UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DETECCIÓN DE FUGA DE OPERARIOS EN UNA EMPRESA DEL SECTOR MINERO UTILIZANDO MINERÍA DE DATOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

ROBERTO IGNACIO JARA DÍAZ

PROFESOR GUÍA
LUIS ABURTO LAFOURCADE

MIEMBROS DE LA COMISIÓN
RICHARD WEBER HAAS
SEBASTIÁN CONDE DONOSO

SANTIAGO DE CHILE
2015

RESUMEN DE LA MEMORIA PARA OPTAR
TITULO DE INGENIERO CIVIL INDUSTRIAL
POR: Roberto Ignacio Jara Díaz
FECHA: 06/10/2015
PROF. GUÍA: Luis Aburto Lafourcade

DETECCIÓN DE FUGA DE OPERARIOS EN UNA EMPRESA DEL SECTOR MINERO UTILIZANDO MINERÍA DE DATOS

El sector minero en Chile actualmente vive un periodo complejo que se debe en gran parte a la tendencia a la baja del precio del cobre y el aumento de los costos de explotación, por lo que se hace necesario innovar en los procesos para mantenerse competitivo en la industria. Uno de los desafíos principales se encuentra en la gestión de Recursos Humanos. En particular, Chile presentará un creciente déficit de operarios especialistas en el rubro durante la próxima década, por lo que se hace relevante conocer bien su comportamiento para evitar los desajustes en la planificación de explotación de un yacimiento. Este estudio afronta el problema desde la minería de datos, aplicados a una faena de la mediana minería chilena.

Siguiendo la metodología KDD se analizaron datos históricos de los operarios de la empresa correspondientes a datos demográficos, capacitaciones y licencias, además de fuentes gratuitas de datos externos como el valor del cobre, para poder conocer de mejor manera cómo ha sido el comportamiento de renuncia de los operarios durante los años de funcionamiento de esta faena minera.

Para el desarrollo del proyecto, se utilizaron tanto modelos paramétricos como no paramétricos sobre una variable objetivo de tres clases: Permanencia en la empresa, Renuncia Voluntaria y Despido. Si bien, el objetivo del proyecto no es predecir los despidos dentro de la compañía, se hace necesario distinguir este comportamiento del resto para tener una mejor comprensión del problema global: La rotación del personal. El modelo seleccionado para este caso fue Incremento de Gradiente (Gradient Boosting), el cual basa su funcionamiento en modelos "débiles" para construir secuencialmente un modelo que minimice una función de pérdida. Con esto, se obtuvo una precisión general del modelo de un 86% sobre datos de prueba.

Las principales conclusiones de este estudio apuntan hacia la actual gestión de capacitaciones de los empleados de la faena. Se observa que aquellas personas que presentan mejor índice de aprobación y asistencia a las capacitaciones financiadas por la empresa poseen una probabilidad mayor a renunciar a su cargo. Según estudios de psicología laboral, casos como estos se pueden encontrar en empresas que son vistas como "semilleros" de empleados donde los empleados saben que pueden encontrar un buen ambiente de aprendizaje y perfeccionamiento, pero que al mismo tiempo, no permiten mucha movilidad dentro de la organización. En consecuencia, estas personas que poseen ahora una mejor empleabilidad debido a los conocimientos y experiencia adquirida, comenzarán a buscar un nuevo empleo donde se les ofrezcan estas oportunidades de crecimiento.

Agradecimientos

En primer lugar agradecer a mis padres, Verónica y Cristián, por su constante apoyo, compañía, consejo y cariño. Gracias por todo, ya que sin ustedes nada de esto sería posible. A mi hermana Bárbara, porque desde el día que me enseñó a leer ha tratado de ayudarme en todo lo que está a su alcance. Gran parte de esta memoria fue escrita en el hogar que formaron con Nicolás. A Anita, por acompañarme durante prácticamente toda mi carrera, celebrando los éxitos y apoyándome en los fracasos. Ustedes han hecho de éste un camino mucho más fácil de lo que podría haber sido.

A los profesores que conforman la comisión por su constante preocupación. Sin el Prof. Richard, esta memoria no existiría. Gracias Prof. Sebastián por la gran disposición a siempre contribuir con este proyecto. Gracias Prof. Luis por las recomendaciones que semana a semana realizaba a este trabajo.

A mis amigos de la Universidad, por todas las risas y anécdotas que a partir de hoy son parte del lindo recuerdo de esta etapa que finaliza. Gracias por el agxante. También agradecer a mis amigos, esos que conocí cuando apenas estábamos aprendiendo a sumar, y que hasta el día de hoy siguen a mi lado. Ustedes saben quiénes son.

A mi tío Fernando, por recibirme en Antofagasta y ayudarme cuando más lo necesité en mis primeras semanas de estadía.

Finalmente, agradecer a Nibaldo y al personal de la compañía por facilitar los datos, interesarse en este proyecto y permitir su realización.

Tabla de Contenido

Índice de Tablas

Índice de Ecuaciones

Índice de Ilustraciones

1. Introducción

1.1. Contexto nacional de la industria

Es de conocimiento general que la industria minera es el motor económico de Chile, siendo el mayor productor de Cobre, Litio y Yodo del mundo. La industria minera representó el 11,2% del PIB nacional durante 2014 [1], manteniéndose sistemáticamente durante la historia del país dentro de los primeros lugares, siendo popularmente conocido como el *sueldo de Chile*. Junto a lo anterior, Chile es además el país que mayor cantidad de reservas de cobre posee en el mundo [2].

El año 2011 el precio del cobre experimentó su mayor valor (nominal y real) en toda su historia. Ese año se esperaba que la producción de cobre aumentaría incluso en un 50% entre 2011 y 2018 [3]. Sin embargo, luego de ese año el precio de este metal ha tendido a la baja, alejándose del auge que vivió ese año. Esto, sumado al aumento en los costos de explotación [4], la inestabilidad de los principales mercados compradores de cobre y los cambios regulatorios a nivel nacional en la industria, ha propuesto a la minería nuevos desafíos para mantenerse competitiva y sustentable durante los próximos años. Uno de ellos atañe directamente a la gestión de sus recursos humanos.

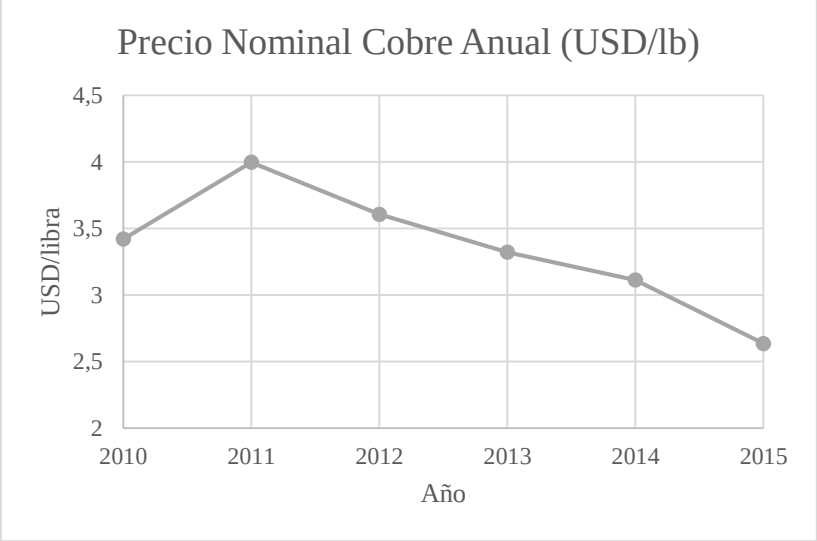


Ilustración 1 - Valor del cobre anual [5]

Para cuantificar el fenómeno de fuga en las empresas, se utiliza el concepto de tasa de fuga, el cual se encuentra en la ecuación 1 a continuación.

$$Tasa\ de\ fuga = \frac{EF}{(EPP+EFP)/2} * 100$$

Ecuación 1 - Tasa de Fuga

Donde:

- EF: Número de empleados fugados durante el año.
- EPP: Número de empleados al principio del año.
- EFP: Número de empleados al final del año.

El sector minero ha enfrentado durante los últimos años una alta rotación de su personal, declarando que también ha impactado directamente en sus costos [6]. Durante los años 2011 y 2012, las tasas de rotación para los trabajadores contratistas de la minería alcanzaron el 30%, llegando incluso al 80% en algunos casos. Esta cifra es bastante superior al 6,1% que promedió la gran minería para sus trabajadores internos en 2011 [7]. En la empresa donde se realiza el estudio, perteneciente a la mediana minería, el porcentaje de rotación para el año 2011, considerando solo empleados internos, fue de 32% superando en valor relativo a sus pares de la gran minería, asemejándose al caso de las empresas que prestan servicios a la minería.

Poseer una tasa de fuga alta con respecto a la competencia, en una industria muy especializada como lo es la minería, produce la adquisición de operarios por parte de otras operaciones mineras, otorgando recursos a la competencia la cual será bastante reñida durante los próximos años. Según el estudio "Fuerza Laboral de la Gran Minería Chilena 2013-2022" [8], el mercado laboral minero chileno tendrá una brecha entre oferta y demanda de 11.000 personas. Del mismo estudio se extrae que la demanda por operarios y mantenedores de equipos mineros tendrá un aumento progresivo durante esta década, como se observa en el siguiente gráfico.

Ilustración 2 - Demanda por Operarios y Mantenedores [9]

Es importante señalar que la demanda proyectada por el estudio contempla además los cargos relacionados con Geología, Ingeniería y Supervisores. Sin embargo, para cada año proyectado por el estudio, los cargos de Operario y Mantenedor ocupan en promedio aproximadamente el 80% de la futura demanda por personal.

Además, dada la edad promedio de los actuales trabajadores chilenos, el número de éstos que estarán en edad de retiro en los próximos años irá en crecimiento, tal como se muestra en la siguiente ilustración.

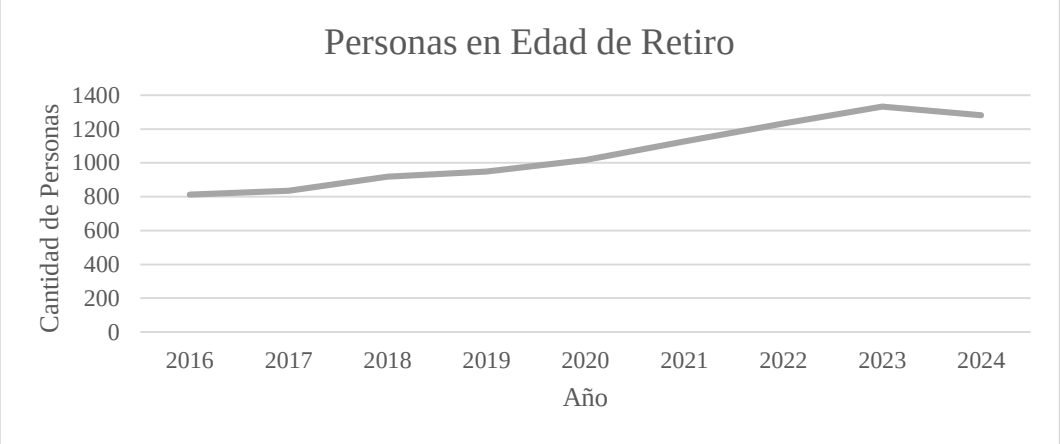


Ilustración 3- Personal en edad de retiro [9]

En el mismo estudio, se menciona además que la formación actual de los trabajadores de la minería, sobre todo a nivel técnico no cubre la demanda que se requerirá en el futuro. Si bien, existe conocimiento y se han abierto nuevas carreras en el área de minería y metalurgia, el mantenimiento y la operación siguen siendo un punto débil.

Por esta razón, es que las empresas se preocupan constantemente de capacitar a su personal para desarrollar las actividades específicas de sus cargos. La inversión realizada por las empresas en términos de capacitación viene a suplir entonces una deficiencia de la oferta por carreras técnicas en el país para la industria minera, sin embargo, también representa un riesgo para la empresa, ya que al capacitar a su personal, éste puede emigrar a la competencia, entregándoles personal capacitado sin que esta última incurra en gastos más que los asociados a la captación de personal.

1.2. Rotación de empleados

Actualmente uno de los activos más importantes que tienen las empresas de cualquier rubro son los trabajadores que la conforman. En ellas reside el conocimiento del funcionamiento de la organización, y en particular, de cómo ejecutar de forma apropiada su cargo debido a la experiencia adquirida en su puesto de trabajo. En conjunto, estas personas son las que marcan la diferencia entre grupos de alto y bajo rendimiento. Por lo tanto, es una preocupación para las organizaciones mantener su capital humano dentro de la organización, puesto que a la larga es más económico retener a su personal que reemplazarlo constantemente [10].

El estudio de la rotación de capital humano para el sector productivo ha sido una constante preocupación tanto para la industria como para la academia, desde el punto de vista del impacto económico que genera como también desde el punto de vista de la psicología laboral.

Existen diversos tipos de costos asociados a la fuga de un empleado, los cuales se pueden agrupar [11] como se observa en la tabla 1.

Costo	Descripción
Costos Pre Fuga	
Baja de productividad	Empleado insatisfecho con su trabajo rendirá menos que en condiciones normales
Absentismo	Número de días ausente previo a la fuga
Influencia en compañeros	Baja productividad individual disminuye la productividad del grupo
Inversión en empleado	Total invertido en el empleado durante su estadía en la empresa, por ejemplo, en capacitaciones
Costos de Vacante	
Costo de RR.HH.	Tiempo requerido para realizar el finiquito del trabajador
Indemnización	Monto a pagar al trabajador por años de servicio, en caso de despido
Contacto con nuevo empleado	Tiempo requerido para encontrar una persona que reemplace el cargo, afectando la productividad del área
Costo de Reemplazo	
Avisaje	Costo asociado a la captación mediante avisos de un nuevo puesto de trabajo
Evaluación	Costos asociados a la evaluación para el cargo al que se postula, tanto como el test sicológico como test médicos
Adaptación	Pérdida de eficiencia dado el período de adaptación del nuevo empleado a la actividad que anteriormente se desempeñaba en la empresa
Entrenamiento	Costo asociado a la capacitación del nuevo empleado para desempeñar su nuevo cargo

Tabla 1 - Costos asociados a la fuga

Los costos pre fuga están asociados a todos los costos que tiene la empresa al mantener a un empleado que piensa en renunciar voluntariamente. El proceso de renuncia voluntaria para un empleado comienza mientras este se encuentra trabajando para su actual empleador, por lo tanto, influye en el rendimiento de la compañía dado los factores mencionados: baja productividad, desmotivación, absentismo. El proceso, según describe Mobley [12] consta de las siguientes etapas:

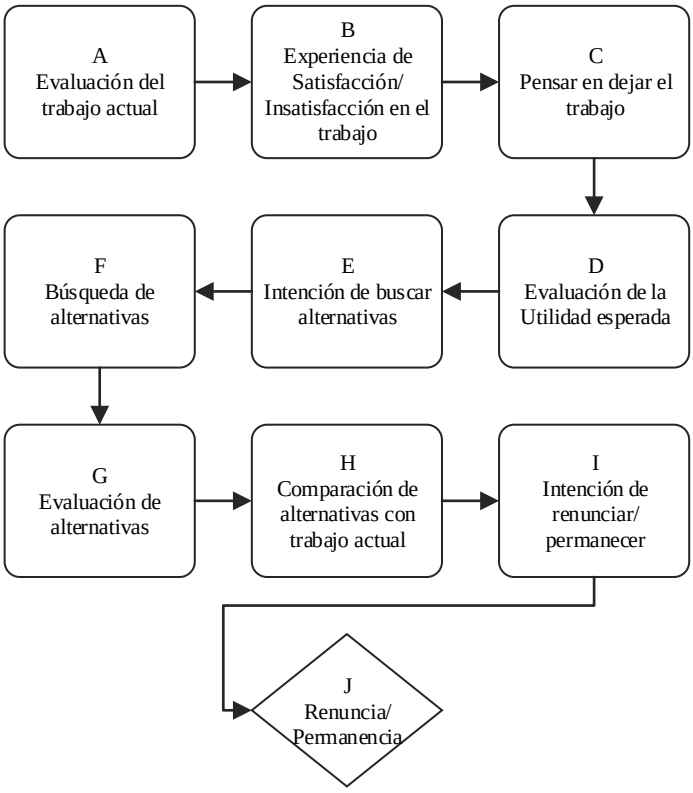


Ilustración 4 - Proceso de Renuncia Voluntaria

Como se puede apreciar, este proceso no es trivial para la persona, puesto que se debe tomar el tiempo para estudiar todas sus posibles opciones fuera del trabajo que ya posee, el cual le otorga un ingreso (en este caso, de los operarios mineros) seguro. Por esta razón es que la decisión es tomada con bastante cautela puesto que, de lo contrario, la persona cambiará su ingreso seguro por algo incierto.

En segundo lugar, los costos de vacante son todos los relacionados con mantener el puesto de trabajo vacío mientras se busca un reemplazo para la persona que se retiró de la empresa. Estos costos además se transmiten hacia sus compañeros que permanecen en la empresa, debido a que ellos serán quienes remplazarán momentáneamente a la persona que renunció, bajando la productividad en las tareas que ellos deberían desempeñar. Incluso, de ser una alta tasa de renuncias o despidos, esto puede influir en la reputación de la empresa, lo que afecta directamente sus ventas y/o tasación bursátil [13].

Finalmente, los costos de reemplazo se relacionan con la búsqueda de un nuevo empleado y su adaptación al cargo y a la organización. Estos se relacionan con los costos de captar a una nueva persona desde el mercado laboral que tenga las competencias necesarias para desarrollar el cargo que se busca. Esto último incluye los costos de reclutamiento y selección, además del avisaje en el mercado laboral de la nueva vacante. Además, cuando se encuentra una nueva persona para el cargo, se debe considerar que la adaptación al cargo no es instantánea. En este sentido, en la literatura se define la "curva de aprendizaje" [14] de un nuevo empleado, que es la relación del Aprendizaje en la nueva tarea (que se relaciona directamente con la productividad al realizar esta tarea) y el tiempo que se ha empleado para aprender. En este sentido, este modelo indica la cantidad de tiempo que tomará a una persona sin experiencia en la tarea a aprender a realizar cierta tarea de manera que cumpla un cierto estándar, llamado "Estándar del trabajador experimentado". En inglés, se utiliza la sigla EWS (Experienced worker's standard). A continuación, se presenta un esquema de una curva de aprendizaje típica, la cual posee una línea horizontal que indica el estándar del trabajador experimentado.

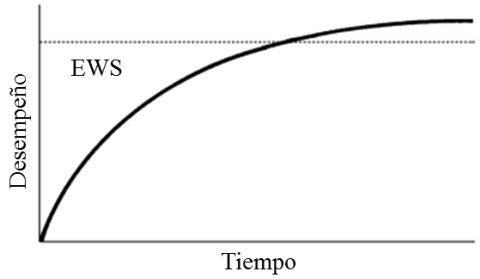


Ilustración 5 - Curva de Aprendizaje Típica

Si bien, los costos mencionados no son directos y fáciles de estimar, se puede apreciar que no son despreciables. Por lo tanto, es importante para los departamentos de RR.HH. tomar medidas para la retención de su personal, especialmente de quienes sean apreciados por la empresa en la labor que cumplen.

Existen diversos factores por los cuales una persona puede tomar la determinación de dejar voluntariamente una empresa. En la literatura se puede encontrar diversos estudios que tratan de explicar este fenómeno [11], pudiendo agrupar las variables en tres grandes grupos: variables demográficas, variables del trabajo y variables psicométricas.

