestrero (AB), prevencionista (AF), refinador (AI), soldador(AJ), técnico (AN), asistente (E), ayudante (H), coordinador (M), espectroscopista (Q), gerente (U). De estos datos el 55,63% pertenecen a contrato vigente (CV), mientras que el 44,36% restante pertenece a término de contrato (RE). Este es un nodo terminal con la clasificación contrato vigente (CV).

Nodo 16: A este nodo llegan 52 datos que no cumplen con la condición del nodo 13 (TC ≠ A, AB, AF, AI, AJ, AN, E, H, M, Q, U), es decir, el trabajador pertenece a alguno de los cargos que están mencionados en esta regla. El 15,21% de los datos pertenece a contrato vigente (CV) y el 84,78% es de término de contrato (RE), por lo tanto, este pasa a ser un nodo terminal con la clasificación término de contrato (RE).

Como observación a la descripción de los nodos, podemos ver claramente que cuando se presentaba una gran diferencia de porcentajes entre las dos clasificaciones (CV y RE), el algoritmo terminaba de dividir el árbol y a dicho nodo le asignaba la clasificación con mayor porcentaje. Mientras que cuando los porcentajes de un nodo son similares, el algoritmo generaba otra división para intentar dejar una sola clasificación dominante dentro del nodo.

5.1.1. Matriz de Confusión

Al construir el modelo en el software R y generar la predicción, este arrojo la siguiente matriz de confusión. En cada columna se representan el número de predicciones de cada clase, mientras que en las filas se representan las instancias reales de cada clase.

Tabla 5: Matriz de confusión del modelo A

		Predicción	
		CV	RE
Real	CV	692	64
	RE	140	218

Fuente: elaboración propia.

En la tabla 5 se muestra el número de trabajadores que el modelo corrigió correcta e incorrectamente. Podemos interpretar la tabla 5 de la siguiente forma:

- El modelo predijo correctamente que 692 trabajadores tienen contrato vigente. Además la predicción estimo a un total de 140 trabajadores con contrato vigente, siendo que en la realidad estos trabajadores se encontraban con término de contrato.
- La predicción arrojo a 64 trabajadores con término de contrato, siendo que en la realidad estos trabajadores se encuentran con contrato vigente. Por otro lado el modelo predijo correctamente que 218 trabajadores están con término de contrato.

Podemos decir que los verdaderos positivos y verdaderos negativos superan en número a los falsos positivos y falsos negativos, cumpliendo con la regla para poder decir que el modelo tiene una predicción acertada, con una menor cantidad de datos erróneos.

5.1.2. Validación Modelo A

Para validar el modelo A se realizaron 10 repeticiones distintas en la ejecución del modelo, donde se escogía aleatoriamente particiones de la base de datos (considerando siempre un 75% de los datos para entrenamiento y el otro 25% para las pruebas del modelo). Todo ello con el fin de garantizar que los resultados sean independientes tanto de los datos de entrenamiento como los de prueba. Luego a cada medida de evaluación se le calculo la media aritmética para estimar la precisión y error en las predicciones del modelo.

Tabla 6: Calculo de promedio

	Accuracy	Precisión	Recall	Error
1	0,779	0,806	0,862	0,221
2	0,807	0,798	0,937	0,193
3	0,809	0,819	0,904	0,191
4	0,798	0,801	0,921	0,202
5	0,791	0,775	0,943	0,209
6	0,804	0,792	0,934	0,196
7	0,809	0,81	0,921	0,191
8	0,795	0,794	0,935	0,205
9	0,81	0,804	0,942	0,19
10	0,813	0,815	0,915	0,187
Promedio	0,802	0,801	0,922	0,198
	80,2%	**80,1%**	**92,2%**	**19,8%**

Fuente: elaboración propia.

Podemos decir, de la tabla 6 que el modelo presenta un error promedio de 0,1948, es decir el modelo se equivoca en predecir un 19,47% del total de datos utilizados en la prueba. Además la exactitud del modelo (accuracy) da como resultado que el 80,2% de las predicciones son correctas, mientras que la precisión promedio del modelo es 0,801, es decir, el 80,1% de los casos con contrato vigente que fueron predichos están correctos. Por otro lado la tasa de verdaderos positivos (recall) del modelo es de 0,922, es decir, que el 92,2% de los datos con contrato vigente fueron identificados correctamente.

5.2. Resultados del Modelo B

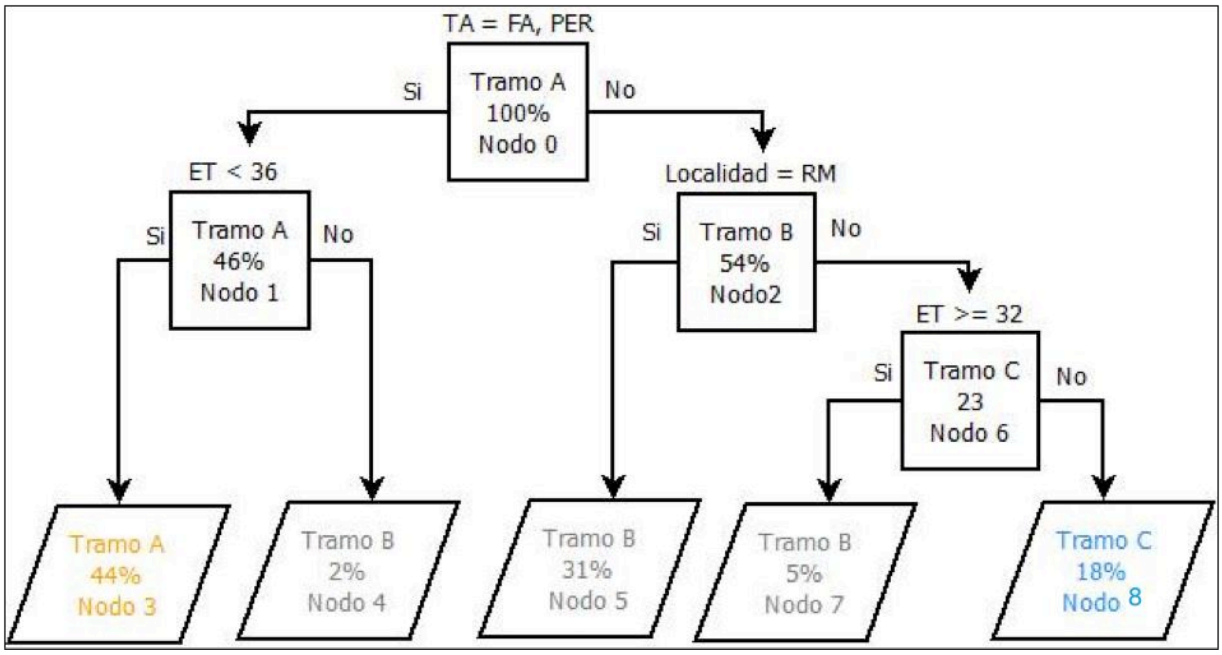
Para comprender mejor las diferentes variables y condiciones del modelo, se presenta la siguiente ilustración.