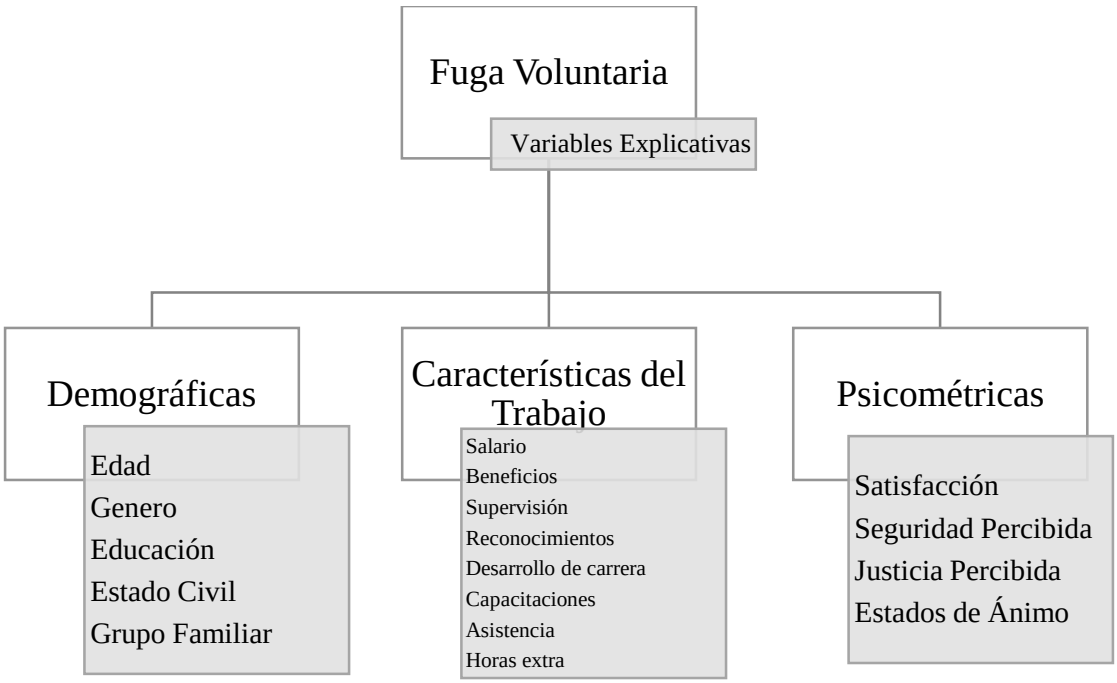


Ilustración 6- Variables explicativas

Los primeros dos grupos pueden ser analizados en base a información histórica de la empresa. Sin embargo, el último grupo es bastante difícil de medir, ya que para esto se deben aplicar encuestas al personal acerca del clima laboral y su sentimiento frente a la tarea desempeñada. Difícilmente una persona que está por dejar su cargo contestará de manera sincera una encuesta de este tipo [15], por lo que las respuestas se encontrarían sesgadas.

Es necesario agregar que las encuestas de clima laboral son generalmente anónimas y entregan solo un promedio por área de la empresa para cada uno de los tópicos de la encuesta, por lo que se carece de datos a nivel de individuos. Dado que en la empresa estudiada no existe registro histórico de encuestas que reflejen este tipo de variables, sumado a que no se tiene información a nivel de individuo, éstas no serán consideradas dentro de las variables a utilizar dentro del presente trabajo.

Además de los conjuntos de variables mencionadas anteriormente, existen variables externas que afectan el comportamiento de los empleados. Por ejemplo, el sueldo que paga la competencia por un cargo similar o indicadores a nivel país, como la tasa de desempleo. Para este caso particular, se considerará el precio del cobre, puesto que cuando éste aumenta, se abren vacantes en el sector y al disminuir, generalmente se reduce el personal.

Con respecto a las variables demográficas y del trabajo, éstas son registradas al momento de ingreso de una persona a la compañía y durante su estadía, con propósitos de remuneración

principalmente. Cuando un empleado deja la empresa o es despedido, se registra su egreso, quedando la historia de la persona en las bases de datos del departamento de Recursos Humanos. Con esto, se pueden buscar relaciones básicas del estilo "las personas más jóvenes son más propensas a cambiar de empleo" o "personas que viven más alejadas del lugar de trabajo son más propensas a dejar la empresa". Sin embargo, y dada la cantidad de variables que se registran de las personas, se plantea utilizar el enfoque de modelos predictivos que ayuden a prever este comportamiento y adelantarse a la acción, que pueden ser utilizados como apoyo para los especialistas de RR.HH. de la empresa y supervisores.

Es importante diferenciar que la rotación de empleados puede suceder tanto por el despido como por la renuncia voluntaria de éste. Para el presente trabajo se considerará esta diferencia, considerando el caso de que la persona permanezca en la organización, teniendo un problema con tres clases posibles: permanencia en la empresa, despido o renuncia voluntaria.

1.3. Minería de Datos y Recursos Humanos

En base a la problemática planteada de Recursos Humanos, se utilizarán modelos cuantitativos para la detección de patrones comunes a la fuga de un empleado de la compañía, para poder comprender el fenómeno, poder explicarlo y en consecuencia, anticiparse a éste. Para esto, se propone utilizar métodos de minería de datos, los cuales ya han sido abordados para generar conocimiento a partir de sistemas de Recursos Humanos. En la literatura se conoce esto como "*Human Resources Analytics*".

Existen diversas tareas de *Human Resources Analytics*, las cuales se pueden resumir [16] como se expone a continuación:

- Clustering: Sirve para agrupar empleados de características similares, buscando agrupaciones naturales de los datos. De esta manera se encuentran grupos de empleados de similares características los cuales sirven para identificarlos y tomar distintas acciones o tratamientos con cada uno de ellos, acorde a lo encontrado.

- Clasificación y Predicción: Contribuyen a predecir el comportamiento de los empleados, su rendimiento futuro en la empresa, la probabilidad de fuga, cuáles beneficios son más atractivos para futuros postulantes, entre otros.

- Asociación: Ayuda a encontrar cuáles atributos son más atractivos para atraer a nuevos empleados y asociarlos con los objetivos estratégicos de la empresa, o analizar cómo se mueve dentro de la organización el personal de características similares.

- Uso de métodos interdisciplinarios: Reportar el conocimiento encontrado, además de resumir la data para su análisis. Por ejemplo, se puede ver cómo se caracteriza el personal de planta, sus compensaciones, la tasa de fuga además de otros elementos de la Gestión de Personas.

En el caso particular de la fuga de empleados, se puede encontrar en la literatura el caso de la industria tecnológica en China [17], la industria de los semi conductores en Taiwán [18], empleados de una oficina de ventas [11] y la industria prestación de servicios a la minería en Chile [19]. En estos documentos se explica cómo el uso de modelos de Minería de datos, en particular Self Organized Maps (un híbrido entre Redes Neuronales y Clustering), Rough Set Theory (modelo que busca relaciones en la data y las expresa en forma de reglas de decisión) y Support Vector Machines (que permite separar en base a atributos quienes dejan o no la empresa, mostrando los atributos que influyen en el comportamiento). Además se detalla la medida de precisión de estos modelos (accuracy, recall, true positive rate, entre otras) comparados con otros, viendo cual tiene mayor poder predictivo para ese caso particular. Las conclusiones en estos casos no son iguales, debido principalmente a que el comportamiento del personal va a depender de diversos factores del contexto en el que se encuentra trabajando. Por mencionar algunos, la posición de la empresa del mercado donde compite, el rubro, el perfil de personas analizadas, entre otros.

Estos estudios tienen una idea en común: el uso de datos de la empresa para poder anticipar comportamiento de empleados para tomar mejores decisiones estratégicas en lo que respecta a sus recursos humanos. Así, por ejemplo, se puede predecir el desempeño de un empleado entrante, lo que lleva a replantearse las políticas de contratación de nuevo personal.

La idea principal de este trabajo es comprender el fenómeno de la fuga en la empresa, encontrando que factores influyen en que un empleado deje la empresa de manera voluntaria y los patrones comunes a este comportamiento. Esto permitirá identificar anticipadamente a las personas propensas a renunciar voluntariamente, evitando las potenciales ineficiencias mencionadas en la sección anterior.

Se contempla una variable dependiente que posee tres clases como se menciona anteriormente: permanencia en la empresa, fuga voluntaria o despido. Los estudios mencionados anteriormente poseen dos categorías: fuga/permanencia o fuga voluntaria/despido. El principal objetivo de esta diferenciación, pasar de un problema binario a un problema de multi-clasificación, es tomar en cuenta que ambos comportamientos (despido y renuncia) son muy distintos y por lo tanto, se necesita generar la distinción. Si bien, el objetivo de este proyecto no es predecir los despidos, estos serán separados como una variable "clase" para poder diferenciarlos tanto de las personas que permanecen en la empresa en el periodo estudiado, como de las que renuncian voluntariamente.

1.4. Marco Legal de Capacitaciones

En Chile, el marco legal acerca de las capacitaciones se crea con el propósito de promover el mejoramiento de las habilidades y conocimientos de los empleados de la industria nacional, mejorando los procesos productivos y contribuyendo al desarrollo del país. La ley 19.518 [20] en relación a esto declara lo siguiente:

Artículo 1º.- El sistema de capacitación y empleo que establece esta ley tiene por objeto promover el desarrollo de las competencias laborales de los trabajadores, a fin de contribuir a un adecuado nivel de empleo, mejorar la productividad de los trabajadores y las empresas, así como la calidad de los procesos y productos.

Artículo 3º.- En materia de fomento del empleo, el sistema comprende acciones encaminadas a:

a) Fomentar el desarrollo de aptitudes y competencias en los trabajadores que faciliten su acceso a empleos de mayor calidad y productividad, de acuerdo a sus aspiraciones e intereses y los requerimientos del sector productivo, y

b) Estimular el desarrollo y perfeccionamiento de mecanismos de información y orientación laboral, así como la asesoría técnica y la supervisión de los organismos que desarrollen dichas funciones.

Para la ley, una capacitación se define de la siguiente manera.

Artículo 10.- Se entenderá por capacitación el proceso destinado a promover, facilitar, fomentar, y desarrollar las aptitudes, habilidades o grados de conocimientos de los trabajadores, con el fin de permitirles mejores oportunidades y condiciones de vida y de trabajo y de incrementar la productividad nacional, procurando la necesaria adaptación de los trabajadores a los procesos tecnológicos y a las modificaciones estructurales de la economía.

Se considerarán también capacitación, las actividades destinadas a desarrollar las aptitudes, habilidades o grados de conocimientos de los dirigentes sindicales, cuando éstas sean acordadas en el marco de una negociación colectiva o en otro momento, y tengan por finalidad habilitarlos para cumplir adecuadamente con su rol sindical.

El programa y financiamiento contemplados en este artículo para programas de capacitación orientados a trabajadores que tengan la calidad de dirigentes sindicales, serán sin perjuicio de otros programas y fuentes de financiamiento público, contemplados en otros cuerpos legales.

La ley establece ciertas obligaciones para las empresas, entre las cuales destacan:

Artículo 13.- Las empresas podrán constituir un comité bipartito de capacitación. Ello será obligatorio en aquellas empresas cuya dotación de personal sea igual o superior a 15 trabajadores. Las funciones del comité serán acordar y evaluar el o los programas de capacitación ocupacional de la empresa, así como asesorar a la dirección de la misma en materias de capacitación.

Artículo 30.- Incumbe a las empresas, por sí o en coordinación con los Comités Bipartitos de Capacitación, en todos los niveles jerárquicos, atender las necesidades de capacitación de sus trabajadores. Los programas de capacitación que desarrollen en conformidad al Estatuto darán lugar a los beneficios e impondrán las obligaciones que señala este cuerpo legal.

Para los efectos de lo dispuesto en este artículo, el término trabajador comprende también a las personas naturales y socios de sociedades de personas que trabajan en las empresas de su propiedad.

Expuesto lo anterior, el artículo 13 de la ley obliga a las empresas en el caso de tener más de quince trabajadores a contar con un comité de capacitación, manteniendo su responsabilidad frente al tema de la capacitación de sus trabajadores. Este comité tiene la obligación de evaluar los cursos que se llevan a cabo, como también promover la instauración de nuevos cursos que respondan a las necesidades de los empleados o eliminar los programas que ya no respondan a éstas. Dentro de las obligaciones de la empresa se encuentra no impedir la elección de los representantes de los trabajadores.

Las capacitaciones deben ser llevadas a cabo por un Organismo de Técnico Capacitación (OTEC) acreditados por el Servicio Nacional de Capacitación y Empleo (SENCE). Este organismo tiene como objetivo aumentar la competitividad de las empresas en territorio nacional y la empleabilidad de las personas. La capacitación de los empleados por medio de las OTEC tiene un beneficio tributario para la empresa que lo realiza, además de una acción subsidiaria que otorga recursos para la capacitación de empleados mediante programas de becas financiadas con recursos públicos. Generalmente, estos se realizan a través de un Organismo Técnicos Intermedios para Capacitación (OTIC) que se encargan de ser un nexo entre las empresas y las OTEC, supervisando los programas de capacitación que realizan los empleados de la empresa y llamando a licitaciones para llevarlos a cabo.

Un curso muy común para de los operarios de la minería es el llamado “Baprever” que tiene relación con los riesgos laborales. Esta acreditación no puede ser exigida por los empleadores

mineros, debido a que se califica como un acto discriminatorio según el dictamen 0352/024 de la Dirección del Trabajo del año 2004.

En el caso de las renuncias voluntarias el dictamen 4924/269 de la Dirección del Trabajo [21], establecido en 1997, declara que el empleador no puede pedir fianzas u otras formas de garantías por las capacitaciones entregadas al trabajador. Por lo tanto, el empleador asume un "riesgo" al momento de capacitar a sus empleados debido al alza que genera en su empleabilidad.

1.5. Objetivo General

"Determinar los factores que influyen en la renuncia voluntaria de operarios en el sector minero, encontrando patrones comunes a este comportamiento".

1.6. Objetivos Específicos

- Identificar los operarios mina con mayor propensión a renunciar voluntariamente
- Caracterizar el comportamiento de distintos tipos de operarios que renunciaron voluntariamente, mediante clustering.
- Describir los distintos tipos de fuga (despido y renuncia) en base a los atributos de la información histórica.
- Determinar qué modelo de minería de datos describe mejor el comportamiento estudiado.
- Proponer acciones para retener a los empleados con propensión a renuncia.

2. Caracterización del problema

En el presente capítulo se lleva a cabo la revisión de los datos disponibles para el estudio, donde se muestran distintas visualizaciones que ayudan a comprender, a priori, el comportamiento histórico de los empleados de la empresa a estudiar. Luego, se hace referencia al funcionamiento general de la empresa, explicando el proceso en el cual trabajan las personas estudiadas, para luego mostrar datos acerca de las renuncias y despidos dentro de la compañía.

2.1. Datos disponibles

Dentro de las fuentes internas de datos, se encuentran tanto los datos demográficos de las persona, como sus licencias médicas y las capacitaciones realizadas durante sus años de servicio dentro de la empresa. Existen también datos obtenidos de otras fuentes, tales como una matriz de distancias de ciudades en Chile[1] y el precio de la libra de cobre transada en la LME[2]. Se cuenta con datos históricos de la empresa desde el año 1998 hasta finales del mes de abril del año 2015. Para el caso del precio del cobre, se cuenta con registros mensuales desde enero de 1999 hasta la fecha. Las remuneraciones de empresas que pertenecen al mismo sector industrial es un dato de carácter privado, por lo que no será considerado para este estudio. Se debe considerar además que este trabajo se centrará en los operadores mina pertenecientes a la empresa y no quienes son personal de empresas contratistas.

A continuación se enunciarán las variables disponibles en la base de datos de la compañía y las variables calculadas a partir de éstas, las que se encuentran categorizadas de acuerdo a lo expuesto en la sección 1.2. Se indicará la tasa de renuncia considerando toda la historia de la compañía. Si se considera toda la historia de la empresa, del total de su dotación histórica un 30% de operarios mina renunciaron voluntariamente a su cargo, dato que se usa como referencia (recta horizontal) en las ilustraciones a continuación.

2.1.1. Variables Demográficas

- Género: variable binaria que considera el valor 1 para hombre y 0 para mujer. Un 95% de la muestra corresponde a hombres. Si bien las mujeres presentan una tasa de fuga bastante más baja que el promedio, igual a 9%, existen muy pocos registros en comparación a los hombres, los cuales al ser mayoría, poseen una tasa de fuga prácticamente igual al promedio de la empresa.

[1] Disponible en: http://www.vialidad.cl/productosyservicios/paginas/distancias.aspx
[2] Disponible en: http://www.cochilco.cl/estadisticas/precio-metales.asp

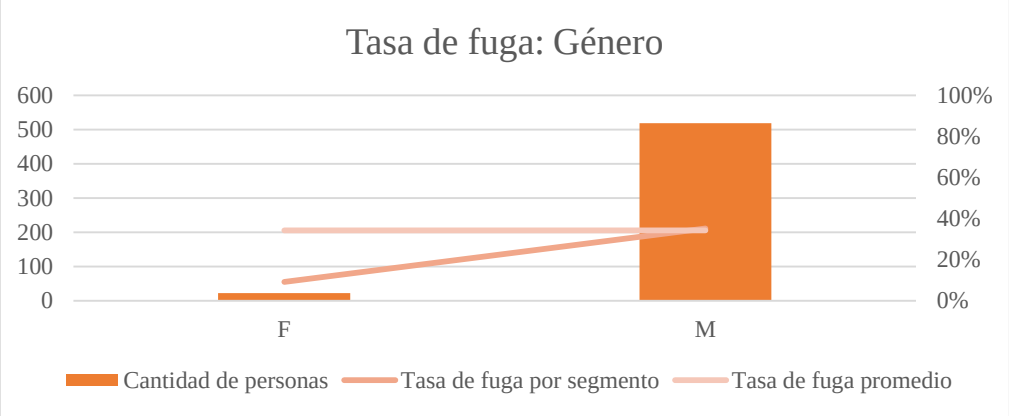


Ilustración 7 - Tasa de fuga por género

- Distancia Hogar: utilizando una matriz de distancias disponible en la Dirección de Vialidad, se agregó la distancia desde la faena a la ciudad de residencia. La media es aproximadamente 290 km. La distancia más común corresponde a 120 km., correspondiente a la distancia de la faena a la ciudad de Antofagasta, siendo la ciudad más cercana y por ende, es el lugar donde vive la mayoría de los trabajadores. La tasa de fuga oscila junto al promedio para todos los tramos de distancias, por lo que en una primera instancia no se podría asegurar que la distancia sea una variable relevante al tomar la decisión de renuncia.

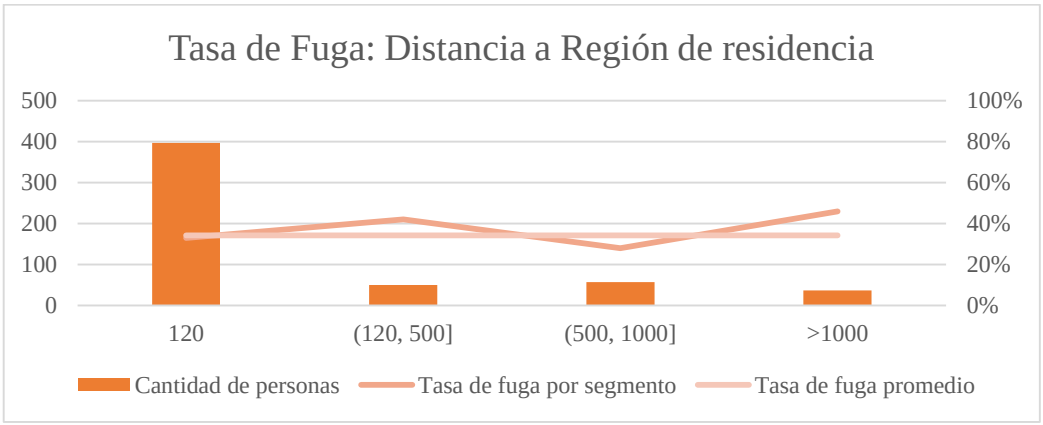


Ilustración 8 - Tasa de fuga: Distancia al hogar

- Estado Civil: variable categórica que considera los estados “casado”, “convivencia”, “soltero” y “divorciado”. Aproximadamente un 50% de la muestra se encuentra casada. Similar al caso de género, los segmentos “Convivencia” y “Divorciado” muestran una tasa de fuga baja con respecto al promedio (12% y 1% respectivamente).