Tabla 7: Tabla de variables para el modelo B

Nombre Variable	Sigla	Descripción	Valor de la variable
Tipo de Empresa	TE	CHILE	CH
		CIMM	CI
		MIN	MI
Descripción del Ausentismo	TA	Accidente de trabajo	ADT
		Falta	FA
		Licencia medica	LI
		Permiso	PER
		Maternal	MA
Localidad	Localidad	Región Metropolitana	RM
		Zona Centro	ZC
		Zona Norte	ZN
		Zona Sur	ZS
Edad del Trabajador	ET	Edad del trabajador	[1,00[

Fuente: Elaboración propia.

Ilustración 10: Modelo B para predecir los días de ausentismo (DA)



Fuente: Elaboración propia

La interpretación de los resultados del modelo B es bastante sencilla y consistente con las condiciones reales dentro del ámbito laboral, ya que, los trabajadores que presentan un permiso o falta a su lugar de trabajo, lo hacen como máximo por dos días. Por otro lado para el caso de accidentes de trabajo o licencias médicas, los días de ausencia son mayores porque el trabajador debe cumplir con unos días de recuperación.

La ilustración 10 representa un árbol de clasificación que permite predecir los días que faltara un trabajador a su lugar de trabajo siguiendo determinadas condiciones, es decir, dependiendo del tramo A, B o C el trabajador faltara de 0 a 1 día, de 2 a 7 días y entre 8 o más días. Para la construcción de este árbol se utilizaron las variables TA (tipo de ausentismo), ET (edad del trabajador), Localidad (zona sur, zona norte y región metropolitana). Para el entrenamiento del modelo se utilizó un total de 3717 datos que corresponden al 75% de la base de datos total.

La ilustración 10 contiene 4 divisiones con diferentes condiciones que conducen a un trabajador a un tramo determinado, para predecir los días de ausencia, dichos tramos se encuentran en los nodos terminales del árbol de clasificación.

También podemos observar que en los nodos terminales se indica el tramo de días de ausencia y el porcentaje de datos que según las determinadas condiciones llegan a ese nodo. Los nodos terminales son el 3, 4, 5, 7 y 8. De estos nodos el 3 y 5 son los que representan un mayor porcentaje de datos, siendo el nodo 3 el que contiene 44% de los datos que se clasifican

dentro del tramo A (0 a 1 día de ausencia). Mientas que el nodo 5 representa el 31% de la base de datos con la clasificación tramo B (2 a 7 días de ausencia).

De las divisiones que se observan en el modelo B, podemos generar 5 reglas de decisión:

- Si $(TA = FA, PER) \wedge (ET < 36) => Tramo\ A$
  Es decir, si un trabajador presenta un tipo de ausencia (TA) igual a falta (FA) o permiso (PER) y además tiene menos de 36 años de edad, este faltara de 0 a 1 día.

- Si $(TA \neq FA, PER) \wedge (ET > 36) => Tramo\ B$
  Es decir, si un trabajador presenta un tipo de ausencia igual a accidente de trabajo o licencia médica y además tiene más de 36 años de edad, este faltara entre 2 a 7 días a su lugar de trabajo.

- Si $(TA \neq FA, PER) \wedge (Localidad = RM) => Tramo\ B$
  Es decir, si un trabajador presenta un tipo de ausencia igual a accidente de trabajo o licencia médica y su lugar de trabajo está dentro de la región metropolitana, este presentara una ausencia entre 2 a 7 días de ausencia.

- Si $(TA \neq FA, PER) \wedge (Localidad \neq RM) \wedge (ET \geq 32) => Tramo\ B$
  Es decir, si un trabajador presenta un tipo de ausencia igual a accidente de trabajo o licencia médica, su lugar de trabajo está en la zona norte o sur y además tiene mas de 32 años, este presentara entre 2 a 7 días de ausencia.

- Si $(TA \neq FA, PER) \wedge (Localidad \neq RM) \wedge (ET < 32) => Tramo\ C$
  Es decir, si un trabajador presenta un tipo de ausencia igual a accidente de trabajo o licencia médica, su lugar de trabajo está en la zona norte o sur y además tiene menos de 32 años, este presentara 8 o más días de ausencia.

A continuación se describe cada nodo generado en el árbol de clasificación, para entender cómo se distribuyen los datos, como se generan las reglas de decisión y como estas dividen a la base de datos para llegar a un nodo terminal que comprenda el mayor porcentaje de una sola variable.

- Nodo 0: este nodo representa un total de 3717 datos, que representan el 75 % de la base de datos utilizada para entrenar el modelo. A partir de este nodo se clasifican los nodos restantes. Este modelo contiene 8 nodos con un total de 4 nodos terminales.

- Nodo 1: contiene 1702 datos que cumplen con la condición del nodo 0. Donde TA (Tipo de Ausentismo) es igual a FA (Faltas) y PER (Permisos). El 91,06 % de los datos pertenecen al tramo A, el 4,64 % al tramo B y 4,28 al tramo C.

- Nodo 2: contiene 2015 datos que cumplen con la condición del nodo 0. Donde TA (Tipo de Ausentismo) es igual a FA (Faltas) y PER (Permisos). El 1,88 % de los datos pertenecen al tramo A, el 66,50 % al tramo B y 31,61 al tramo C.

- Nodo 3: contiene 1620 datos que cumplen con la condición de un trabajador de edad menor a 36 años (ET < 36). El 95,18 de los datos pertenecen al tramo A, el 0,61 al tramo B y 4,19 % al tramo C.

- Nodo 4: contiene 82 datos que cumplen con la condición de un trabajador de edad menor a 36 años (ET <36). El 9,75 % de los datos pertenecen al tramo A, el 84,14 % al tramo B y 6,09 % al tramo C.

- Nodo 5: contiene 1150 datos que cumplen con la condición de que una persona trabaje dentro de la Región Metropolitana. El 0,17 % de los datos pertenecen al tramo A, el 93,65 % al tramo B y 6,17 % al tramo C.

- Nodo 6: contiene 865 datos que cumplen con la condición de que una persona trabaje en otras zonas, diferentes a la Región Metropolitana. El 4,16 % de los datos pertenecen al tramo A, el 30,40 % al tramo B y 65,43 % al tramo C.

- Nodo 7: contiene 203 datos que cumplen con la condición de un trabajador con una edad mayor a 32 años de edad. El 0% de los datos pertenecen al nodo A. El 96,05 % al tramo B y el 3,94 % al tramo C.

- Nodo 8: contiene 662 datos que cumplen con la condición de un trabajador con una edad menor a 32 años de edad. El 5,43% de los datos pertenecen al nodo A, EL 10,27 % al tramo B y el 84,29 % al tramo C.

Con el porcentaje de cada variable que llega a un nodo determinado, se puede obtener la matriz de confusión, para determinar si el modelo está prediciendo correcta o incorrectamente.

5.2.1. Matriz de Confusión Modelo B

Tabla 8: Matriz de confusión del modelo B

		Predicción		
		Tramo A	Tramo B	Tramo C
Real	Tramo A	520	10	15
	Tramo B	0	435	22
	Tramo C	17	26	194

Fuente: Elaboración Propia

En la tabla 8 se muestra el número de trabajadores que comprenden días de ausentismo clasificados según los tramos A, B y C que el modelo predijo correcta e incorrectamente. Podemos interpretar la tabla 8 de la siguiente forma:

- El modelo predijo correctamente que los días de ausentismo para 520 trabajadores

comprendidos en el tramo A, en realidad pertenecían a ese tramo.
- El modelo predijo incorrectamente que los días de ausentismo para 17 trabajadores comprendidos en el tramo A, en realidad pertenecían al tramo C.
- El modelo predijo incorrectamente que los días de ausentismo para 10 trabajadores comprendidos en el tramo B, en realidad pertenecían al tramo A.
- El modelo predijo correctamente que los días de ausentismo para 435 trabajadores comprendidos en el tramo B, en realidad pertenecían al tramo B.
- El modelo predijo incorrectamente que los días de ausentismo para 26 trabajadores comprendidos en el tramo B, en realidad pertenecían al tramo C.
- El modelo predijo incorrectamente que los días de ausentismo para 15 trabajadores comprendidos en el tramo C, en realidad pertenecían al tramo A.
- El modelo predijo incorrectamente que los días de ausentismo para 22 trabajadores comprendidos en el tramo C, en realidad pertenecían al tramo B.
- El modelo predijo correctamente que los días de ausentismo para 194 trabajadores comprendidos en el tramo C, en realidad pertenecían al tramo C.

Podemos decir que los verdaderos positivos y verdaderos negativos superan en número a los falsos positivos y falsos negativos, cumpliendo con la regla para poder decir que el modelo tiene una predicción acertada, con una menor cantidad de datos erróneos.

5.2.2. Validación del modelo B

Para realizar la validación del modelo B, se realizaron 10 iteraciones diferentes en la creación del árbol de clasificación, el objetivo es seleccionar aleatoriamente los datos, tanto para entrenamiento del modelo (75%) como para las pruebas realizadas al modelo (25%) y así asegurar la solidez del modelo y comprobar la variabilidad de los resultados obtenidos.

Luego de realizar 10 veces el modelo con datos diferentes y obtener 10 diferentes valores de cada medida de evaluación, se obtuvo la media aritmética de cada uno. Esta media representa el valor con el cual es evaluado el modelo.