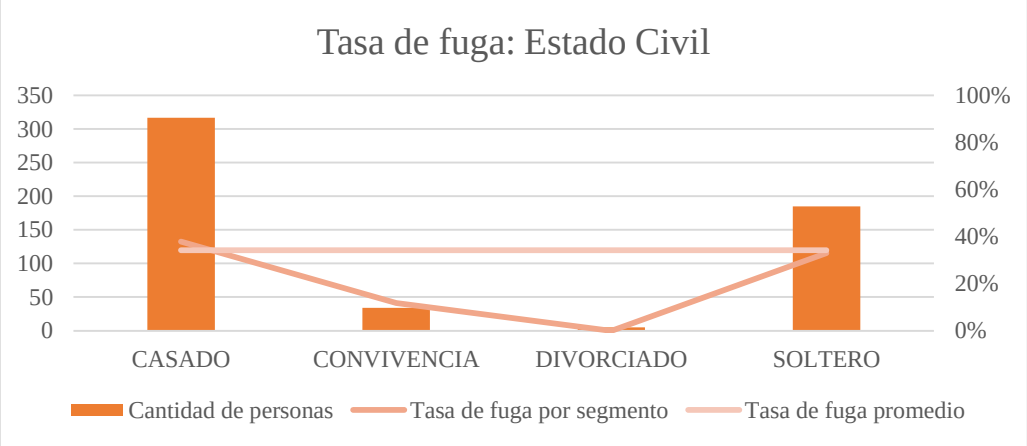


Ilustración 9 - Tasa de fuga por Estado Civil

- Mínimo Nivel de Escolaridad: variable categórica que indica el nivel mínimo registrado de escolaridad de la persona. Dado que es el mínimo del que se tiene registro, se asume que es el nivel de escolaridad que tenía el trabajador al momento de ingresar a su puesto de trabajo.

- Máximo Nivel de Escolaridad: nivel máximo registrado de escolaridad de la persona. Se asume que es su estado actual. A simple vista, no se aprecian segmentos representativos que muestren una tasa de fuga distinta al promedio.

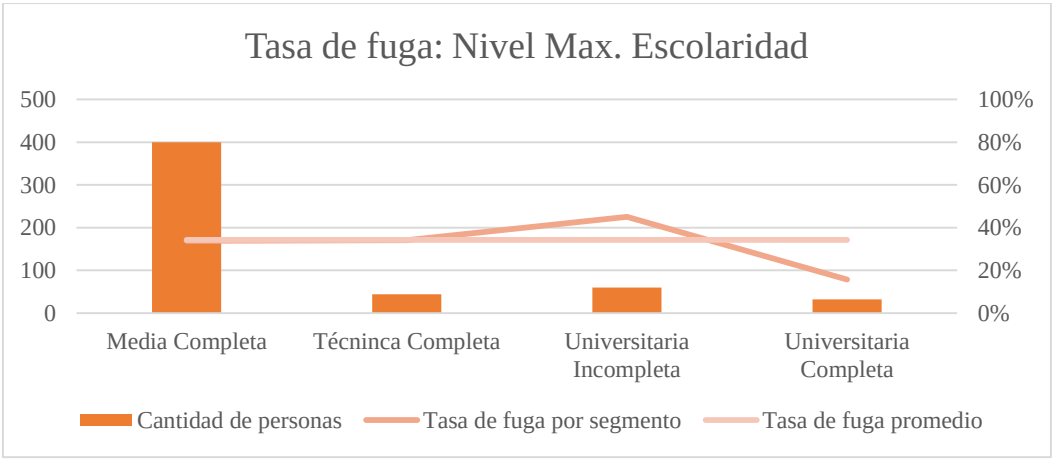


Ilustración 10 - Tasa de fuga por máximo nivel de escolaridad

- Edad en la empresa: esta variable numérica considera la diferencia entre la fecha de nacimiento de la persona y el último día trabajado por ésta. Así, si un empleado sigue trabajando en la empresa, se calcula su edad utilizando la diferencia entre su fecha de nacimiento y Abril de 2015, fecha en que fue extraída la información de la base de datos para este estudio. En cambio, si la persona ya no está en la empresa, se considera su fecha de egreso para calcular su edad. La media de edad es 42 años. La tasa de fuga más alta se presenta entre los 35 y 40 años, sin embargo, no se aprecia a simple vista un comportamiento uniforme a través de los años.

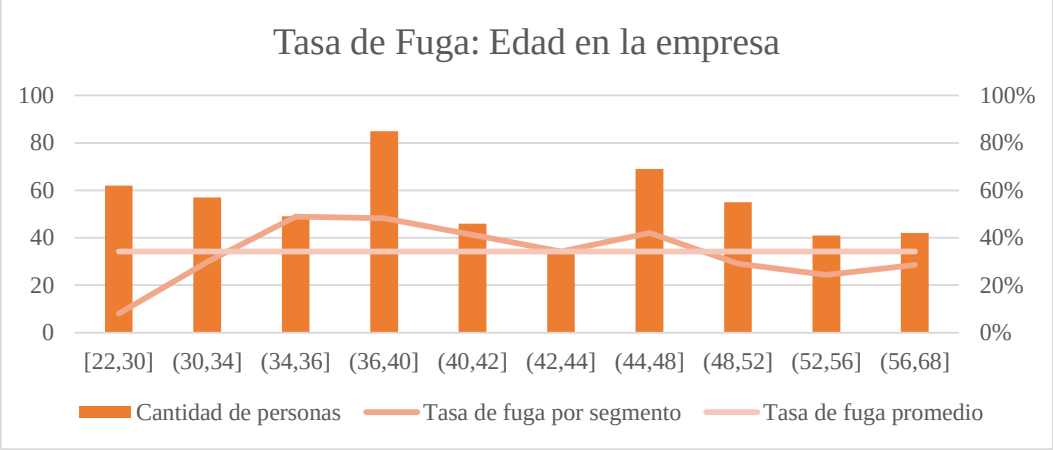


Ilustración 11 - Tasa de Fuga por edad

- Licencia Profesional: variable binaria que registra el valor 1 para la persona que tiene licencia de conducir profesional y 0 si no la posee. Un 54% de la muestra posee una licencia profesional. Ambos grupos son bastante homogéneos, presentando una tasa de fuga del 29% quienes no poseen una licencia profesional y 39% quienes sí la poseen.

2.1.2. Variables Características del Trabajo

- Condición sindical: variable binaria que indica si la persona adhiere al sindicato de operadores de la empresa. Cerca de un 70% de los operarios adhiere al sindicato. Cada segmento es homogéneo relativo a la fuga, teniendo una tasa del 36% quienes no adhieren al sindicato y un 34% quienes sí lo hacen.

- Beneficio por renuncia: según lo estipulado en el contrato colectivo sindical, una persona que adhiera al sindicato y posea más de 6 años de servicio en la empresa, podrá acceder a una bonificación al momento de renunciar voluntariamente, recibiendo su último sueldo mensual multiplicado por la cantidad de años de estadía. Al igual que el caso anterior, aunque parezca contra intuitivo, ambos segmentos no presentan mayores diferencias del promedio de la tasa de fuga.

- Tiempo en la empresa: al igual que lo realizado con la edad, se tiene esta variable que indica el tiempo que lleva la persona trabajando para la empresa. En el caso de los fugados voluntariamente, se considera desde su fecha de ingreso hasta su egreso. En el caso de quienes permanecen, desde su ingreso hasta el día del último registro. El promedio de permanencia para el total de empleados es de 5 años y medio. Las mayores tasas de fuga se ven en los primeros años de servicio. Sin embargo, existe una crecida entre los 5 y 6 años de permanencia (el promedio de la empresa) que no permite establecer una regla clara acerca de la tasa de fuga con respecto la cantidad de años en la empresa.

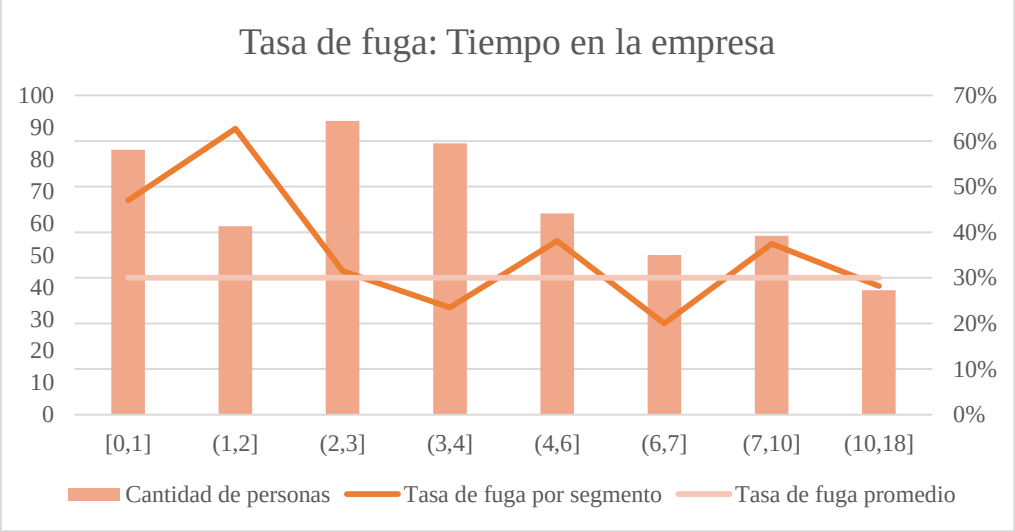


Ilustración 12 - Tasa de fuga por Años de Servicio

- Tramo de Sueldo: los sueldos en la compañía se ordenan mediante niveles y sub niveles de sueldo. En cuanto a los niveles, los operadores pueden tener nivel A, B o C, donde se cumple que $C < B < A$ en valor monetario. Luego, los subniveles van desde el 1 al 5, en orden ascendente. De esta manera, se cumple para todo nivel X que $X1 < X2 < X3 < X4 < X5$. Por lo tanto, se tiene una variable con 15 categorías, que va desde el C1 al A5. Finalmente, se debe mencionar que una persona que ha alcanzado el subnivel máximo correspondiente a su nivel, no puede acceder a un nivel superior. El segmento A es quien posee la mayor tasa de fuga.

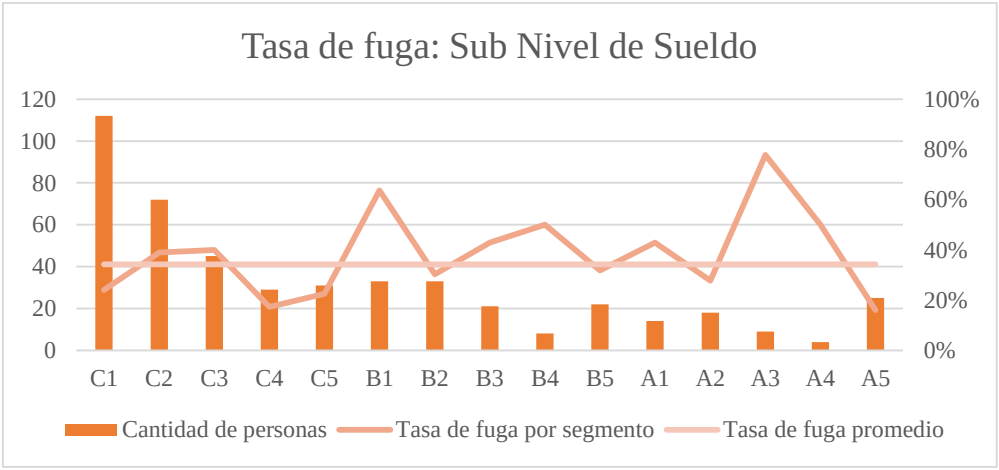


Ilustración 13 – Tasa de fuga por subnivel de sueldo

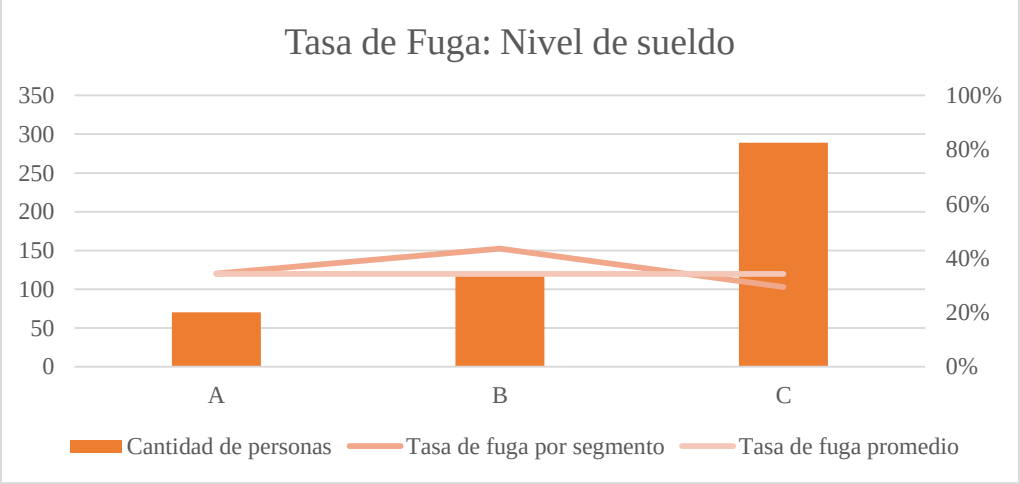


Ilustración 14 - Tasa de fuga por nivel de sueldo

- Máximo subnivel: variable binaria que indica 1 si la persona alcanzó el máximo subnivel posible (A5, B5 o C5) y 0 de lo contrario. Si bien estas personas no pueden seguir ascendiendo en la escala de sueldos, se les entrega un bono de compensación a fin de año por esta razón. Al parecer, es una medida que ha tenido éxito puesto que la tasa de fuga para el grupo de personas que poseen el máximo de su escala de sueldos es menor que para el resto de las personas estudiadas.

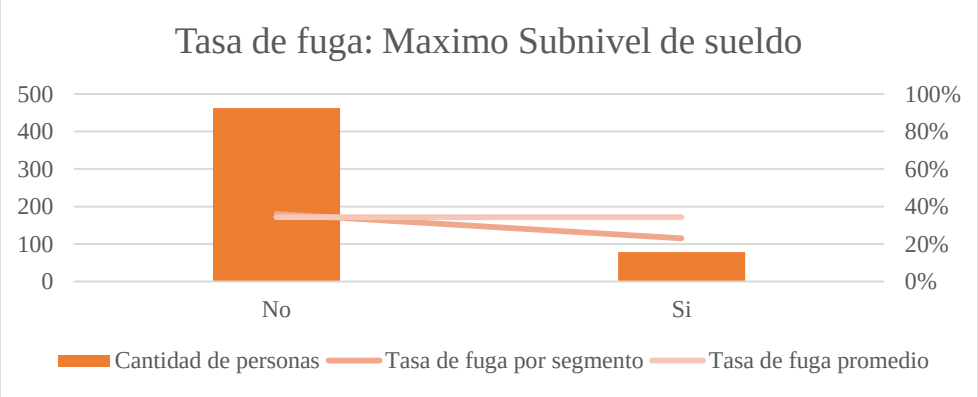


Ilustración 15 - Tasa de fuga: Máximo Subnivel de sueldo

- Franquicia SENCE: variable categórica que indica el porcentaje del valor que franquicia SENCE de una capacitación. Este porcentaje, al ser mayor, indica que el sueldo de la persona es menor.

- Capacitaciones anuales: cantidad promedio anual de capacitaciones que ha realizado el empleado durante su estadía en la empresa. Las personas que realizan menos de una capacitación en promedio al año (por ejemplo, realizan una capacitación cada dos años) son quienes tienen una mayor tasa de fuga.

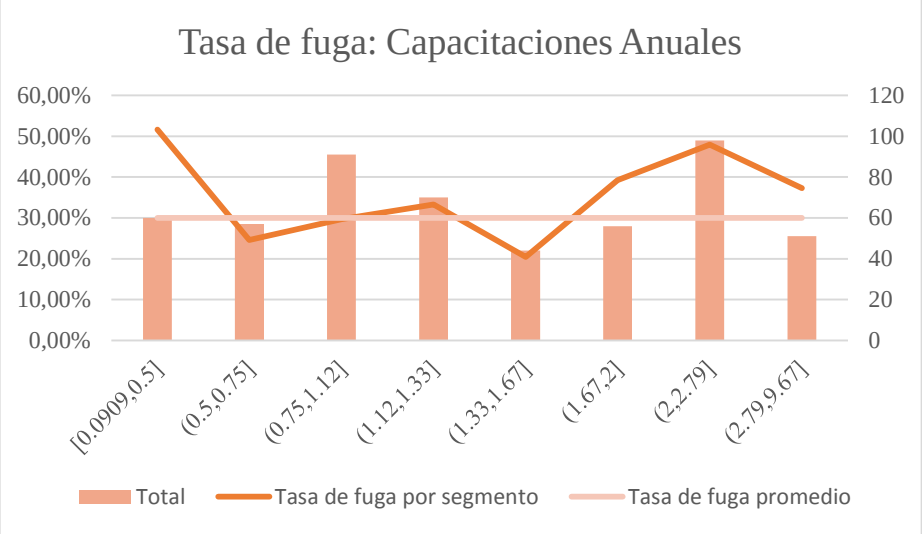


Ilustración 16 - Tasa de fuga por número de capacitaciones anuales

- Inversión anual en capacitación: indica el valor invertido, en promedio, en capacitaciones anualmente por parte de la empresa en el empleado durante la estadía del operario. Es claro ver que, dado lo anterior, en personas en las cuales se invierte menos en términos anuales en capacitación poseen una mayor tasa de fuga que en quienes se invierte una mayor suma.

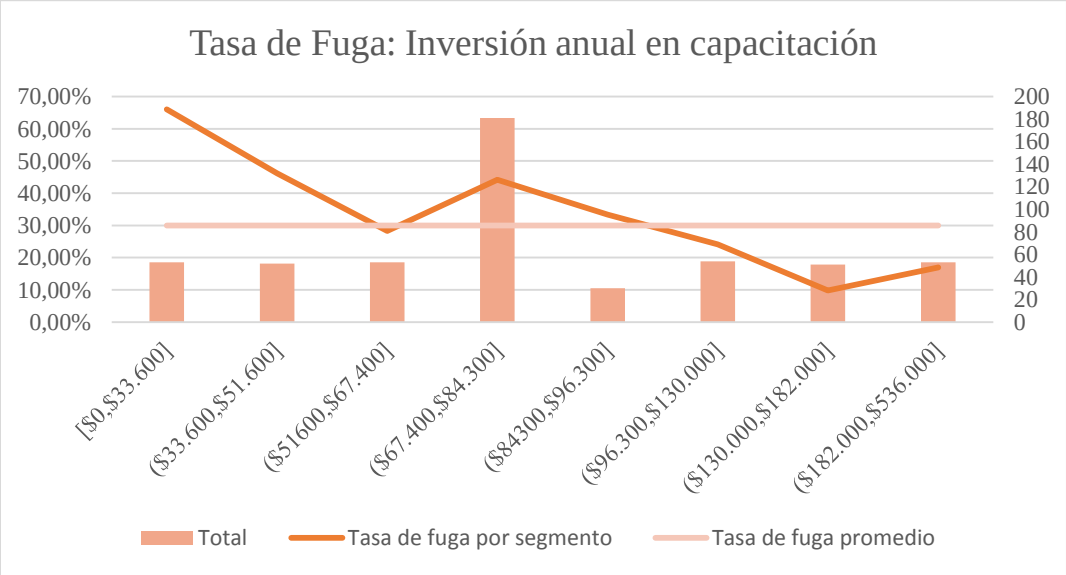


Ilustración 17- Tasa de fuga por inversión anual en capacitación

- Total invertido en capacitación: total de dinero invertido por la empresa en el empleado en términos de capacitación.