Tabla 9: Calculo de promedio

	Acurracy	**Precisión Tramo A**	**Precisión Tramo B**	**Precisión Tramo C**	**Recall Tramo A**	**Recall Tramo B**	**Recall Tramo C**	**Error**
1	0,927	0,968	0,924	0,840	0,954	0,952	0,819	0,073
2	0,929	0,949	0,938	0,862	0,975	0,950	0,791	0,071
3	0,917	0,948	0,928	0,813	0,968	0,939	0,753	0,083
4	0,921	0,981	0,882	0,815	0,959	0,954	0,753	0,079
5	0,927	0,935	0,935	0,886	0,981	0,948	0,769	0,073
6	0,926	0,981	0,898	0,855	0,957	0,968	0,766	0,074
7	0,937	0,987	0,915	0,854	0,954	0,961	0,837	0,063
8	0,914	0,971	0,888	0,837	0,944	0,956	0,760	0,086
9	0,927	0,957	0,926	0,854	0,968	0,950	0,791	0,073
10	0,913	0,952	0,922	0,802	0,961	0,938	0,758	0,087
promedio	**0,924**	**0,963**	**0,916**	**0,842**	**0,962**	**0,952**	**0,780**	**0,076**
	92,4%	96,3%	91,6%	84,2%	96,2%	95,2%	78,0%	7,6%

Fuente: elaboración propia.

De la tabla 9 se desprende que el modelo B tiene una exactitud (accuracy) del 0,924, es decir, el 92,4% de las predicciones son correctas, dejando para el error solo el 7,6% de los casos. Además se observa que en promedio el modelo predice correctamente cada tramo un 91% de las veces. Por otro lado la tasa de verdaderos positivos (Recall) en promedio de cada tramo es de 0,898, es decir, que el 89,9% de los datos de cada tramo son identificados correctamente por el modelo.

5. Sugerencias

Con el objetivo de reducir la rotación y ausentismo laboral, se pretende desarrollar una disminución sobre las siguientes reglas de decisión:

- Árbol de Clasificación Modelo A:

Si $(HM \neq t) \wedge (AT < 0{,}5) => RE$ (Termino de Contrato).
Es decir, si un trabajador no tiene un hijo menor a 18 años y sus años de trabajo no superan los 0,5 años (6 meses), entonces el trabajador va a renunciar a la empresa.

Esta regla de decisión fue seleccionada por representar el mayor porcentaje de trabajadores que terminaban contrato (12%).

- Árbol de Clasificación Modelo B:

Si $(TA \neq FA, PER) \wedge (Localidad \neq RM) \wedge (ET < 32) => Tramo\ C$

Esta regla de decisión fue seleccionada debido a que es el nodo en donde los trabajadores presentan un número mayor o igual a 8 días de ausencias. Esta regla está enfocada en los trabajadores menores a 32 años de edad, que trabajan en la zona norte o sur y que se ausentan debido a una licencia médica o a un accidente de trabajo. Además dicha regla representa el 18% de la base total de trabajadores que presentan una ausencia.

Esta sugerencia tiene como propósito mantener al trabajador durante los primero 6 meses de trabajo y crear una conciencia respecto a las licencias médicas. Para este efecto, Se pretende llevar a cabo un programa de incorporación que integre a todos los nuevos colaboradores (trabajadores).

Este programa tiene como objetivo alistar rápidamente al trabajador con su nuevo rol, para un desempeño exitoso como parte del equipo de Empresa de Servicios.

- Contempla 6 etapas.
- Se extenderá por 6 meses, es decir, desde el inicio de su contrato hasta su plazo indefinido.
- Contará con 2 cursos específicos distribuidos entre las etapas de exploración y compromiso
- Participaran 3 roles claves:
    - Nuevo colaborador.
    - Asesor (padrino).
    - Jefe directo.

Roles

Jefe directo

- Asigna un Compañero Mentor apropiado.
- Suministra el equipo necesario para el primer día del colaborador.
- Se asegura que el colaborador se sienta bienvenido.
- Se asegura que el colaborador sepa qué se espera de él.
- Amplía el conocimiento que el colaborador tiene de SGS y le ayuda a crear su red de trabajo.
- Explicar cómo el rol del nuevo colaborador encaja en todo el panorama.
- Dar y recibir retroalimentación.
- Establecer objetivos personales.

Asesor o padrino

- Responsable de la administración del programa.
- Da apoyo a los jefes inmediatos, nuevos colaboradores y compañeros mentores.
- Monitoreo e informes.

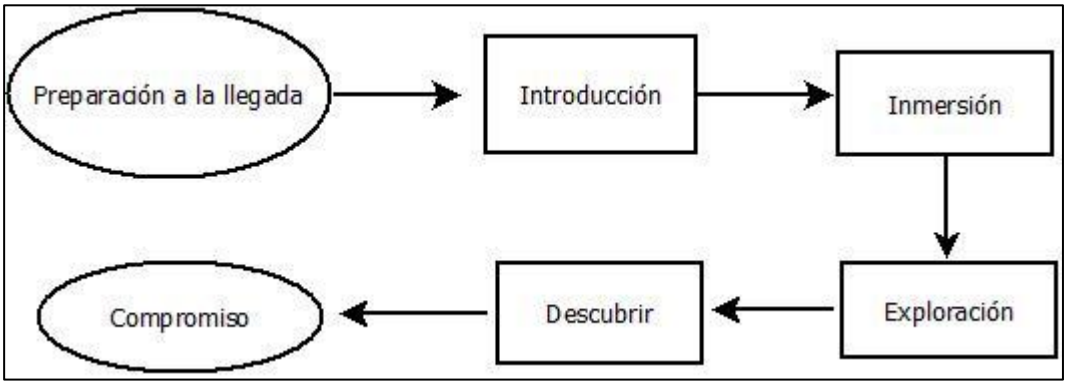
Nuevo colaborador

- Responsable de aprender acerca de SGS, del rol y expectativas de su cargo, y de conocer a sus colegas personalmente.
- Realizar todo el entrenamiento requerido.
- Completar las actividades del Programa de Incorporación.
- Dar y recibir retroalimentación.
- Buscar consejo y ayuda.

Etapas del Proceso

- Preparación Previa a la llegada.
- Primer Día: Introducción.
- Primera semana: Inmersión.
- Primer Mes: Exploración.
- Primeros Tres Meses: Descubrir.
- Primeros seis Meses: Compromiso.

Ilustración 11: flujo de proceso de Incorporación



Fuente: elaboración propia

En la ilustración 12 se puede observar cada etapa que comprende la integración del nuevo colaborador, siendo estas implementadas dentro de los 6 primeros meses de trabajo. Las etapas se clasifican en:

Etapa 1: Preparación Previa a la llegada.

Esta etapa implica el primer contacto en términos operativos del trabajador con la organización. Esto conlleva considerar un equipo y las instalaciones necesarias para que el trabajador pueda desenvolverse en su lugar de trabajo. Dentro de los accesorios podríamos encontrar (teléfono, notebook, lugar físico (escritorio), EPP (equipo de protección personal de ser necesario), marca de huella, ficha, entre otros.

Etapa 2: Introducción.

Esta etapa implica la bienvenida y presentación general del jefe directo al nuevo colaborador para enfocarlo en los objetivos y metas claves del sector en donde se desenvolverá el trabajador. Significa introducir al trabajador en el contexto de la empresa y que entienda el sentido y lo que hace la empresa para su desarrollo. Es una etapa muy importante, ya que una mala introducción deja al trabajador Confundido y sin claridad respecto de lo que se pretende. Durante esta etapa es importante el descriptor del cargo y planteamiento de problemas del área para el que se fue contratado.

Etapa 3: Inmersión

En esta etapa el trabajador deberá ya conocer más su entorno y se desenvuelve con mayor comodidad, comienza a desarrollar ciertas cercanías y se introduce en el mundo de la compañía.

Etapa 4: Exploración

En esta etapa el trabajador ya genero lazos y conoce más menos donde apunta su función e intentar captar y adquirir mayor conocimientos para la elaboración de sus tareas cotidianas.

Etapa 5: Descubrir

En esta etapa el trabajador sigue construyendo su camino en la organización, ya conoce bien su entorno y ejecuta sus funciones con normalidad e intenta tomar decisiones que añadan un valor objetivo a su participación laboral.

Etapa 6: Compromiso

En esta etapa el trabajador adquiere una identidad y compromiso con la organización. Tiene un sentido de pertenencia y le dedica tiempo y dedicación a su trabajo. Es durante esta etapa en donde el trabajador se puede entregar por completo a la compañía y aportar de forma eficiente y eficaz.

El objetivo de esta sugerencia es provocar la permanencia del trabajador (menor rotación laboral) y un compromiso por parte de él hacia la compañía. Se pretende desarrollar esta iniciativa generando dos cursos que aporten al desarrollo del trabajador dentro de la empresa (conocimiento y aprendizaje). Además se otorgara un bono por participación (5% del sueldo) a los trabajadores que asistan a estos cursos y aprueben la evaluación final de este. Por lo tanto aquellos trabajadores que obtengan una nota igual o superior a 5.0 en ambos curso, obtendrán dicho bono.

Además se pretende desarrollar una cultura organizacional, respecto a los días de ausencia generados por las licencias médicas y accidentes de trabajo.

Para la primera causal de ausentismo (licencias médicas) se pretende generar la contratación de doctores de especialidad general que ocupen un lugar dentro del sector de OI (Operación Integral) con el objetivo de controlar de forma inmediata los motivos y causantes de enfermedades o problemas de salud de trabajadores en particular. Otorgando de forma transparente y adecuada el uso de licencias médicas al interior de la compañía. De esta manera se controlara tanto el correcto uso de licencias médicas y por lo tanto disminuirán los días de esta causal. Ya que utilizaran correctamente estos métodos. También seria de vital importancia utilizar el vínculo de SGS con el IST, para generar presentaciones mensuales respecto al correcto uso de licencias médicas, además de la fiscalización de estas por medio de un personal que controle el día a día en que estas se generan. Con el objetivo de disminuir este número y entregarle a la compañía una mayor confianza respecto al uso de las licencias médicas.

Por otro lado se recomienda instaurar políticas de prevención y control más eficientes y eficaces para el área de OI (Operación Integral) que permita desarrollar publicidad comunicacional hacían el correcto uso de EPP, además de los usos de los recursos presentes dentro de la organización (maquinarias, instrumentos, etc.). También se pretende instaurar cursos de uso de extintores y zonas de alertas y prevención en caso de incendios o fenómenos naturales ocasionados durante las jornadas laborales.

7. Evaluación Económica

Para la propuesta del Modelo A: "Modelo de Clasificación para predecir el Motivo de Retiro", el objetivo es reducir el porcentaje de rotación en la organización, que corresponde a un 32,31% para el 2014. Para este efecto nos enfocaremos principalmente en la reducción del nodo mayor, que corresponde a un 12 %, que se produce específicamente cuando un trabajador de la empresa no posee hijo menor a 18 años y tiene menos de 6 meses de antigüedad en la empresa. Se pretende reducir este porcentaje a un 0% (Reducción del 12%) con la finalidad de disminuir las salidas y términos de contratos de los trabajadores.

En este caso el total de datos utilizados para construir el modelo fueron 4455 trabajadores. De estos se utilizaron el 75%, siendo 3.341 datos para calcular los costos asociados a la situación actual v/s situación propuesta.

Tabla 10: Situación actual vs Situación propuesta.

	Situación Actual	Situación Propuesta
% Total de Termino de Contrato	1 + 5 + 3 + 2 + 12 = 23 %	1 + 5 + 3 + 2 + 0 = 11 %

En la tabla 10 se muestra la suma de porcentajes correspondientes a los trabajadores que terminan su Contrato de trabajo, cada uno de los factores considerados en la suma corresponden a los porcentajes de cada nodo en donde el árbol arrojo como resultado Termino de Contrato. Con la Sugerencia el nodo mayor de un 12 % se reducirá a un 0%. Por lo tanto el porcentaje de trabajadores que terminaran su contrato será de 11%, respecto al 23% de la situación actual

Por otro lado, los costos involucrados en el proceso de rotación laboral (proceso efectuado entre la salida de un trabajador y el ingreso de otro a la organización) son los que se observan en los cálculos siguientes.