- Promedio de notas en capacitación: promedio ponderado por el número de horas de la capacitación. De esta manera los cursos de mayor cantidad de horas, que tienden a ser más complejos, cobran mayor relevancia dentro del promedio de la persona.

- Índice de aprobación: indica el porcentaje de capacitaciones aprobadas en relación al total de horas de capacitación realizadas. Como se puede ver, quienes renuncian voluntariamente son quienes aprueban la mayor cantidad de horas realizadas.

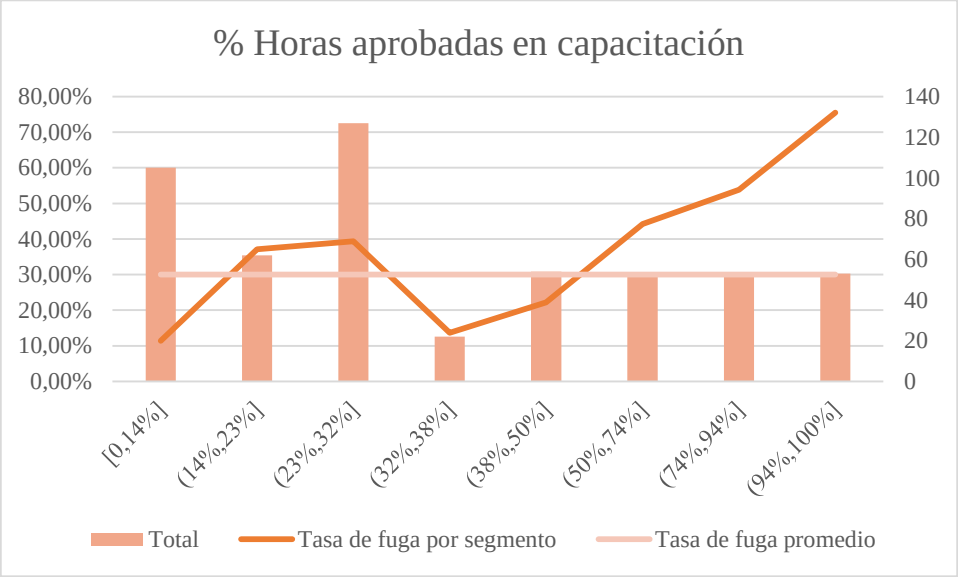


Ilustración 18- Tasa de fuga por cantidad de horas de capacitación

- Periodo de prueba: variable binaria que toma el valor 1 si la persona estuvo en periodo de prueba antes de ser contratado y 0 si no. La muestra de personas que hicieron periodo de prueba a la empresa es de alrededor de 100 personas, quienes poseen una tasa de fuga igual a 24%, que se encuentra por debajo del promedio. El resto del universo de personas tiene una tasa de fuga muy similar al promedio.

- Demora de contrato indefinido: tiempo en meses que tardó que la persona terminara su periodo de prueba y firmara contrato indefinido. Si es igual a 0, entonces la persona no hizo período de prueba. Al aumentar la cantidad de meses de demora, no se aprecia una gran variación en la tasa de fuga, la que oscila el 34% promedio.

- Número de cuotas pagadas: se refiere al número de cuotas pagadas del beneficio habitacional que ofrece la compañía a sus trabajadores, con un máximo de 60 cuotas. Cuando el número de cuotas aumenta, se puede apreciar un aumento en la tasa de fuga. Sin embargo, cuando el crédito se encuentra pagado en su totalidad la tasa se estabiliza en el promedio.

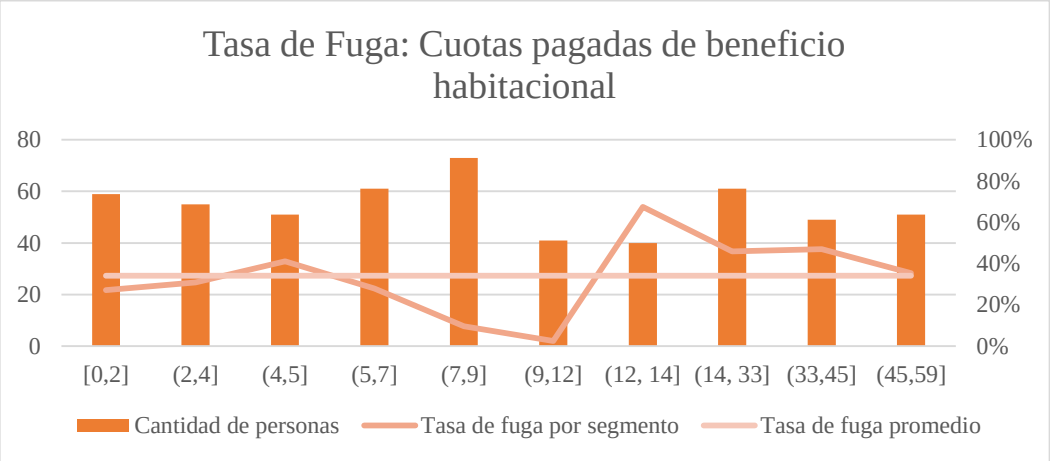


Ilustración 19 - Tasa de fuga por cuotas pagadas de beneficio habitacional

- Pago completo: variable binaria que indica 1 si la persona ya efectuó el pago completo de su beneficio habitacional y 0 si no. Si bien, un crédito puede tener menor cantidad de cuotas que el máximo de 60, la tasa de fuga de quienes tienen un pago completo de su crédito es solo un 6% mayor que la del promedio.

- Reingreso a la empresa: si la persona posee un ID en la empresa distinto a su Rut, quiere decir que ésta ya había trabajado previamente en la compañía y volvió. Por lo tanto, se tiene una variable binaria que indica 1 si se hizo reingreso y 0 si no. Solo existen 16 casos de personas que hicieron reingreso, de las cuales dos personas solamente dejaron la empresa de manera voluntaria luego de reingresar.

- Causa de la última licencia: esta variable indicará la causa de la última licencia de la persona, la cual es una variable categórica que puede tomar el valor “médica”, “falta”, “permiso” u “otra”.

- Cantidad de días totales de licencias por tipo: este ítem incluye la cantidad de días totales de licencias por tipo (falta, permiso, licencia médica u otra) durante los años de servicio.

- Target: variable objetivo del estudio. Posee tres categorías: permanencia en la empresa, despido y renuncia voluntaria. Se puede ver que las clases se encuentran bastante equilibradas, lo cual facilita el trabajo en comparación a casos donde las clases se encuentran desbalanceadas.

Clase	Frecuencia
Permanencia	162
Despido	194
Renuncia Voluntaria	185
Total	541

Tabla 2 - Variable Objetivo

Dado que las variables enunciadas son estáticas, se agregará la siguiente información disponible para ver un cambio en el comportamiento:

- Meses transcurridos desde la última capacitación: se considera, para el caso de una persona que dejó la empresa, la diferencia en meses desde la última capacitación realizada hasta su egreso de la compañía. En el caso de personas que permanecen en la empresa, se considera la diferencia entre Abril de 2015 y la última capacitación realizada. Se puede apreciar que las personas que han realizado recientemente una capacitación son propensas a dejar la empresa, y también se tiene el mismo efecto cuando transcurre más de 1 año y medio sin éstas.

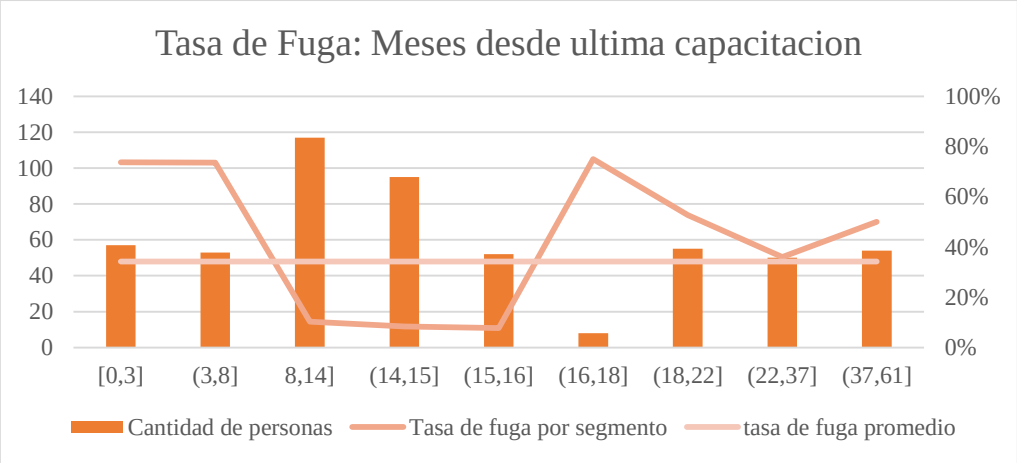


Ilustración 20 - Tasa de fuga por meses desde última capacitación

- Meses desde última licencia: se considera la cantidad de meses desde la última licencia hasta el último mes trabajado. En el caso de las personas que permanecían en la empresa, se consideró abril de 2015 como último mes trabajado.

- Valor del precio del cobre al momento del egreso: en el caso de quienes permanezcan en la empresa, se tomará el valor correspondiente al último registro de la base de datos, correspondiente a abril de 2015.

- Valor del precio del cobre 6 meses antes del egreso: se consideró un plazo de 6 meses, ya que la decisión de renunciar a un trabajo no se toma, en la mayoría de los casos, de un momento a otro, ya que las personas comienzan a buscar nuevas opciones en otro lado para mantener su estabilidad económica.

- Variación del precio del cobre a 6 meses: se considera la variación porcentual de ambos valores para poder observar cómo las alzas o bajas del precio del cobre afectará la salida de personas de la empresa. Se puede observar que durante el periodo estudiado la variación del cobre es cercana al -10%, lo cual se refleja en la figura a continuación. Para casos en

que el precio sube, es posible que las personas renuncien para tomar posición en alguna empresa de la competencia, puesto que altos precios del cobre implican la apertura de nuevos puestos de trabajo en minería. Para el caso de la baja del cobre, el ambiente de inestabilidad puede generar inseguridad en las personas, impulsando a tomar la decisión de cambiar de trabajo, pero en otra industria.

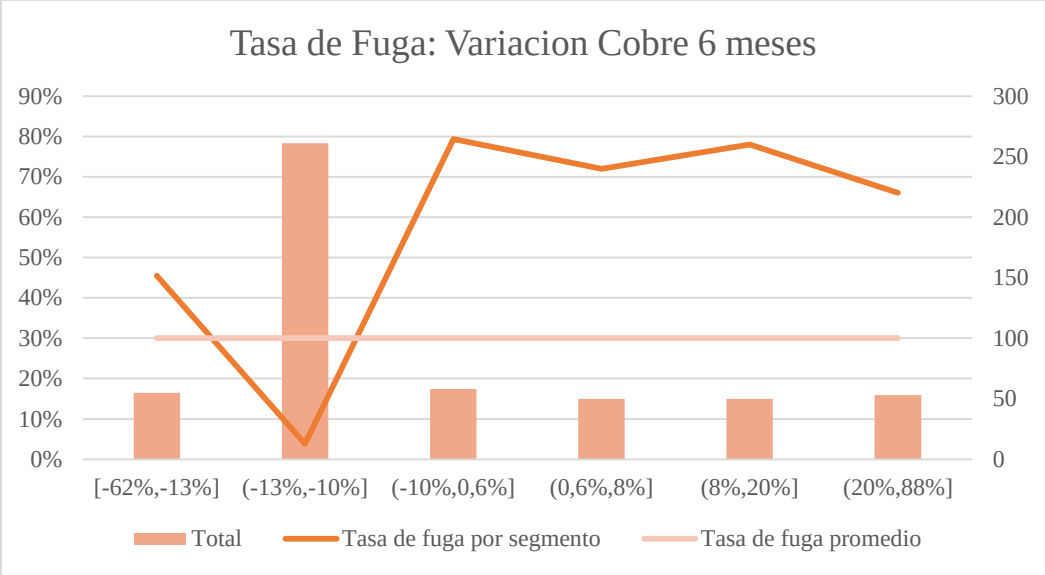


Ilustración 21 - Tasa de fuga por variación en el precio del cobre

- Variación del precio del cobre a 3 meses: se consideró además la variación a tres meses del valor del precio del cobre, de la misma manera anterior. No existen mayores diferencias en el comportamiento de ambas variables.

Tomando en cuenta que el proceso de renuncia voluntaria de una persona no es instantáneo y que obedece a un proceso de evaluación, se consideraron estos dos periodos de tiempo para comparar los efectos de cada uno en la población estudiada y observar qué variable (3 o 6 meses) era más significativa para el modelo.

En cuanto a las variables respectivas a capacitaciones, se debe mencionar que estas en ningún caso corresponden al curso “Baprever” mencionado en la sección 1.4. Si bien este curso se realiza dentro de la empresa, no está considerada dentro de las capacitaciones que se integran dentro de las variables del problema en estudio, por lo tanto, las capacitaciones consideradas para este trabajo son de corte específico para una tarea que desempeñe el operario.

2.2. Contexto de la organización

La empresa a estudiar, perteneciente a la mediana minería del país, dio inicio a sus operaciones el año 1998. Se dedica a la obtención de cátodos de cobre mediante minería a rajo abierto, proceso que a grandes rasgos, funciona de la siguiente manera:

- Chancado: proceso en el cual la piedra es triturada. Aquí se separan el material de alta ley, llamado Heap, y de baja ley, llamado Rom.
- Lixiviación en pilas: se agregan químicos al material triturado, generando un líquido denominado PLS.
- Extracción por Solventes: el PLS se dirige a la planta SX para preparar la obtención de cátodos.
- Electro-obtención: gracias a la utilización de energía eléctrica, el cobre adopta forma de plancha. Este proceso dura aproximadamente 5 días.

Dentro de este proceso es de suma importancia el traslado del material para cada una de sus etapas, así como la operación de la maquinaria correspondiente. Es por esta razón que la ausencia de operadores de equipo de alto tonelaje como estos impacta directamente en el proceso productivo.

La planificación de producción minera, tal como en muchas industrias que requieren de gran inversión inicial, se hace durante la etapa de evaluación del proyecto cuando aún no se ha comenzado a instalar la faena. Aquí, se estima la cantidad de recursos y su vida útil mediante sondajes, que dirán la cantidad de mineral que puede ser explotado durante un periodo de tiempo determinado. Debido a que los recursos minerales no son inagotables, se estudia la cantidad de activos necesarios para su explotación a través del tiempo, considerando un proceso de turnos de trabajo continuos durante las 24 horas del día durante todo el año. Estos activos van desde los equipos de alto tonelaje (palas, camiones, equipos de apoyo), como los químicos necesarios, el suministro de agua y energía, además de la dotación de personal necesaria para llevar a cabo la operación. Desajustes en la planificación, como por ejemplo cambios en la dotación de personal debido a renuncias voluntarias, terminan por dificultar el cumplimiento de los objetivos estratégicos de la compañía.

Dicho lo anterior, durante el año 2012 se realizó un estudio de pre factibilidad que permitirá ampliar la producción de la compañía en un 45%, lo que demandará un aumento de la dotación de personal en al menos 300 personas que se contemplan para iniciar sus actividades en 2017.

Hacia el año 2013, en la empresa trabajaban alrededor de 150 operarios de mina. Durante el año 2011, cuando la libra de cobre registró un valor histórico al alcanzar los $4 USD por libra en la Bolsa de Metales de Londres (LME) fue cuando se registró una mayor cantidad de egresos de la compañía, como se puede apreciar en la figura 22.

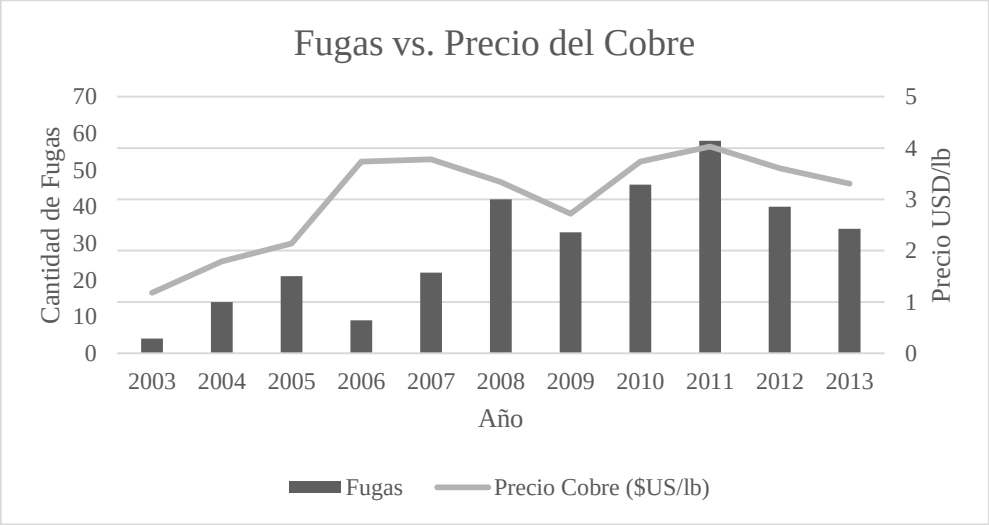


Ilustración 22 - Fuga vs. Precio del cobre

Si bien es claro que existe una relación del valor del precio del cobre con las fugas a partir de 2009, donde se ve claramente cómo ambas variables crecen, no es claro que esto sea una regla estricta, puesto que se puede apreciar que para el periodo 2005-2008 esto no se cumple, teniendo una baja cantidad de fugas con respecto años posteriores y un precio para el metal rojo cercano a lo registrado entre 2010 y 2012. Por lo tanto, y dado que existen registros, es que se plantea incorporar variables internas tanto de los trabajadores (como los son las variables demográficas) como de la compañía (características del puesto de trabajo). Si se hace la diferenciación entre egresos voluntarios y no voluntarios para este periodo, se observa lo siguiente en la figura 3.

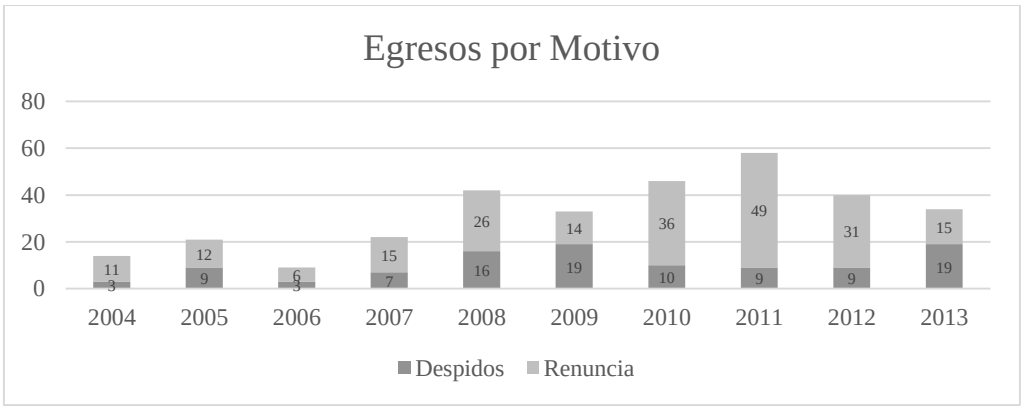


Ilustración 23 - Egreso de personal por motivos

Es claro ver que las renuncias superan en cantidad a los despidos de la compañía. También, se puede apreciar que en épocas donde disminuye el precio del cobre, aumenta la cantidad de despidos, lo cual hace sentido pensando en que las compañías deben disminuir sus gastos en periodos donde el ingreso se ve afectado.

La compañía a estudiar se ha destacado durante su funcionamiento por mantener a su personal capacitado constantemente, lo cual declaran sus propios funcionarios. Sin embargo, al tener una alta tasa de renuncia, sobre todo en periodos del alza del precio del cobre donde se abren nuevas vacantes en empresas pertenecientes a la competencia, se puede deducir que se están perdiendo esfuerzos de capacitación ya que finalmente, se está facilitando personal calificado a competidores. Sólo entre los años 2007 y 2013, se invirtieron aproximadamente 50 $MM (millones de pesos) en empleados que renunciaron voluntariamente a la compañía en términos de capacitación, lo cual no considera el valor franquiciado SENCE. Éste último se define como:

"…el incentivo tributario establecido por la ley N° 19.518 que se otorga a las empresas clasificadas por el Servicio de Impuestos Internos como Contribuyente de Primera Categoría de la Ley de Impuesto a la Renta y que tengan una planilla anual de remuneraciones imponibles superior a 35 UTM."[3]

El gobierno, a través del Ministerio del Trabajo, crea este incentivo para la capacitación de trabajadores en las empresas, en la cual franquicia un porcentaje del costo del curso de capacitación.

Es importante señalar en este punto que aunque la empresa capacite permanentemente a su personal, los sueldos son más bajos que en empresas que tienen un mayor tonelaje de extracción anual del metal rojo. Por lo tanto, es posible asumir que al aumentar el número de vacantes en otras compañías con remuneraciones más altas, el personal de la compañía se cambie a la competencia. Esto se reafirma al generar visualizaciones de los datos de renuncias voluntarias, viendo como las tasas de fuga aumentan al variar el precio del cobre. Igualmente con el tema de capacitaciones, se puede observar que quienes permanecen en la empresa, si bien poseen un mayor número de horas realizadas, su asistencia y aprobación es menor que la de las personas fugadas, dando a entender que quienes realizan un cambio hacia la competencia adquirieron de buena manera el conocimiento, haciéndolo suyo y no parte de la empresa, lo que les da una mayor independencia a la hora de buscar un nuevo puesto de trabajo. No obstante lo anterior la información correspondiente a los sueldos de la competencia es de carácter confidencial y tiene un alto costo de obtención a través de estudios de benchmark, por ende, no será considerada para este estudio.

[3] Disponible en: http://www.dt.gob.cl/consultas/1613/w3-article-60462.html

2.3. Alcances

Los esfuerzos de este trabajo de Memoria estarán enfocados a la utilización e implementación de modelos de minería de datos para describir el fenómeno planteado y el estudio de los patrones involucrados en la fuga, solo generando recomendaciones a partir de los resultados para la retención. Así, el trabajo propuesto entregará un modelo escalable el cual servirá para la toma de decisiones tanto de RR.HH. como la oficina de Planificación Estratégica acerca de los recursos que se deben orientar para mantener a la sus actuales trabajadores en la empresa.

3. Metodología

Esta sección pretende sentar las bases desde las cuales se desarrollará el trabajo, haciendo una breve descripción acerca del marco de trabajo general del proyecto, el KDD, para luego hacer referencia a los distintos modelos que se utilizarán para la detección de patrones de comportamiento que permitan explicar el problema.

3.1. Knowledge Discovery in Data Bases (KDD)

La metodología estándar empleada en proyectos de Data Mining es el KDD [22], que entrega una serie de pasos a seguir para incluir el pre procesamiento de la data para su posterior análisis. La ilustración 24 muestra el esquema base del KDD.

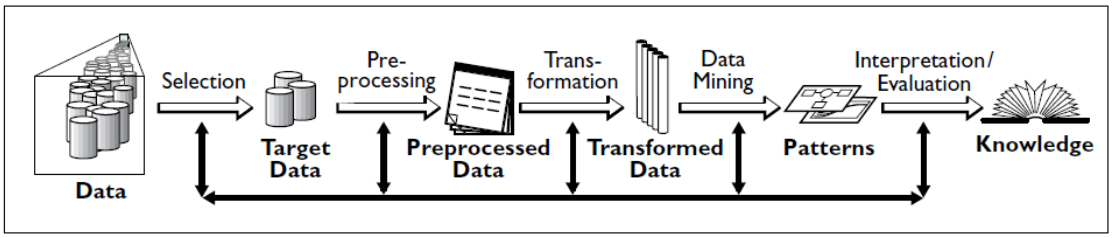


Ilustración 24 - Proceso KDD

- Selección de datos, que consiste en escoger el subconjunto que se desea estudiar de la base de datos original. En este caso particular, de la base de datos original se realizó un filtro para seleccionar solo los operarios de la mina que son la población de la cual se desea comprender su comportamiento. Para esta etapa, se considera crear conjuntos de datos para entrenamiento del modelo y su correspondiente testeo, los cuales contendrán registros de empleados que presentan despidos, renuncias voluntarias y permanencia en la empresa. No obstante, para éste último grupo se seleccionará personal que esté por debajo del promedio de permanencia en la empresa del total de observaciones, es decir, 5 años y medio, para aplicar el modelo y predecir el comportamiento de estas personas.

- Pre Procesamiento de Datos, lo cual incluye verificar la calidad de los datos, valores perdidos y la sumarización de datos. Con respecto a los valores perdidos, estos se pueden trabajar mediante la eliminación o imputación de éstos, dependiendo de la cantidad que sean. Si son pocos datos perdidos, estos se pueden imputar utilizando la media o moda (dependiendo si la variable es numérica o categórica), pero en el caso de que sea un número considerable, estos no deben ser imputados de esta forma, ya que se estaría cargando la muestra hacia el centro de

su distribución. En estos casos, se puede utilizar algún modelo predictivo utilizando como variable objetivo el dato perdido.

- Transformación de Datos, fase en la cual se busca agregar valor a ella, moldeándola de manera conveniente para su posterior análisis. Aquí, se calculan nuevas variables a partir del set de datos original, las cuales puedan ayudar a comprender de mejor manera el fenómeno estudiado (por ejemplo, calcular la edad del individuo a partir de su fecha de nacimiento). Además, incluye la normalización de datos para poder hacer que las variables sean comparables en términos de distancia, así como la categorización de atributos no numéricos.

- Minería de Datos, donde se realiza la implementación de modelos para extraer conocimiento de la base de datos, buscando los que mejor expliquen el comportamiento y tengan un mayor poder predictivo. Estos modelos pueden ser supervisados o no supervisados, es decir, utilizan una variable dependiente (clase) o no. Se abordará con mayor detalle la metodología en las siguientes secciones 2.4.2. y 2.4.3.

- Interpretación, donde se busca explicaciones lógicas de los patrones descubiertos que permitan explicar el fenómeno. De esta manera, se pueden generar propuestas de mejora para la actual toma de decisiones en la organización

3.2. Aprendizaje supervisado

Los modelos de aprendizaje supervisado son los que utiliza una variable dependiente o clase. El objetivo principal de estos modelos es predecir, en base a los atributos independientes, la variable clase.

Para realizar estos métodos, generalmente se particiona la base de datos en tres, generando un set de entrenamiento del modelo, un set de validación y un set de testeo, con el fin de evitar sobreajuste del modelo a los datos. Esto último quiere decir que el modelo desarrollado se ajuste de tal manera a los datos utilizados, que al momento de replicarlo con las mismas variables para otro set de datos, el modelo no sea capaz de predecir de buena manera, perdiendo generalidad. Dada la naturaleza de los proyectos de Minería de Datos, los cuales están asociados a grandes volúmenes de datos (Big Data) es que cobra sentido realizar esta triple partición. Si bien, el concepto de Big Data es algo ambiguo y va a depender de la empresa que lo enfrente, se consideran volúmenes de datos del orden de Petabytes hacia arriba dentro de esta categoría. En el caso de este estudio, el cual se lleva a cabo con los datos de operarios de una faena de la mediana minería, se optó solo por utilizar una partición de entrenamiento y validación, debido a la poca cantidad de datos, utilizando un 70% y 30% de la base de datos respectivamente.

Se busca entonces probar con distintos modelos para poder encontrar el que mejor ajuste tenga a los datos, comparando sus medidas de rendimiento. Estas se calculan generando una matriz de confusión comparando los resultados que genera el modelo con el set de validación que se reservó antes de comenzar. Una matriz clásica de clasificación binaria tiene la siguiente estructura:

		Valores predichos	
		1	0
Valores Reales	1	tp	fp
	0	fn	tn

Tabla 3 - Matriz de confusión en clasificación binaria

Donde los True Positives (tp) son los casos donde el modelo predijo de manera correcta a quienes renuncian voluntariamente. Los otros casos de la diagonal, True Negatives (tn) se refiere a los que el modelo clasificó correctamente como no – renuncia voluntaria. El resto de los valores son False Negatives (fn) que se refiere a elementos mal clasificados como negativos y False Positives (fp), que son elementos mal clasificados como positivos.

A partir de esto se pueden obtener varias medidas de rendimiento (como *accuracy*, *recall* o *precision*) de los modelos que servirán para compararlos y hacer una buena elección para su futura implementación.

$$Accuracy = \frac{tp + tn}{tp + tn + fn + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$Precision = \frac{tp}{tp + fp}$$

Ecuación 2 - Medidas de Rendimiento: Clasificación Binaria

Una forma de combinar estos dos de estos indicadores es a través del *F Score* [23], el cual alcanza su mejor valor en 1, su peor valor en 0 y se define de la siguiente manera:

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

Ecuación 3 - F Score

Donde el sub índice 1 indica que se trata del llamado *F Score* balanceado, el cual pondera de igual manera ambos indicadores.

Debido a que el problema a resolver para este caso tiene tres clases y no dos como la mayoría de los casos similares estudiados, se debe adoptar un enfoque que pueda capturar este comportamiento. En la literatura se distinguen 4 distintos tipos de problemas de aprendizaje supervisado [24], los cuales se enumeran a continuación:

- Binarios (Binary): problema clásico donde existen dos clases disjuntas (C_1, C_2), donde cada objeto solo puede pertenecer a una clase. Un ejemplo sería el caso de detección de fraude, donde existe la clase "*Fraude*" y "*No Fraude*".

- Multi-Clase (Multi-class): cada objeto puede pertenecer solo a una de las n clases $(C_1, C_2, \dots, C_n)$ existentes, las cuales son disjuntas, donde $n \in \mathbb{N}$, $\{n: 3,4,5 \dots\}$. Por ejemplo, el caso de estudio de este trabajo puede ser categorizado en *"Permanencia en la empresa"*, *"Despido"* o *"Renuncia Voluntaria"*.

- Multi-Etiqueta (Multi-labelled): cada objeto puede pertenecer a varias clases de un conjunto disjunto de clases $(C_1, C_2, \dots, C_n)$. Un caso de ejemplo podría ser la clasificación de los temas que se abordan en un documento, los cuales pueden ser diversos en un solo caso.

- Jerárquicos (Hierarchical): existe una jerarquía predeterminada de clases en el problema de clasificación, la cual no puede variar durante el transcurso del problema. De esta forma, cada objeto de estudio puede pertenecer a sólo una clase C_j. Por ejemplo, un animal puede ser mamífero, ave o un pez. Dentro de cada una de estas clases se encuentran sub-clases, donde cada animal puede pertenecer a solo una.

Tal como se menciona anteriormente, el problema que se aborda es del tipo "Multi-clase", ya que cada operario puede pertenecer a solo una de las tres clases existentes. Por lo tanto la matriz de confusión para este caso tomaría la siguiente estructura:

		Valores Predichos		
		Permanencia	Despido	Renuncia
Valores Reales	Permanencia	n_{11}	n_{12}	n_{13}
	Despido	n_{21}	n_{22}	n_{23}
	Renuncia	n_{31}	n_{32}	n_{33}

Tabla 4 - Ejemplo de Matriz de Confusión

Un enfoque para medir rendimiento en estos casos es considerar la "Tasa de éxito promedio" (Overall Success Rate) [23] que se define de la siguiente forma:

$$OSR = \frac{1}{n}\sum_{i=1}^{k} n_{i,i}$$

Ecuación 4 - OSR o Tasa de éxito promedio

Además de lo anterior, para este tipo de problemas existen generalizaciones del caso binario para las métricas vistas anteriormente [24]. Se consideran las métricas anteriormente mencionadas para el caso binario, pero se calculan para cada clase i, para luego calcular el promedio, donde I corresponde al número de clases.

$$Accuracy = \frac{\sum_i^I \frac{fp_i + fn_i}{tp_i + fn_i + fp_i + tn_i}}{I}$$

$$Recall = \frac{\sum_i^I \frac{tp_i}{tp_i + fn_i}}{I}$$

$$Precision = \frac{\sum_i^I \frac{tp_i}{tp_i + fp_i}}{I}$$

Ecuación 5 - Medidas de Rendimiento: Multi clasificación

Se observa que en el numerador se tiene la misma expresión del caso binario, sin embargo, esta se calcula para cada clase. Por esta razón, dependerá de cada clase qué valores de la tabla 4 vista anteriormente corresponden a *tp*, *fp*, *tn* y *fn*. A modo de ejemplo, para el caso de la clase "Permanencia", estas deberían ser las coordenadas consideradas para calcular sus métricas de rendimiento:

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="3">Valores Predichos</td></tr>
<tr><td>Permanencia</td><td>Despido</td><td>Renuncia</td></tr>
<tr><td rowspan="3">Valores Reales</td><td>Permanencia</td><td>tp</td><td colspan="2">fp</td></tr>
<tr><td>Despido</td><td rowspan="2">fn</td><td colspan="2" rowspan="2">tn</td></tr>
<tr><td>Renuncia</td></tr>
</table>

Tabla 5 - Matriz de Confusión: Ejemplo Clase Permanencia

Con este enfoque, se pueden rescatar métricas para cada una de las clases estudiadas y priorizar la clase de interés, en este caso, la renuncia voluntaria.

Para el modelamiento, se consideraron modelos paramétricos y no paramétricos. Básicamente, un modelo paramétrico busca estimar un set de parámetros $(\sigma_1, \sigma_2, \dots, \sigma_n)$ que definen el modelo. En cambio en un modelo no paramétrico, no solo estos parámetros definen el modelo, sino que además lo definen los datos de entrenamiento. Es necesario hacer la distinción de que no significa que un modelo no-paramétrico no calcule parámetros, si no que los parámetros están determinados por los datos de entrenamiento y no por el modelo en sí. Por ejemplo, el caso de una regresión lineal del tipo $y_i = \beta_0 + \beta_1 x_i$ se trata de un modelo paramétrico que está determinado por los parámetros β_0 y β_1. Para encontrar un nuevo valor para y_i, solo bastará con utilizar ambos parámetros calculados. En el caso de un modelo no paramétrico, este sería del tipo $y_i = f(x_i) + \varepsilon_i$ donde la función $f(\cdot)$ puede ser cualquiera, la cual estará determinada por lo datos. Por lo tanto, para estimar un nuevo valor en un modelo no paramétrico, además de conocer los parámetros del modelo (conocidos como híper parámetros), se utiliza la data anterior.

Junto a lo anterior, es necesario mencionar que los modelos utilizados son todos compatibles con el problema de multi clasificación. Por ejemplo, en el estudio realizado sobre empleados que prestan servicios a la minería [19] se utilizó el método Support Vector Machine, sin embargo, este modelo se complejiza cuando se quiere utilizar con una clase que no es binaria [25].

A continuación se presentan el marco teórico general de los modelos utilizados para el desarrollo del presente trabajo.

3.2.1. Modelos Paramétricos.

- Redes Neuronales [26]: este modelo pretende simular el comportamiento, de manera elemental, de las neuronas del cerebro humano. Basa su funcionamiento en estructuras simples llamadas "perceptrones" o "neuronas", las cuales se encuentran conectadas entre sí. Estas conexiones tendrán asignado un número o "peso" los cuales se ajustarán mediante las iteraciones del modelo, simulando un proceso de aprendizaje, los cuales se denotan como w_{ij}. Así, algunas conexiones tendrán pesos mayores que otros, dando un sentido de mayor relevancia a esas conexiones frente al resto, tal como es el caso de las conexiones del cerebro humano que toman mayor importancia cuando el individuo se encuentra realizando alguna tarea que requiera el funcionamiento de alguna zona del cerebro.

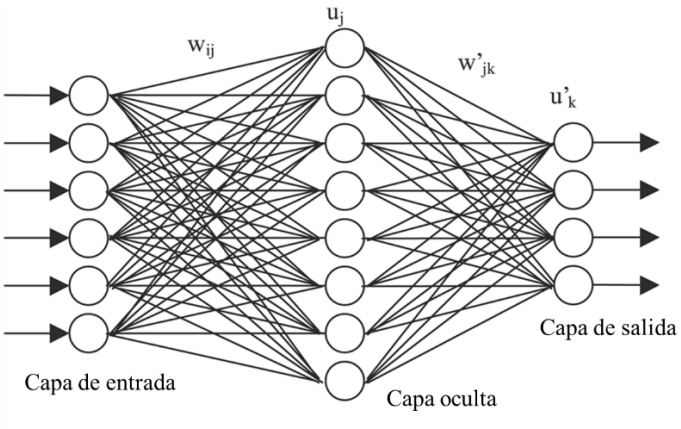


Ilustración 25 - Red neuronal de dos capas

Generalmente se hace alusión a este modelo como de "caja negra", ya que la estimación de estos pesos es un proceso bastante complejo, llamado "Backpropagation". Cuando se cuenta con estructuras de más de dos capas, se dice que se envían las señales "hacia adelante" y los errores se propagan "hacia atrás" (backpropagation). El modelo comienza su proceso de entrenamiento en base a pesos w_{ij} aleatorios y el objetivo es ajustarlos para minimizar el error. Este algoritmo comienza con una "activación" de cada perceptrón, la cual es una suma ponderada de los datos de entrada x_i y el peso w_{ij}.

$$A_j(x, w) = \sum_{i=0}^{n} x_i w_{ij}$$

Ecuación 6 - Función de Activación de Perceptrón

Con esta función de activación, se define una función para los perceptrones de la capa de salida u "outputs".

$$O_j(x, w) = \frac{1}{1 + e^{A_j(x,w)}}$$

Ecuación 7 - Función de salida en redes neuronales

La idea entonces será obtener un resultado esperado cuando un cierto set de datos entra a la red. Dado que el error es la diferencia entre un resultado esperado y el resultado obtenido, el error dependerá de los pesos asignados, por lo que se necesita ajustarlos durante el proceso de aprendizaje o entrenamiento. Típicamente se utiliza el error cuadrático $E_j = \left(O_j(x, w) - d_j\right)^2$. Finalmente, se ajustan los pesos mediante el método de gradiente:

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

Ecuación 8 - Cálculo de pesos en redes neuronales

La intuición detrás de esto es que el ajuste de cada uno de los pesos será una constante negativa η multiplicada por la dependencia del peso anterior en el error E de la red, el cual es la derivada del error con respecto al peso. De esta forma, si el peso aporta mucho al error de la red, el ajuste será más significativo.

- Logit Multinomial [27]: es una generalización del caso clásico del modelo logit, que permite predecir una elección binaria utilizando un set de variables explicativas. Este modelo se basa en la maximización de utilidad u_{ij} de cada individuo i cuando realiza la elección de una de las opciones discretas j que tiene para escoger, la cual se compone de una componente determinística v_{ij} y otra estocástica ε_{ij}. A partir de esto, se modela la probabilidad de elección para cada opción j de la siguiente manera.

$$\mathbb{P}_{ij} = \frac{e^{v_j}}{\sum_{j=1} e^{v_{ij}}}$$

Ecuación 9 - Probabilidad de elección en MLogit

Donde se debe cumplir que $0 < \mathbb{P}_{ij} < 1 \; \forall i, j$ y $\sum_{j=1} \mathbb{P}_{ij} = 1$, que son las propiedades básicas de la probabilidad. De esta manera, se puede calcular la probabilidad de cada uno de los individuos de elegir una de las opciones disponibles que en este caso son permanencia en la empresa, despido o renuncia.

3.2.2. Modelos No paramétricos.

- Árboles de Decisión: en términos generales, sirven para generar clasificación en base a "decisiones" para cada atributo, formando un diagrama con forma de árbol invertido. Generan información del tipo "si el empleado tiene sobre 50 años y vive a más de 300 km. de la faena, entonces dejará la empresa". Además, se puede observar de manera visual, lo cual entrega una interpretación sencilla. Existen diversos algoritmos para generarlo. A continuación, se explicará el algoritmo ID3, el cual viene implementado en diversas herramientas computacionales.