Tabla 11: Remuneraciones

Remuneración Personal de Selección	Costo
Psicólogo (2 cargos)	$1.268.385
Encargado de Selección	$1.028.522
Encargado de Capacitación y Desarrollo	$1.071.152
Total	$3.368.059
Costo Remuneraciones Anual	$40.416.708,00

En la tabla 11 se describen las remuneraciones del personal encargado de la selección y capacitación del trabajador que ingresa a la empresa.

Tabla 12: Costos Asociados al Trabajador (Situación Actual)

	costo unitario	costos total	costo mensual
Costos Curso Inducción	$ 70.000	$ 53.794.125	$ 4.482.843,75
Costos exámenes pre-ocupacionales (IST)	$ 2.000	$ 1.536.975	$ 128.081,25
Costos publicidad Vacantes		$ 24.000.000,00	$ 2.000.000,00
Costo Total	$ 72.000	$ 79.331.100	$ 6.610.925

De la tabla 12 se desprende los costos unitarios que genera un trabajador a la empresa, al momento de su ingreso, relacionados a la documentación, exámenes e inducciones propuestas por el reglamento interno de la empresa.

$$C.Total\ Reclutamiento\ anual\ = Costo\ Remuneraciones + Costos\ por\ trabajador$$

Costo Total de Reclutamiento anual situación actual	$119.747.808

Este costo total se calcula considerando la situación actual de la empresa (23 % trabajadores con Término de Contrato).

Por otro lado la propuesta del modelo B: "Modelo de clasificación para predecir el número de días de ausencia que presentara un trabajador", tiene como finalidad reducir el porcentaje presente en el nodo 8. A este nodo llegan el 18% de los trabajadores que presentan una ausencia debido a una licencia médica o accidente de trabajo, además de trabajar en la zona norte o sur y tener menos de 32 años de edad. Todos los trabajadores presentes en este nodo faltaran a su trabajo 8 o más días. Con la sugerencia propuesta se pretende reducir a un 8% el número de trabajadores que falten 8 o más días debido a una licencia médica o accidente de trabajo.

Para generar el cálculo de las ausencias de los trabajadores, se utilizó un promedio de las remuneraciones (sueldo base) de cada trabajador, sin considerar a los gerentes y se dividió por 30 días. Además se utilizó un total de 1024 trabajadores que presentaban una ausencia. También se consideró solo el nodo 8 del modelo B, ya que, son los casos en donde los trabajadores presentan una licencia médica o accidente de trabajo y faltan 8 o más días al lugar de trabajo.

El sueldo base promedio de la empresa es de $578.002. Dicho valor se usara como referencia para calcular el valor que se le paga a un trabajador cuando se ausenta debido a una licencia médica o un accidente de trabajo, como se observa en la tabla 13.

Tabla 13: Costo de un día de ausencia

Costo de Ausencia (1 día)	Días	Costo total ausencia
$19.267	8	$154.134

En la tabla 13 se observa que la empresa debe cubrir un costo promedio por día de ausencia de $19.267, es decir, la empresa debe pagar $154.134 a un trabajador que se ausente a su lugar de trabajo 8 días, presentando la licencia médica correspondiente.

De la predicción el modelo arrogo que al nodo 8 llegan 18% de los datos, esto implica que los trabajadores faltaran 8 días o más. Por lo tanto la situación actual de la empresa indica que 139 trabajadores son clasificados en el nodo 8. Para calcular el costo total de las ausencias generadas por dichos trabajadores es de $21.424.626.

Tabla 14: Costo de la situación actual

Costo Total de Reclutamiento anual situación actual	$119.747.808
Costo total anual para el ausentismo de la situación actual	$21.424.626
Costos total situación actual	**$141.172.434**

De la tabla 14 se desprende que la suma de los costos asociados a la rotación y ausentismo laboral es de $141.172.434, siendo este el costo total de la situación actual.

Calculo de los costos para la situación mejorada

Con la sugerencia desarrollada, la cual tiene como objetivo principal reducir la cantidad de trabajadores que se van de la empresa a un 11 % (reducciones del 12 %). La disminución en términos de costos considerando las mismas remuneraciones, pero diferentes costos asociados al trabajador seria:

Tabla 15: Costos Asociados al Trabajador (Situación Propuesta)

	costo unitario	costos total	costo mensual
Costos Curso Inducción	$70.000	$25.727.625	$2.143.968,75
Costos cursos propuesta	$40.000	$30.720.000	$2.560.000,00
Bono Participación (5%)	$28.900	$22.195.286	$1.849.607,16
Costos exámenes pre-ocupacionales (IST)	$2.000	$735.075	$61.256,25
Costos publicidad Vacantes		$24.000.000,00	$2.000.000,00
Costo Total	$72.000	$103.377.986	$4.205.225

Por lo tanto el costo de la situación propuesta seria según la tabla 15 de $103.377.986, considerando solo al 11% de los trabajadores que terminan contrato.

Además con la sugerencia se pretende disminuir el 18% de los trabajadores que están faltando 8 o más días, es decir reducir el nodo 8 de 18% a 8% del modelo B. Por lo tanto los costos de la situación propuesta son considerando que solo el 8% de los trabajadores se ausentaran 8 o más días es:

Costos ausencia situación propuesta (8%)	$9.556.308

Para calcular los costos totales de la situación propuesta, se suman los costos de rotación con la propuesta y costos de ausencia con la propuesta.