Se comienza con un todas las observaciones en un mismo set.

1. Se busca el mejor atributo para generar la separación.
2. Eliminar ese atributo del set de atributos disponibles a usar.
3. Separar en base a ese atributo las distintas clases.
4. Si todas las observaciones son de la misma clase en un nodo, detenerse.
5. Volver a 2.

Para elegir el mejor atributo para realizar la separación, se busca el que genere la menor medida de información (IM) en el nodo resultante mediante la ecuación donde:

- $IM = \sum_{i=1}^{m} p_i * E_2(K_i)$
- $E_2(K) = -p^+ * \log_2(p^+) - p^- * \log_2(p^-)$
- K: Nodo (hoja)
- p^+, p^-: frecuencia relativa de postivos/negativos en una hoja
- $p^+ + p^- = 1$
- $E_2(K) \geq 0$
- $E_2(K) = 0 \Leftrightarrow p^+ = 0 \ \vee p^-$
- $\max\big(E_2(K)\big) \Leftrightarrow p^+ = p^-$

Ecuación 10 - Definiciones Árbol de Decisión

Este proceso puede dejar algunas variables fuera del modelo, utilizando solo las variables que separen mejor las observaciones entre las clases. Esto se conoce como "prunning" o "poda" del árbol.

- Random Forest [28]: perteneciente a la familia de algoritmos de *boosting*, basa su funcionamiento en la creación de una serie de modelos "débiles" o poco complejos para crear un modelo más robusto. En este caso, se crean múltiples arboles de decisión, utilizando sub muestras de la base de entrenamiento, concepto conocido como *Bagging*.
  Se generan entonces árboles utilizando una muestra de $X = x_1, x_2, \dots x_n$, los vectores de observaciones del training set e $Y = y_1, y_2, \dots, y_n$ la clase de cada vector.
  - Para iteraciones $b = 1, \dots, B$
  - Muestrear con reposición n ejemplos de X, Y llamandolos X_b, Y_b , donde X_b además contiene solo algunos de los atributos del total.
  - Entrene un árbol de decisión f_b en X_b, Y_b sin poda o *prunning*.
  - Para desplegar estos múltiples árboles como un solo modelo, se considera que cada árbol generado tiene el mismo peso para el modelo resultante, donde se considera para la observación y_i cuantos árboles predijeron si ésta es positiva o negativa, asignando la clase que la mayoría de los árboles predicen. Para extraer conocimiento de este modelo, se hace complejo observar todas las reglas generadas por los distintos árboles creados, por lo tanto se utiliza el enfoque de rankear variables por su importancia en el modelo predictivo. Una

de ellas es observar qué tanto aporta una variable al Accuracy del modelo, llamado *Mean Decrease Accuracy*. Así, se permutan al azar los valores de una variable a lo largo de la base de datos, midiendo el impacto en el rendimiento final del modelo. Cuando el impacto es mayor, quiere decir que la variable es más relevante para el modelo. Otra forma de ver esto es a través del índice de Gini, midiendo qué tan desigual es un nodo hijo al separar por una variable.

- Gradiente Incremental (Gradient Boosting) [29]: al igual que el modelo anterior, este método utiliza el concepto de *boosting*. La idea principal de este método es construir estos modelos "débiles" para estar correlacionados con el gradiente negativo de una función de pérdida o error $L(\cdot)$ del modelo final. Este algoritmo puede ser utilizado con modelos débiles $F(\cdot)$ y funciones de pérdida $L(\cdot)$ a elección del analista.

Por lo tanto, se busca encontrar una relación funcional de la forma $F^*(x) = y$ tal que sea capaz de describir el problema estudiado. Para esto, se crea a partir de la adición de modelos débiles $h(\cdot)$ que minimice la función el valor esperado de la función de pérdida $F^* = \arg\min \mathbb{E}_{x,y}[L(y, F(x)]$. Estimando híper parámetros γ se tiene que la función que describa el caso estudiado será la siguiente suma ponderada.

$$F^*(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + const.$$

Ecuación 11 - Modelo General de Gradiente Incremental

Debido a que el problema de optimización que conlleva esto podría llegar a ser muy complejo, es que este método propone realizar un método *greedy* (en español, avaro). En otras palabras, realizar una heurística (algoritmo que siempre converge pero no siempre encontrará el óptimo) que en cada paso del algoritmo busque un óptimo local para proponer una solución final cercana al óptimo. El algoritmo, de manera genérica, procede de la siguiente manera:

- Inicializar la función en una constante $F_0(x) = cte.$
- En el primer paso, buscar el óptimo local $F_1(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$
- Iterar M veces sobre:$F_m(x) = F_{m-1}(x) + \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + f(x_i))$

No obstante, el problema de optimización para funciones de pérdida más complejas, como es el caso de estimación de multi clases, puede ser bastante complejo computacionalmente de resolver. En consecuencia, se propone la idea de calcular el gradiente negativo (la dirección de máximo decrecimiento de la función) en un punto para encontrar un mínimo local, el cual

corresponde a la derivada parcial de la función de pérdida que se desea minimizar con respecto al modelo débil. Así, los híper parámetros γ para cada iteración pueden ser calculados de la siguiente manera.

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{N} L\left(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial h(x_i)}\right)$$

Ecuación 12 - Estimación de parámetro de ajuste en Gradiente Incremental

En este caso particular, donde se presenta un problema de multi clase, se utiliza la función propuesta para el caso de multi clasificación [29]:

$$L\left(\{y_k, F_k(x)\}_1^k\right) = -\sum_{k=1}^{K} y_k \log p_k(x)$$

Ecuación 13 - Función de pérdida para multi clasificación

Donde el índice k hace referencia a cada clase $k = \{1, \dots, K\}$. Lo siguientes parámetros quedan definidos como $y_k = 1(clase = k)\, \epsilon\, \{0,1\}$ un indicador para clase, $p_k = \mathbb{P}(y_k = 1|\, x\,)$ la probabilidad de que una observación pertenezca a la clase k dado un set de parámetros x y F el modelo "débil" sobre el cual se construirá el modelo final.

Típicamente para este caso se utilizan como modelos débiles árboles de decisión, sobre los cuales se construye el modelo final. El cálculo de los gradientes y las probabilidades expuestas a continuación se pueden encontrar en el documento introductorio al modelo [29]. El algoritmo entonces para el caso multi clase es el siguiente.

Entrada:

- Datos de entrada $(x, y)_{i=1}^{N}$
- Número de iteraciones M
- Función de pérdida $L(\cdot)$
- Modelo débil $F(\cdot)$

Algoritmo:

i. Inicializar $F(\cdot)$ como una constante
ii. Desde $t = 1$ hasta M desarrollar:
    a. Calcular la probabilidad de pertenencia a cada clase de cada observación $p_k(x) = \frac{\exp(F_k(x))}{\sum_i \exp(F_i(x))}, k = \{1, \dots, k\}$
    b. Desde $k = 1$ hasta K desarrollar:
        i. $\widehat{y_{ik}} = y_{ik} - p_k(x_i), i = \{1, \dots, N\}$
        ii. $\{R_{jkm}\}_{j=1}^{J}$ = J-nodo terminal del Árbol
        iii. $\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{x_i \epsilon R_{jkm}} \widehat{y_{ik}}}{\sum_{x_i \epsilon R_{jkm}} |\widehat{y_{ik}}|(1-|\widehat{y_{ik}}|)}, j = 1, \dots, J$
        iv. $F_{km}(x) = F_{k,m-1}(x) + \sum_{j=1}^{J} \gamma_{jkm}\, 1(x \in R_{jkm})$[4]
        Terminar.
    Terminar.

Fin del Algoritmo.

3.2.3. Aprendizaje no supervisado

El aprendizaje no supervisado es aquel que no utiliza la variable clase, utilizando solo las variables independientes. Esto servirá para poder caracterizar a los empleados fugados, separándolos del resto de las observaciones, y poder generar segmentos que sirvan para poder comprender el comportamiento de mejor manera.

El objetivo entonces del aprendizaje no supervisado para este caso será la creación de grupos de observaciones que sean semejantes entre sí y que al mismo tiempo, estos grupos sean distintos entre sí. Los algoritmos de aprendizaje no supervisado buscan patrones dentro de la estructura natural de los datos, generando grupos basados en la distancia entre ellos. Las medidas de distancia a utilizar varían dependiendo de la aplicación que se busca.

En este caso particular, se utilizan la distancia euclidiana entre puntos para ver qué observaciones se asemejan entre sí. Uno de los métodos más utilizados es k-medias [30], el cual crea clusters rígidos, es decir, que un elemento sólo puede pertenecer a un cluster. Cada uno de estos cluster estará determinado por un centroide m_k, el cual será un punto ficticio ubicado en el centro del cluster generado. Cada elemento x entonces pertenecerá al cluster k si se encuentra cerano al centroide m_k.

[4] $1(x \in R_{jkm})$ se refiere a la función indicatriz, la cual toma el valor 1 cuando la condición se cumple que la observación pertenece a la hoja R_{jkm} y 0 si no se cumple.

El número de segmentos k debe ser elegido a priori. Si bien no existe una manera correcta de elegir el número de segmentos, este debe ser tal que permita al analista interpretar los resultados pero que a la vez minimice el error. La manera típica de encontrar el número k de clusters es realizar un proceso iterativo desde $k = 2$ en adelante y graficar la suma de errores cuadráticos dentro de los clusters.

$$SSE = \sum_{k=1} \sum_{x \epsilon C_k} (x - m_k)^2$$

Ecuación 14 - Suma de errores cuadráticos: clustering

Luego, se puede generar la curva de errores cuadráticos y buscar un "codo" o punto de inflexión. Este punto representa un punto donde aumentar el número de clusters no aporta en términos de disminución del error, por lo tanto, no vale la pena seguir aumentando el número de clusters.

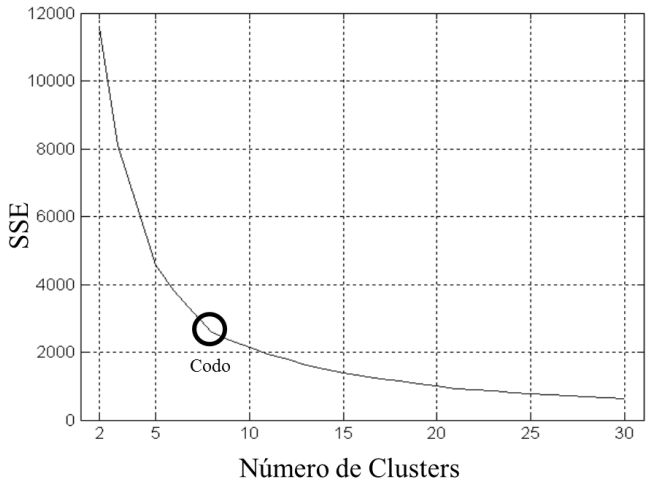


Ilustración 26 - Regla del Codo

Para iniciar el algoritmo, se necesita un número k de segmentos, y un conjunto $m = \{m_1(0), m_2(0), \dots, m_k(0)\}$ de centroides para cada segmento, los cuales se eligen arbitrariamente para la instancia $t = 0$. Luego, se asigna a la observación X_i el segmento del cual esté más cerca del centroide, lo cual se repite para todo i. Después de esto, se recalculan los centroides dados los nuevos puntos pertenecientes al conjunto y se crea el conjunto $Z = \{m_1(1), m_2(1), \dots, m_k(1)\}$. El criterio de parada se cumple cuando, para un ε dado, se cumple:

$$m_i(t) - m_i(t-1) < \varepsilon \,, \forall i$$

Ecuación 15 - Criterio de parada de clustering

Uno de los objetivos de este estudio es caracterizar el comportamiento de los empleados que ya renunciaron voluntariamente para poder observar distintas características en ese grupo de la población a estudiar. El método de k-medias permite diferenciar entonces distintos segmentos de empleados fugados, para los cuales se pueden proponer distintas medidas de retención de acuerdo a sus características.

Además de cumplir con funciones de agrupamiento de datos, el aprendizaje no supervisado además puede cumplir la función de detección de valores fuera de rango u *outliers* [31], los cuales son puntos dentro de la base de datos a estudiar que contienen valores atípicos. Ellos dificultarán la implementación de modelos, por lo tanto, se requiere modificar estos puntos para que no afecten el análisis. Para realizar esto, basta con encontrar configuraciones de número de clusters que minimicen el error y que posean clusters con un bajo número de elementos. A esto se le llama en la literatura "external outlier", que en un caso bidimensional sería similar a la ilustración 27 a continuación. Si al eliminar estos puntos, se realiza nuevamente el algoritmo de k-medias y los mismos grupos anteriores son significativos, se detiene el procedimiento de eliminación de outliers.

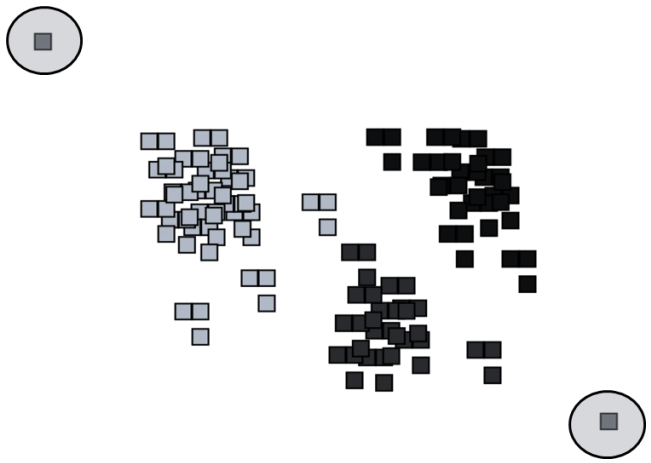


Ilustración 27- Detección de Outliers: Ejemplo bidimensional

4. Desarrollo Metodológico

Este capítulo abarca los pasos vistos de la metodología KDD, desde el proceso de limpieza de datos, selección de variables y la implementación de modelos mostrando sus correspondientes errores en la base de entrenamiento y prueba. Con esto se da paso a la selección del modelo de aprendizaje supervisado que se utilizó para describir el comportamiento y los resultados encontrados. Finalmente, se da paso a mostrar los resultados de la clusterización propuesta para empleados renunciados, explicando las características que describen a cada uno de los segmentos encontrados.

4.1. Preparación de la base

En primera instancia, se debe crear una base de datos para análisis. En este caso, se utilizaron diversas fuentes de datos internas de la empresa para crear una base de datos que tuviera para cada ID's de la persona las variables expuestas. Para asegurar la confidencialidad de los datos, las ID's utilizadas fueron enmascaradas y los nombres de las personas fueron eliminados del análisis, siendo solo utilizadas para generar uniones consistentes entre bases de datos y poder obtener comportamiento individual de las personas.

Para el comienzo del proyecto fue necesario entonces unificar los datos de las diversas fuentes de la empresa. A continuación se presenta un diagrama de la fase inicial del proyecto, llevado a cabo en el software *SAS Enterprise Guide*, a excepción de la eliminación de outliers e imputación de datos faltantes, que se desarrolló utilizando *SAS Enterprise Miner*.

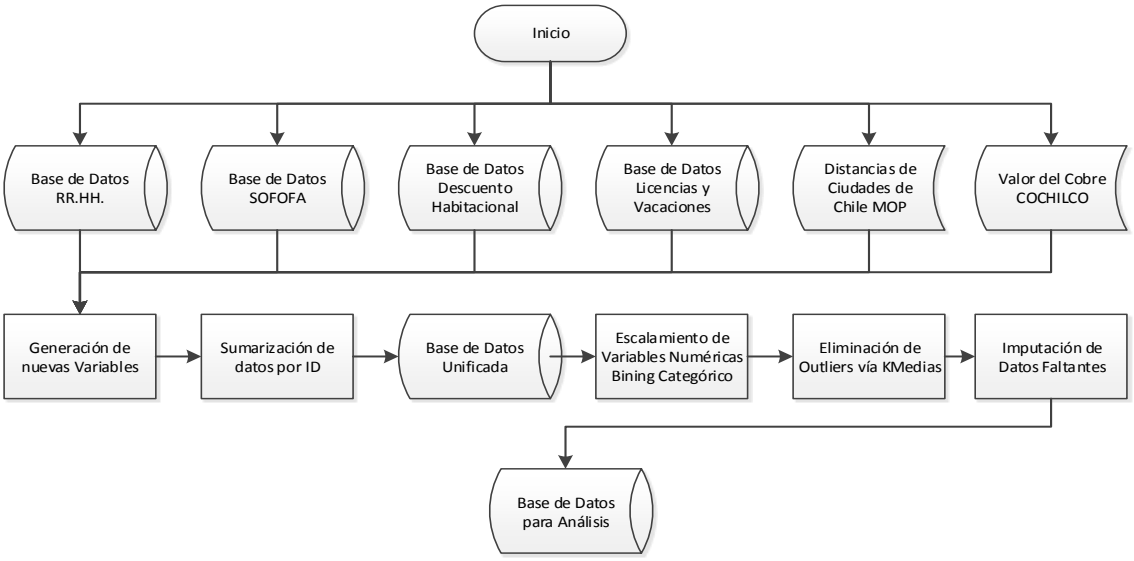


Ilustración 28 - Diagrama: Inicio del proyecto

En la mayoría de los casos, los datos registrados por la empresa (a excepción de los datos demográficos) se encontraban basados en *timestamps*, lo que quiere decir que se registra en una fecha determinada cierto acontecimiento acerca un empleado de la faena. Por lo tanto, se realizó una sumarización vía ID de la persona para poder capturar el comportamiento histórico e individual de cada trabajador de la empresa, como se ejemplifica en la siguiente imagen.

	ID	unidad	cod_sin	ames	descitm
1	12	203	3	200401	Descto. Ptmo. Habitacional Ct.:21 de 60
2	12	203	3	200402	Descto. Ptmo. Habitacional Ct.:22 de 60
3	12	203	3	200403	Descto. Ptmo. Habitacional Ct.:23 de 60
4	12	203	3	200404	Descto. Ptmo. Habitacional Ct.:24 de 60
5	12	203	3	200405	Descto. Ptmo. Habitacional Ct.:25 de 60

	ID	unidad	cod_sin	cuota_min	cuota_max	cuotas_pagadas	pago_completo
1	87	202	5	0.01667	0.8333333333	49	0
2	449	203	1	0.01667	1	59	1
3	165	202	5	0.16667	1	50	1
4	181	0	3	0.95000	1	3	1
5	441	205	5	0.05000	1	57	1

Ilustración 29 - Sumarización por ID

Sobre el cálculo de nuevas variables, en su mayoría corresponden a operaciones algebraicas básicas entre las variables disponibles entre las distintas fuentes de datos. Las variables generadas se encuentran detalladas en la sección 2.1. de datos disponibles.

Con la nueva base de datos unificada se hace necesario eliminar los valores fuera de rango para evitar problemas posteriores con el ajuste de modelos, los cuales podrían llevar a soluciones de bajo poder predictivo. Se utilizó el método de k medias explicado en la sección 3.2.3. de modelos no paramétricos. Dado que este modelo basa su funcionamiento en el cálculo de distancias, las variables que debe recibir como entrada deben ser todas numéricas, por lo que las variables categóricas deben ser transformadas a numérica mediante "binarización". Esto quiere decir que para cada categoría de la variable se crea una nueva variable que indica 1 si pertenece a esa categoría y 0 si no. Además de esto, dado que entre variables las distancias no son comparables entre sí (por ejemplo, el caso de la variación del precio del cobre que está en orden porcentual versus el dinero invertido en capacitación por la empresa, que está en el orden de los cientos de miles) se debe realizar un escalamiento de variables numéricas para que las distancias tengan sentido. Luego de esto, es posible desarrollar el procedimiento de k-medias para detectar los elementos fuera de rango, los cuales al representar menos del 1% de la base de datos fueron eliminados.

La naturaleza del proyecto, el cual contempla tanto aprendizaje supervisado como no supervisado, requiere para el primero una variable objetivo o target para la cual el modelo pueda generar una predicción. En el registro de egreso de una persona en la empresa se encuentra si ésta fue despedida (para lo cual existen diversos motivos) o renunció voluntariamente. A partir de esto, se genera la siguiente variable categórica multi clase:

$$y_i = \begin{cases} 0 & \text{Si el empleado permanece en la empresa} \\ 1 & \text{Si el empleado es despedido de la empresa} \\ 2 & \text{Si el empleado renuncia voluntariamente} \end{cases}$$

Ecuación 16 - Variable categórica objetivo para el problema

Para comenzar con la implementación de modelos, se imputaron los datos faltantes mediante la técnica de árboles de decisión, los cuales pertenecían principalmente a datos correspondientes a las capacitaciones, los cuales no tenían una cantidad de datos perdidos mayor a un 25% en ninguna de las variables. En el caso de los datos correspondientes a beneficios habitacionales (pago de cuotas de beneficio habitacional, pago completo del beneficio o cuotas pactadas), no fueron imputados y fueron descartados de la base de datos debido a que la ausencia de estos datos superaba un 75% de la base de datos. El procedimiento de imputación mediante árboles de decisión utiliza como variable objetivo el valor perdido, de la misma manera explicada en la sección 3.2.2. de modelos no paramétricos.

4.2. Implementación de Modelos

Con la consolidación de una base de datos unificada e individualizada para el análisis, es entonces cuando se pueden desarrollar los modelos de minería de datos planteados. Se experimentó con diversos modelos siguiendo el orden de la ilustración 30, mediante la utilización del software *SAS Enterprise Miner*.

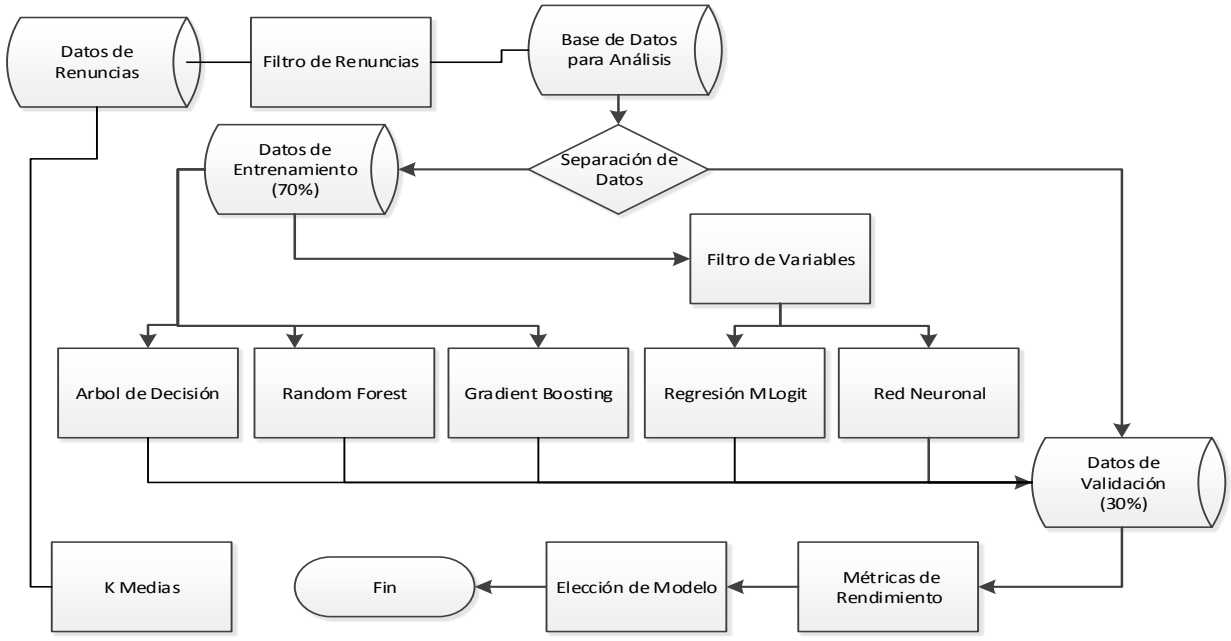


Ilustración 30 - Diagrama: Implementación de Modelos

Es necesario agregar que tanto el modelo Logit Multinomial como Redes Neuronales fueron sometidos a un filtro de variables. En el caso de ambos, agregar muchas variables al problema puede generar sobreajuste de los datos y pérdida de la interpretabilidad de los resultados. Tanto los árboles de decisión, Random forest y gradient boosting no tienen problemas con esto. Los dos primeros, se conocen como "métodos embebidos", lo que quiere decir que el filtro de variables se realiza dentro del mismo algoritmo. En el caso del árbol de decisión, esto se conoce como poda o *prunning* y en el caso del random forest se conoce como *bagging*, ambos conceptos explicados en la sección 3.2.2. de modelos no paramétricos. Para el caso del gradient boosting, debido a que se basa en la suma de modelos de árboles de decisión, tiende a ser robusto frente a la adición de variables irrelevantes [29].

En adición a lo anterior, para realizar perfiles de personas que renunciaron voluntariamente, se separaron de la base de datos inicial a las personas que presentaran esta característica para realizar análisis de cluster mediante el método de k medias.

4.2.1. Filtro de Variables: Test de Kolmogorov Smirnov

Los modelos de minería de datos buscan que las distribuciones acumuladas de los datos no se encuentren perfectamente correlacionadas para cada una de las clases a estudiar, puesto que esto dificulta la detección de patrones para poder predecir comportamiento. Por lo tanto, se aplicó el test de Kolmogorov-Smirnov (KS) para determinar si existen diferencias significativas entre las distribuciones de quienes renuncian voluntariamente a la empresa y quienes no lo hacen. Dado que

los esfuerzos de este trabajo de memoria se encuentran enfocados en encontrar patrones a la renuncia voluntaria, es que se define la siguiente variable objetivo para realizar el test.

$$y_i = \begin{cases} 1 \; Si \; renunció \; a \; la \; empresa \\ 0 \qquad\qquad\quad \sim \end{cases}$$

Ecuación 17 - Variable Objetivo para Test KS

El test KS es una prueba no paramétrica que básicamente busca la mayor distancia entre las distribuciones acumuladas de los datos de ambas clases (y_i igual a 1 o 0), y evalúa si la distancia D es significativa estadísticamente, como en el siguiente diagrama.

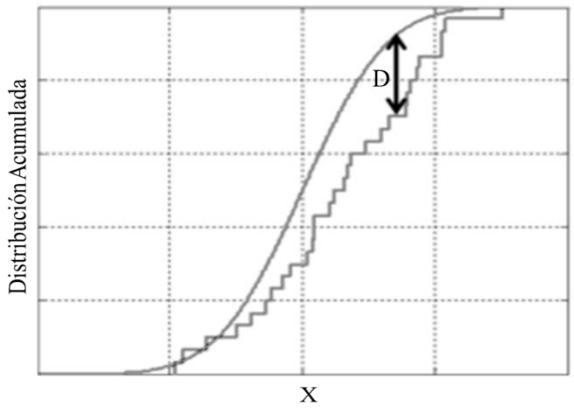


Ilustración 31 - Test KS

La hipótesis nula del test es que las dos muestras dibujan la misma distribución. Para realizar esto, se busca la mayor distancia entre el valor de la función para ambos grupos [32]. En este caso, F_1 y F_2 representan la función descrita para cada población en una misma variable y "sup" representa el supremo o valor máximo obtenido.

$$D_{n,n'} = \sup_x \left| F_{1,n}(x) - F_{2,n'}(x) \right|$$

Ecuación 18 - Distancia test KS

El p-valor se puede interpretar como la probabilidad de equivocarse al rechazar la hipótesis nula. Para el caso de este estudio, se consideró que si el p-valor es menor de 0.05, se rechaza la hipótesis nula y se acepta la hipótesis alternativa: las distribuciones son distintas.

Los valores coloreados corresponden a variables en las cuales se tienen diferencias significativas entre las distribuciones de ambas clases (renuncia voluntaria y no renuncia voluntaria). Se puede apreciar que, dado el análisis exploratorio mostrado en la sección 2.1. "Datos Disponibles", las variables que mostraban alta tasa de fuga en grupos suficientemente grandes de la población son los que presentan un p-valor menor a 0,05, por lo que se puede decir de ellos al 95% de confianza que sus distribuciones son distintas, y por lo tanto, servirán a los modelos para discriminar comportamiento. Esto se tiene para el caso de:

Test Kolmogorov-Smirnov	
Variable	p-valor
edad_empresa	0,006425000000000
tiempo_empresa	0,000143600000000
demora_contrato_indef	0,061980000000000
meses_ultima_capacitacion	0,000000000000011
total_invertido_capacitacion	0,000000016760000
max_nivel_escolaridad	0,933800000000000
periodo_prueba	0,215500000000000
reingreso_empresa	1,000000000000000
licencia_profesional	0,072840000000000
maximo_subnivel	0,071200000000000
beneficio_renuncia	0,047800000000000
min_nivel_escolaridad	0,785500000000000
sindicalizado	1,000000000000000
distancia_hogar	0,955300000000000
variacion_cobre_6meses	0,000000000000000
inversion_empresa_anual	0,486600000000000
indice_aprobacion_horas	0,000000000000000
capacitaciones_año	0,002448000000000
promedio_asistencia_capacitaciones	0,000140000000000
variacion_cobre_3meses	0,000000000000000
meses_desde_licencia	0,004302000000000
cantidad_licencia_por_falta	0,204900000000000
cantidad_licencia_por_Permiso	0,133200000000000
cantidad_licencia_por_licencia_medica	0,149100000000000
dias_licencia_anual	0,984300000000000

Tabla 6 - p-valor para test KS

Por lo tanto, las variables que ingresan a los modelos de MLogit y Redes Neuronales son:

- Edad en la empresa
- Tiempo en la empres
- Meses desde la última capacitación
- Total invertido por la empresa en capacitación
- Beneficio de renuncia

- Variación del precio del cobre en 6 meses
- Variación del precio del cobre en 3 meses
- Precio del cobre en el último mes
- Índice de aprobación de horas de capacitación
- Capacitaciones por año
- Promedio de asistencia a capacitaciones
- Meses desde licencia

4.2.2. Horizonte temporal del valor del precio del cobre.

En el punto anterior, acerca del filtro de variables, se hace notar una marcada tendencia del valor del cobre como una poderosa variable predictora de renuncias voluntarias dentro de la empresa. Los valores obtenidos para ambos p-valores son del orden de 10^{-16}, lo que muestra una muy baja probabilidad de equivocarse al rechazar la hipótesis nula del test, lo que quiere decir que existe evidencia estadística para afirmar que la diferencia entre ambas funciones (fugados y no fugados) es bastante distinta y que servirá como predictor para los modelos. Cada modelo fue implementado con solo una de estas dos variables ya que de lo contrario se estaría agregando información irrelevante para poder describir el problema de recursos humanos, además de contribuir al sobreajuste del modelo. Se muestran los errores de cada modelo utilizando la misma data, pero variando la temporalidad de variación de precio del cobre: 3 y 6 meses.

$$Error\ General = 1 - OSR$$
$$= 1 - \frac{1}{n}\sum_{i=1}^{k} n_{i,i}$$

Ecuación 19 - Error General de un modelo

	Tasa de Error General	
Modelo	Cobre 3 meses	Cobre 6 meses
Árbol de Decisión	29,7%	25,3%
Random Forest	36,8%	32,7%
Incremento de Gradiente	26,4%	21,8%
Logit Multinomial	44,3%	40,5%
Red Neuronal	38,4%	34,2%

Tabla 7 - Comparación Horizonte Temporal Cobre

Es claro ver que para cada uno de los modelos, el error disminuye cuando se utiliza un horizonte temporal de 6 meses. Si bien, en promedio es un 4% de mejora con respecto a todos los modelos utilizados, se optó por utilizar solo el horizonte temporal de 6 meses debido a que presenta una mejora para todos los modelos probados. Considerando además que el proceso de toma de

decisión de renuncia de una persona no es trivial, ya que en éste la persona estudia todas sus posibles opciones de trabajo, parece razonable considerar un horizonte temporal de esta magnitud versus un periodo de tres meses. Sin embargo, esta evidencia no permite asegurar la cantidad de meses de anticipación con la cual un operario comienza a pensar en renunciar a la empresa, pero si puede ser utilizada como una buena variable predictora.

4.3. Resultados obtenidos

El proyecto contempla, tal como se puede apreciar en los objetivos específicos, técnicas de aprendizaje supervisado y no supervisado para describir el problema de las renuncias voluntarias dentro de la compañía. En esta sección se presentan los resultados obtenidos utilizando estas técnicas.

4.3.1. Resultados aprendizaje supervisado

A continuación se muestran las medidas de rendimiento para cada uno de los modelos probados, los cuales como se menciona anteriormente, utilizaron un horizonte temporal de seis meses para la variación del precio del cobre, además del resto de las variables explicativas nombradas anteriormente.

		Permanencia	Despido	Renuncia	Promedio
Árbol de Decisión	Accuracy	91,1%	75,9%	82,3%	83,1%
	Error	8,9%	24,1%	17,7%	16,9%
	Recall	100,0%	29,6%	96,9%	75,5%
	Precision	74,1%	100,0%	70,5%	81,5%
	Fscore	85,1%	45,7%	81,6%	78,4%
Random Forest	Accuracy	80,0%	67,3%	87,3%	78,2%
	Error	20,0%	32,7%	12,7%	21,8%
	Recall	58,1%	51,3%	90,0%	66,4%
	Precision	66,7%	54,1%	78,3%	66,3%
	Fscore	62,1%	52,6%	83,7%	66,4%
Logit Multinomial	Accuracy	79,7%	64,6%	74,7%	73,0%
	Error	20,3%	35,4%	25,3%	27,0%
	Recall	50,0%	44,4%	78,1%	57,5%
	Precision	62,5%	48,0%	65,8%	58,8%
	Fscore	55,6%	46,2%	71,4%	58,1%

		Permanencia	Despido	Renuncia	Promedio
Red Neuronal	Accuracy	82,3%	72,2%	77,2%	77,2%
	Error	17,7%	27,8%	22,8%	22,8%
	Recall	55,0%	59,3%	78,1%	64,1%
	Precision	68,8%	59,3%	69,4%	65,8%
	Fscore	61,1%	59,3%	73,5%	65,0%
Gradient Boosting	Accuracy	90,6%	78,6%	88,0%	85,8%
	Error	9,4%	21,4%	12,0%	14,2%
	Recall	93,5%	50,0%	95,5%	79,7%
	Precision	76,3%	84,0%	77,8%	79,4%
	Fscore	84,1%	62,7%	85,7%	79,5%

Tabla 8 - Medidas de Rendimiento

Un resultado que se puede apreciar en esta etapa es que tanto las medidas de rendimiento para las clases de permanencia y renuncia son considerablemente más altas que para el caso los despidos. Se puede inferir que, dado que para los modelos utilizados es difícil discriminar la clase "despido" del resto, no existe un patrón demasiado marcado para estos casos.

Una manera de observar que el modelo no se encuentra sobre ajustado a los datos de entrenamiento es observar cómo se comporta la base de datos de prueba. De los resultados obtenidos se obtiene que para todos los modelos, el error cuadrático promedio es menor para el caso de la base de entrenamiento que en la base de prueba, lo cual es un resultado lógico si se considera que el modelo entrenado no utilizó en su construcción valores del set de testeo. No obstante, en ninguno de los modelos se observa una variación mayor al 5% para este valor.

Las medidas obtenidas para el modelo de Gradiente Incremental son superiores en *Accurac*y para todas las clases y por consecuencia, menor en error para cada una de las clases estudiadas. En adición a lo anterior, tanto *precision* como *recall* son superiores para cada una de las clases, lo que lleva también a un F Score más alto. Además, su variación respecto de la base de entrenamiento con la base testeo para el error cuadrático promedio es de un 3,9% lo que cual evidencia que el modelo no se encuentra sobre ajustado. El trazado del error cuadrático promedio para las iteraciones del modelo puede encontrarse en la sección de anexos. Por lo tanto, este modelo fue el escogido para proseguir con el estudio. En la sección de anexos se puede encontrar el detalle de las matrices de confusión que fueron obtenidas al probar estos modelos en los datos de prueba.

El modelo de Gradiente Incremental entrega como resultado tanto una probabilidad de pertenencia a cada clase estudiada, además de un ranking de variables que fueron relevantes a la hora de ajustar el modelo. Gracias a las probabilidades, se pueden desarrollar extensiones del modelo, como se podrá ver más adelante en el presente informe, además de entregar la probabilidad de pertenecer a la clase "renuncia" de los actuales trabajadores de la empresa. Esto se puede desarrollar de manera sencilla utilizando los procesos generados tanto para SAS Enterprise Guide

como SAS Enterprise Miner sobre datos de operarios actuales de la empresa. Con respecto a la importancia relativa de las variables, esta se propone como una manera de poder extraer conocimiento a través de los patrones encontrados con el modelo [29], lo cual es la finalidad de un proyecto de KDD.

Sea una variable j en uno de los árboles T generados por el modelo de Gradiente Incremental, el cual posee L separaciones o *splits* (nivel del árbol que se ramifica en más hojas). Para calcular la importancia de la variable j en el árbol T, se buscan todos los nodos no terminales (donde el algoritmo no cumple el criterio de parada), es decir, desde el nivel $L-1$ hacia arriba y se calcula lo siguiente:

$$Importancia_j(T) = \sum_{i=i}^{L-1} I_i^2 1(S_i = j)$$

Ecuación 20 - Importancia de una variable para un árbol de decisión

Donde I_i^2 es la mejora del error cuadrático como resultado del split en el nivel i y $1(S_i = j)$ es la función indicatriz que indica 1 si la variable j fue seleccionada para separar y 0 si no. De esta manera, la medida de importancia de una variable va a estar dada por el número de veces que fue utilizada para crear una separación, y como esta separación contribuyó a disminuir el error de clasificación en el árbol. Para obtener la importancia de la variable para el modelo de gradiente incremental final, se debe promediar la importancia de todos los árboles generados.

$$Importancia_j = \frac{1}{M}\sum_{i=1}^{M} Importancia_j(T_i)$$

Ecuación 21 - Importancia de una variable para Gradiente Incremental

Si bien esta medida entrega qué tan relevante es la variable para generar la búsqueda de patrones dentro de los datos de entrenamiento, no permite conocer el comportamiento de estas variables frente al problema. Es por esta razón que las variables deben ser analizadas para conocer su efecto en el problema que se quiere dilucidar, en este caso, cómo afectan las variables al comportamiento de renuncia. La siguiente tabla muestra las variables más relevantes para el modelo, encontrándose el detalle en la sección de anexos.

Variable	Importancia
Variación cobre 6 meses	1,00
Tiempo en la empresa	0,62
Porcentaje horas aprobadas capacitación	0,44
Promedio Asistencia a capacitación	0,36

Tabla 9 - Variables más importantes para Incremento de Gradiente

Tal como se puede evidenciar, el valor de la variación del precio del cobre (escogido a 6 meses como se explica en la sección 3.2.2. de Horizonte temporal del valor del cobre) es una variable bastante relevante para la predicción en el modelo. Como ya se mencionó, estas importancias no hablan sobre el comportamiento de los individuos frente a estas variables, por lo tanto, se realizaron las siguientes visualizaciones de datos que permiten evidenciar el comportamiento que se desea explicar.