Tabla 16: Costo total de la situación propuesta

Costo reclutamiento (situación propuesta)	$143.794.693,95
Costo ausencia (situación propuesta)	$9.446.308
Costo total situación propuesta	**$153.351.001**

Podemos observar que en la tabla 16 se presenta el costo total que se generaría en la empresa si esta implementa las sugerencias.

Una vez calculado los costos de antes y después de la propuesta, se puede calcular el ahorro que se generaría si la empresa implementa la propuesta y reduce los porcentajes de ausentismo y rotación laboral que presentan sus trabajadores. El ahorro es:

Costos total situación actual	$331.042.306
Costo total situación propuesta	$153.251.001
Ahorro	**$177.791.305**

En conclusión, si se considera la sugerencia y se dan las condiciones de las situaciones propuestas para los problemas de ausentismo y rotación laboral, la empresa se estaría generando un ahorro de $177.791.305 solo con el hecho de enseñar, crear conciencia sobre el trabajador y controlar las licencias médicas que este genera.

8. Conclusión y Recomendaciones

Analizando los resultados obtenidos en el modelo A, podemos dar a conocer las condiciones necesarias para que un trabajador se mantenga o se retire de la empresa de servicios. Dichas características permiten generar planes de acción tendientes a desarrollar una retención del personal dentro de la compañía. Reduciendo de esta manera el 12 % de la rotación laboral.

Debido a esto se puede concluir que uno de los factores principales para reducir la rotación laboral dentro de la empresa, es que las personas que comienzan a trabajar o llevan menos de 6 meses en la empresa deben ser constantemente motivadas e informadas durante este tiempo. Es necesario realizar cursos durante los primeros 6 meses de trabajo con el objetivo de preparar al trabajador para su comienzo, mantener al trabajador en conocimiento de su entorno laboral y por ultimo generar un compromiso de este hacia su la organización.

Del modelo de clasificación B para predecir los días de ausentismo se obtienen 5 nodos terminarles, de los cuales uno de ellos representa el tramo A (0 a 1 día de ausencia), tres comprenden el tramo B (1 a 8 días de ausencia) y el restante al tramo C (8 o más días de ausencia).

En el nodo 2 se obtuvo el tramo A, donde los trabajadores faltaran entre 0 y 1 día y que corresponden a los trabajadores que presentan una edad menor a 36 años y presentan una falta o permiso para justificar su ausencia. Por lo tanto este grupo de trabajadores se encuentran dentro de las políticas de la empresa, en donde se indica que una persona no puede superar los 2 días de ausencia, de lo contrario es inmediatamente desvinculado de su lugar de trabajo.

En los nodos 3 , 5 y 7 se encuentran los trabajadores que comprenden el tramo B, presentando de 1 a 8 días de ausencia respectivamente, razón que se produce específicamente para aquellas personas menores a 32 años que trabajan en la zona norte o sur. Y además presentan una licencia médica o accidente de trabajo para justificar su ausencia. Por lo tanto resulta fundamental instaurar una política de cumplimento y conciencia en el uso de las licencias médicas, además de incluir en el curso de inducción un programa de capacitación que presente las amonestaciones y explicaciones detallada de cada causal de ausentismo. También desarrollar visitas de inspección por parte los seguros (Metlife y Bice Vida) para controlar y obtener un registro periódico de la presión y condiciones básicas del trabajador con respecto a la salud.

Si se aplican las sugerencias propuestas en esta investigación, la empresa se estaría ahorrando $177.791.305, solo en el concepto de reducción de la rotación y ausentismo laboral. Lo que guiara a la empresa a desarrollar una mayor conciencia y valorización del capital humano (al retener al talento). Permitiendo además generar un ambiente laboral más integral y participativo.

Bibliografía

[Méndez, Leonett., 2005] Daniel Leonett, Oscar Méndez. (2005). Análisis de los factores que generan ausentismo laboral en el personal de enfermería del centro médico docente. Maturín: Universidad de Oriente Escuela de Ciencias Sociales y Administrativas.

Francisca Meza. (2012). Estudio revela que ausentismo laboral en la región es mayor en la banca y sector productivo. Junio, de Inmune Sitio web: http://www.inmune.cl/author/root/page/4/.

[Guihard T., 2012] Thierry Guihard. (2012). América Economía. Como disminuir el ausentismo laboral. Recuperado. Mayo del 2015, de América Economía Sitio web: www.americaeconomia.com/analisis-opinion/como-disminuir-el-ausentismo-laboral

[De Luca M., 2006] Mauricio De Luca. (2006). Plan para enfocar las campañas bancarias utilizando Datamining. Santiago de Chile: Universidad de Chile, Facultad de ciencias físicas y matemáticas.

[Moreno M. et al., 2006] Moreno M. (2006). Aplicación de técnica de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. España: Universidad de Salamanca. Departamento de Informática y Automática.

[Contreras Ferreira 2012] Evelyn Contreras, Francisca Ferreira. (2014). diseño de un modelo predictivo de abandono de clientes para una empresa de telecomunicaciones utilizando arboles de decisión. Santiago de Chile: Universidad de Valparaíso.

[ALTONIVEL, 2014] Altonivel. (2014). Existen problemáticas en las empresas donde la única opción es poner sobre la mesa toda alternativa posible para solucionarlo. El árbol te ayuda a lograrlo.. Junio del 2015, de Revista electrónica Altonivel Sitio web: http://www.altonivel.com.mx/54514-3-valores-para-darle-sentido-a-tu-vida-profesional.html

[Valenzuela O., 2015] Oscar Valenzuela. (2015). Chilenos faltan al trabajo 16 días al año. Abril de 2015, de Inmune Chile Sitio web: http://www.inmune.cl/lun-pillados-chilenos-faltan-al-trabajo-16-dias-al-ano/

[OCDE, 2011] J. Poblete. (2011). Funcionarios chilenos lideran ranking de la OCDE en horas de trabajo y ausentismo laboral. Junio de 2015, de La Tercera Sitio web: http://diario.latercera.com/2011/06/25/01/contenido/pais/31-74061-9-funcionarios-chilenos-lideran-ranking-de-la-ocde-en-horas-de-trabajo-y.shtml

[Chapman et al. 2000] Pete Chapman et al. (2000). Crisp-DM 1.0 "Step-By-Step data minning guide. USA: SPSS.

[Goicochea A., 2009] Aníbal Goicochea. (2009). CRISP-DM, Una metodología para proyectos de Minería de Datos. Mayo 2015, de Blog Aníbal Goicochea Sitio web: http://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-minería-de-datos/

[Hernández et al., 2004] Hernández et al. (2004). Introducción a la Minería de Datos. Madrid: Pearson

[Calle F., 2014] Fernando Calle Alonso. (2015). Técnicas Bayesianas de apoyo a la toma de decisiones y sus aplicaciones. España: Universidad de Extremadura.

[Domínguez I., 2015] Ignacio Domínguez. (2015). CREDIT SCORING. Agosto 2015, de Expansión Sitio web: http://www.expansion.com/diccionario-economico/credit-scoring.html

Luis Pinto. (2015). Modelo algorítmico para la clasificación de documentos de carácter judicial en lenguaje portugués según su contenido. Perú: Pontificia Universidad Católica del Perú.

Pablo Tapia. (2014). Diseño e implementación de un sistema para la clasificación de tweets según su polaridad. Mayo 2015: Universidad de Chile.

[Sandoval F., 2014] Felipe Sandoval. (2014). La importancia del sector servicios en el comercio internacional. Mayo del 2015, de DIRECON Ministerio de Relaciones Exteriores Sitio web: http://www.direcon.gob.cl/2014/12/la-importancia-del-sector-servicios-en-el-comercio-internacional/

Anexos

Nombre variable	Sigla	Descripción	Valor que puede tomar la variable
Tipo de Empresa	TDE	CHILE	CH
		CIMM	CI
		MIN	MI
Número de licencias presentadas	NLP	Es la cantidad de licencias que ha presentado un trabajador desde enero del 2013 hasta diciembre del 2014	1,2,3,….
Edad de Trabajador	Edad	Nº que indica la cantidad de años de un trabajado	1,2,3,….
Cargo	Cargo	Abogado	A
		Administrativo	B
		Analista	C
		Asesor	D
		Asistente	E
		Auditor	F
		Auxiliar	G
		Ayudante	H
		Cajero	I
		Capataz	J
		Chofer	K
		Coordinador	M
		Ejecutivo	N
		Electromecánico	O
		Encargado	P
		Espectroscopista	Q
		Estadístico	R
		Fundidor	S
		Geólogo	T
		Gerente	U
		Ingeniero	V
		Inspector	W
		Jefe	X
		Maestro	Y
		Mecánico	Z
		Muestreo	AB
		Operador	AC
		Planificador	AD

		Preparador de Muestra	AE
		Prevencionista	AF
		Químico	AG
		Recuperador	AH
		Refinador	AI
		Soldador	AJ
		Sub Gerente	AK
		Supervisor	AM
		Técnico	AN
Localidad	localidad	Zona Norte	ZN
		Zona Sur	ZS
		Zona Centro	ZC
		Regio Metropolitana	RM
División	División	Administración	AQ
		Agriculture	AR
		Analítica	AS
		As	AT
		Automotriz	AU
		Comercial	AV
		CTS	CTS
		División Club Deportivo	AX
		Environmental	AY
		Finanzas	AZ
		Geo-metalurgia	BA
		GIS	GIS
		Inspectores	BC
		IT	IT
		Legal	BE
		Management	BF
		Metalurgia	BG
		Mineralogia	BH
		OI	OI
		Oil, gas & chemical	OGC
		Outsourcing	BK
		Proyectos	BI
		RR.HH	RH
		S&SC	SSC
		Servicios Estratégicos	BO
		Sin Clasificación	BP
Sexo	Sexo	Masculino	M

		Femenino	F
Causal de Ausentismo	CA	Accidente	AC
		Enfermedad	EN
		Enfermedad Hijo	EH
		Falta	FA
		Licencia Maternal	LMA
		Licencia Medica	LME
		Permiso	PE
Días de Ausencia	DA	N° de veces que puede faltar el trabajador a su lugar de trabajo desde enero del 2013 hasta diciembre del 2014	
Mes Presentación Licencia	MPL	Meses del año	Enero = 1 , Febrero= 2, Marzo = 3, Abril= 4, Mayo = 5, Junio = 6, Julio = 7, Agosto = 8 , Septiembre= 9, Octubre= 10, Noviembre = 11 , Diciembre= 12
Estado Civil	EC	Casado	CA
		Soltero	SO
		Divorciado	DI
		Separado	SE
		Viudo	VI
Nivel de Estudios	NE	Educación Media	EM
		Educación Superior	ES
		Otros	Otros
Hijo Menor	HM	True	T
		False	F
Numero de Parientes	NP	Cantidad de personas que viven con el trabajador en su vivienda	1,2,3,4,5
Años de Trabajo	AT	Cantidad de años que ha trabajado un empleado en la empresa	1,2,3,…
Cónyuge	Cónyuge	True	T
		False	F
Rut	ID	Identificador único del trabajador	Rut
Tipo de Ausentismo	TA	Accidente de Trabajo	ADT
		Falta	FA
		Licencia	LI
		Licencia Maternal	MA
		Permiso	PER