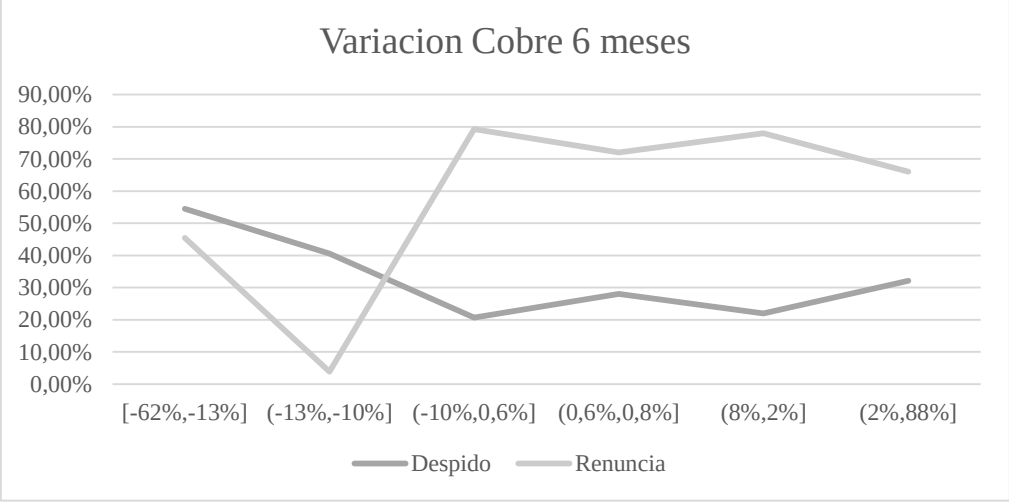


Ilustración 32 - Tasa de rotación por Variación del Precio del Cobre

Es claro ver cómo, a valores positivos del valor del metal transado, la tasa de renuncias aumenta considerablemente, mientras que la curva de despidos disminuye. Este comportamiento no es difícil de explicar, considerando que con el aumento del precio, se abren nuevos proyectos mineros dentro del país. Debido a esto, es que aumenta la tasa de fuga voluntaria de la empresa, debido a que mineras de mayor tonelaje anual comienzan a requerir más gente en sus operaciones, donde los operarios ven una oportunidad de aumentar su remuneración. En el caso de los despidos, es conocido el caso de que cuando el precio comienza a disminuir, existe la necesidad de reducción de costos que generalmente recaen en los recursos humanos [33].

La segunda variable más relevante, el tiempo en la empresa, expresa la cantidad de años de servicio del empleado en la compañía.

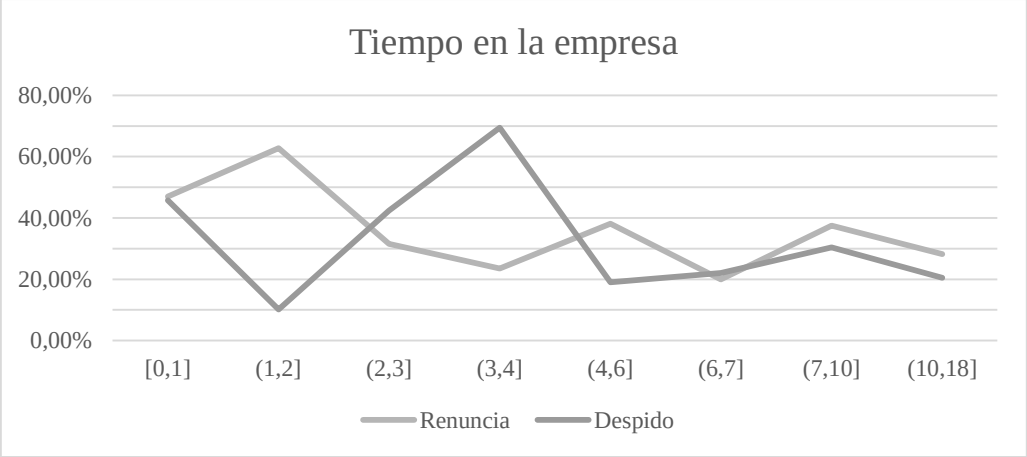


Ilustración 33 - Tasa de Rotación por Tiempo en la empresa

Para el caso de las renuncias, se ve una clara tendencia hacia la baja a medida que la persona tiene más años de servicio en la compañía, evidenciando un efecto de fidelización. Por otro lado, esta variable pareciera tener un comportamiento irregular para la curva de despidos, disminuyendo drásticamente hacia el segundo año y para luego repuntar drásticamente en los años siguientes, tiene explicación debido a las medidas de selección de personal de la empresa. El primer año se considera a prueba a la persona, por lo que puede ser despedida si no cumple con las expectativas de su empleador. Luego de pasar este filtro, difícilmente la persona será despedida en el siguiente año, pues ya demostró encontrarse capacitada para desarrollar la labor. Sin embargo, si los próximos años presenta una baja en su rendimiento, desde el punto de vista de su supervisor, probablemente sea despedido si la empresa así lo requiera, por ejemplo en escenarios donde el valor del precio del cobre no les permita mantener la dotación de personal. Finalmente, se aprecia una clara tendencia a nos despedir a personas que ya sobrepasan los 4 años de servicio en la empresa, lo que evidencia que estos empleados se encuentran en mejor posición frente a quienes llevan menos años, debido a la experiencia que poseen.

La tercera variable en orden de importancia es el porcentaje de horas aprobadas en capacitación. Cada curso que se realiza mediante las OTEC tiene un cierto número de horas lectivas, existiendo cursos que pueden durar unas pocas sesiones a varios meses. La gráfica que muestra esta variable para las clases de renuncia y despido son las siguientes.

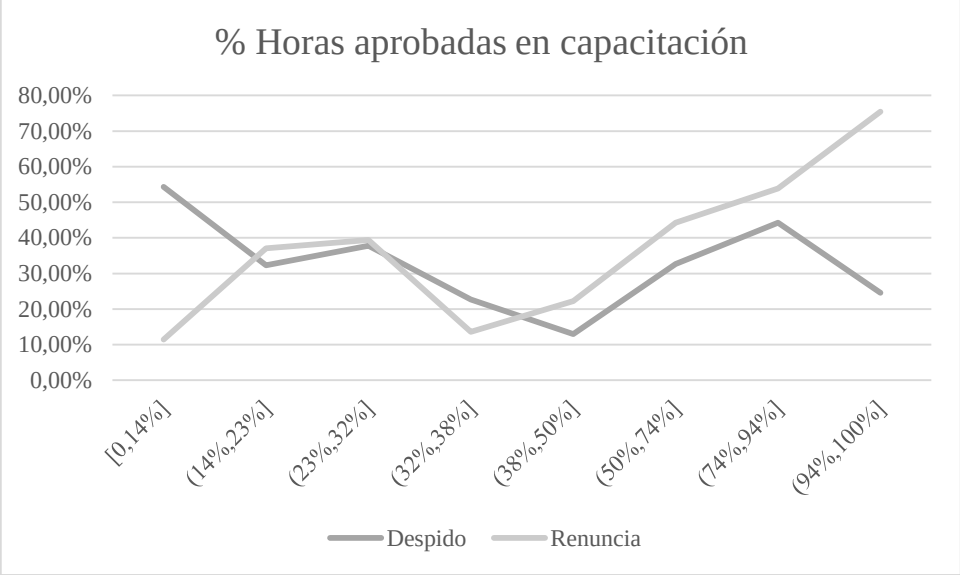


Ilustración 34 - Tasa de rotación por Horas aprobadas en capacitación

Se puede evidenciar que las personas que renunciaron voluntariamente a su cargo presentan una tasa bastante más alta de aprobación de las horas que realizan con respecto a sus pares que fueron despedidos.

Finalmente, la cuarta variable más relevante para el modelo tiene que ver con las capacitaciones que se realizan en la empresa. El promedio de asistencia a capacitaciones dibuja las siguientes curvas con respecto a renuncias y despidos.

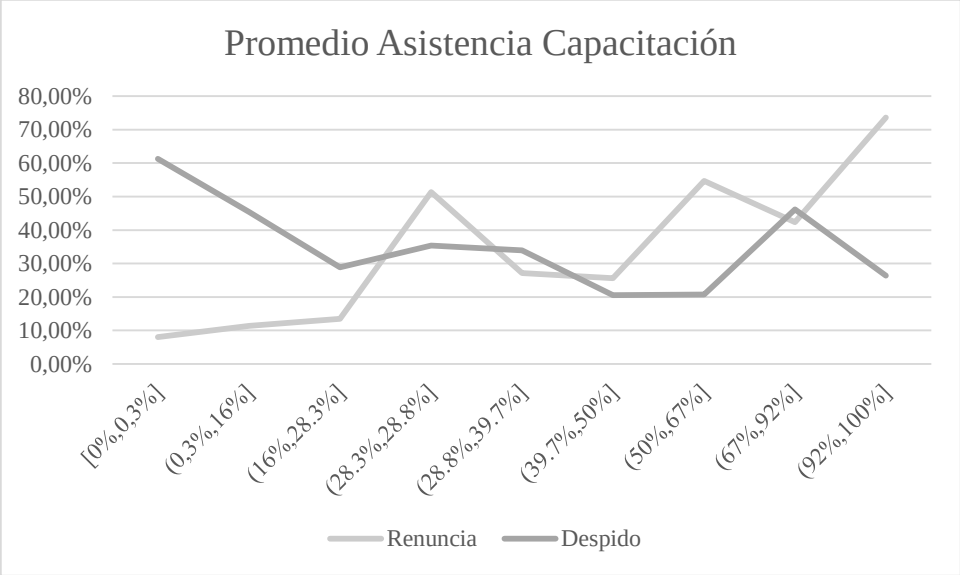


Ilustración 35 - Tasa de Rotación por Asistencia a capacitaciones

Claramente existe una tendencia a la alza en la asistencia a sus respectivas capacitaciones en las personas que renuncian voluntariamente a su cargo. En el caso contrario, existe una tendencia negativa en las personas que son despedidas de su cargo. Por lo tanto, se puede deducir que las personas que renuncian a sus cargos son personas las cuales, durante su tiempo de servicio en la empresa, se preocuparon por asistir a las oportunidades de perfeccionamiento que ésta les ofreció, aumentando de esta manera sus habilidades técnicas. En el caso de los despidos, se ve que estas personas no aprovechan de buena manera estas oportunidades, ausentándose a los cursos que la empresa ofrece. Si observamos la curva, se puede observar una tendencia negativa que muestra lo dicho anteriormente.

Si se ve solamente ambas variables discriminadoras para el modelo, se puede asumir que la empresa en cuestión está invirtiendo dinero en capacitaciones en personas que finalmente renunciarán a la empresa. Sin embargo, esto no es del todo cierto. Si se observa a continuación la curva de inversión en capacitación del personal, se puede ver que para las personas que renuncian, existe una menor inversión realizada, no así con los despidos, en quienes se hizo una mayor inversión.

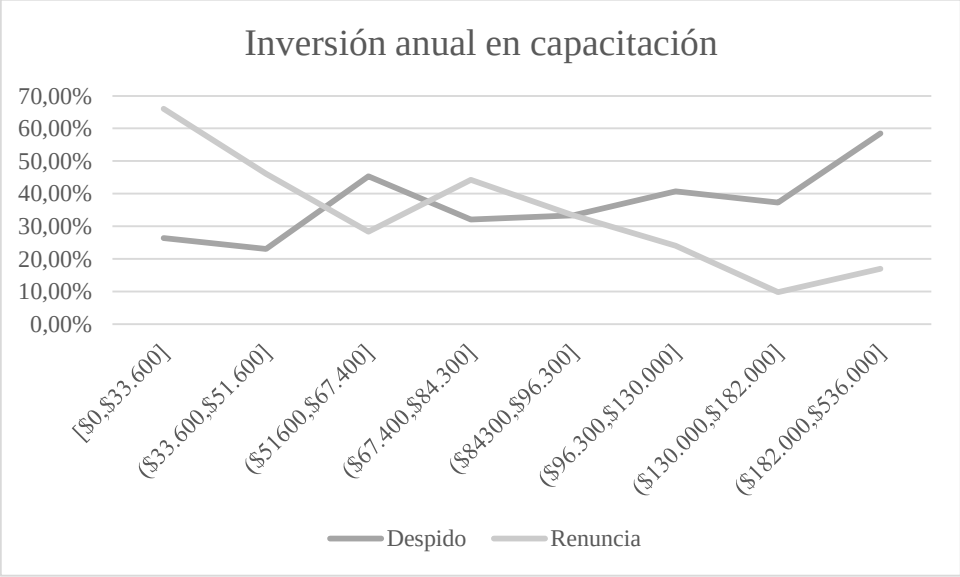


Ilustración 36 - Tasa de rotación por Inversión anual en capacitación

Por lo tanto, se puede deducir que si bien, existe una mayor asistencia y tasa de aprobación de cursos en quienes renunciaron voluntariamente, la inversión que se realiza en ellos es menor en promedio. En el caso de despidos, se ve que la empresa despide a personas que, además de no cumplir tanto con la aprobación y asistencia, se invierte mayor dinero en ellos, lo que evidencia un "desgaste" de la relación del empleado con su empleador, que finalmente se traduce en su desvinculación de la organización.

4.3.2. Resultados aprendizaje no supervisado

En el caso del método k-medias, a continuación se presenta la gráfica de la suma de errores cuadráticos, que permitirá discriminar la cantidad de clusters para este problema, donde la curva sólida representa la suma de errores cuadráticos para cada cantidad de clusters y la curva punteada representa la diferencia de error con respecto al punto anterior. Esto se llevó a cabo utilizando el paquete *Rattle*, el cual es una interfaz gráfica de minería de datos para el lenguaje de programación *R*, de código abierto.

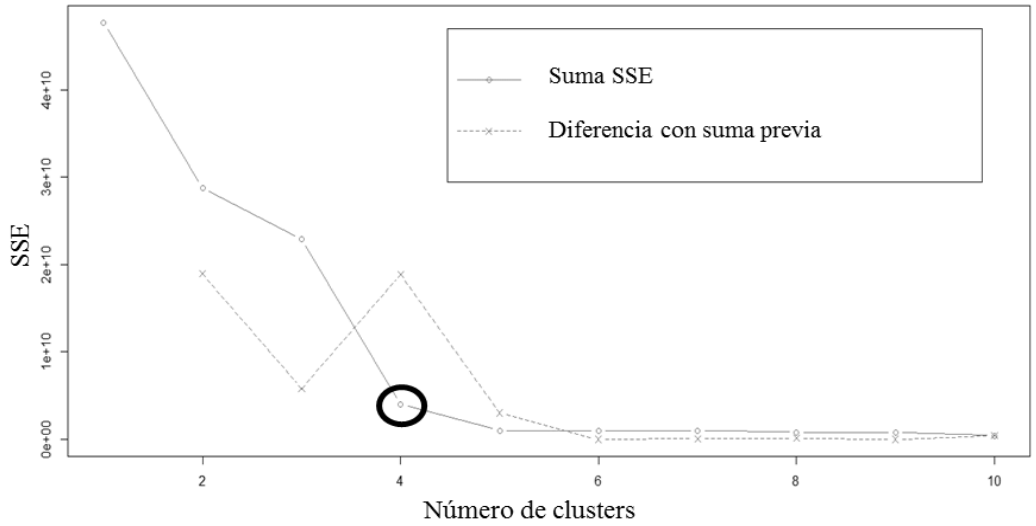


Ilustración 37 - Suma de errores cuadráticos para k-medias

Se puede observar un codo o punto de inflexión cuando se generan 4 clusters. Este número además permite interpretabilidad de los grupos generados, pudiendo "bautizarlos" para generar una mejor explicación del agrupamiento natural de los datos. En caso de que el codo se encuentre en un punto donde los clusters son demasiados, no encontrando una manera lógica de interpretarlos, se debe optar por disminuir su número en desmedro del aumento de error que se genere.

Los clusters muestran el distinto comportamiento de empleados, que se distinguen en gran parte por su edad y su tiempo en la empresa, encontrándose en anexos la tabla con el detalle de los centroides calculados y la interpretación de estos. Estos fueron nombrados de la siguiente manera:

- Empleados Antiguos (18,4%): poseen una alta tasa de sindicalización y en su mayoría cuentan con licencia de conducir profesional, lo que les permite ascenso dentro de la compañía. Renunciaron luego de superar el promedio de años de servicio de la compañía, siendo el

promedio de este segmento cercano a los 6 años de servicio y se encuentran cercanos a los 40 años de edad. Cuentan con buena asistencia y aprobación en capacitaciones

- Empleados Nuevos Senior (23,6%): quienes renunciaron en sus primeros años de servicio, pero tienen edad similar a los empleados del segmento anterior. Ellos presentan nula sindicalización, y frente a sus compañeros, presentan la menor asistencia a capacitaciones y un mayor número de ellas, además de no contar con licencia profesional.

- Empleado promedio (26,3%): quien tanto en edad como en años de servicio se encuentra en el promedio de la empresa, 37 y 4 años respectivamente. Posee una alta tasa de sindicalización y una buena tasa de aprobación y asistencia a capacitaciones, además de poseer con licencia profesional de conducir.

- Empleados Nuevos Jóvenes (31,6%): el cluster más numeroso. Son personas que renunciaron en sus primeros años de servicio, y tienen menos de 30 años. Si bien cuentan con licencia profesional, presentan una baja asistencia y aprobación en capacitaciones.

Considerando que, en el rubro de los operarios mineros, los ascensos en la compañía se dan cuando la persona es capaz de operar maquinaría de más alta complejidad, es importante considerar el tema de la licencia profesional de conducción. Se puede decir entonces que tanto el primer como cuarto segmento tienen mayores posibilidades de optar a un puesto mejor en otra compañía en comparación a quienes pertenecen al segundo y tercero debido a esta razón. Se puede ver además que tanto el primer como tercer segmento, probablemente preocupados por seguir creciendo en el rubro, mantienen una buena asistencia y aprobación de sus capacitaciones, lo que les permite aumentar su empleabilidad y por ende, acceder a un mejor trabajo. No así los últimos dos segmentos, los cuales podrían tener distintos motivos para no hacerlo. En el caso del *Empleado Nuevo Senior*, probablemente se mantenga rotando entre varias empresas y su motivación de cambio de empresa puede deberse a temas personales o de clima laboral, mientras que el *Empleado Nuevo Joven*, como ya posee licencia profesional y se encuentra en sus primeros años, no le es difícil cambiar su trabajo.

Para este último grupo, el cual es el más numeroso, se debe considerar que son personas que nacieron después de 1980, lo cual es llamado en la literatura como *Generación Y* [34]. Estas personas se caracterizan por valorar de distinta forma el balance entre su vida laboral y sus actividades fuera de ella con respecto a generaciones anteriores. Además, presentan mayor tendencia a cambiar de trabajo rápidamente con respecto a generaciones anteriores.

5. Discusión y Recomendaciones

La obtención de variables correspondientes a capacitaciones hace necesario encontrar una explicación de la relación entre éstas y las renuncias voluntarias. En el presente capítulo se muestra como la academia explica el fenómeno descrito, cómo esto se puede aplicar a la empresa estudiada y qué medidas pueden ser tomadas para manejar éste efecto. En conjunto a lo anterior, se muestra cómo se puede utilizar el modelo de Gradiente Incremental para generar perfiles de empleados fugados, además de información para Recursos Humanos. Finalmente, se muestra como un cambio en el esquema de Base de Datos puede ser de gran utilidad para mejorar la Gestión de Recursos Humanos.

5.1. Capacitaciones y Percepción de los empleados

Luego de realizar el modelamiento y obtener los patrones de comportamiento de los empleados que renuncian, se evidencia una clara relación de este fenómeno con las capacitaciones que se llevan a cabo en la empresa. Existe una relación entre las personas que aprovechan las oportunidades de capacitación que otorga la empresa con la renuncia voluntaria, para lo cual la literatura introduce dos conceptos clave [35]:

- Apoyo organizacional para el desarrollo (Organizational support for development): el OSD se define como la percepción que se tiene sobre la empresa acerca de la preocupación por la capacitación de sus empleados.

- Oportunidades de Carrera percibidas (Perceived carrer opportunity): el PCO se define como la percepción del empleado acerca de sus oportunidades de desarrollo de carrera dentro de la empresa.

En el estudio citado se demuestra empíricamente que, cuando una empresa posee un alto OSD y un bajo PCO, esto va a influir en el aumento de las renuncias voluntarias. Esto se explica por el aumento de empleabilidad que se le otorga al empleado, y la escasa posibilidad percibida de poder seguir creciendo dentro de la compañía. Debido a esta libertad otorgada por el conocimiento adquirido a través de las capacitaciones, la persona siente que puede cambiarse de compañía.

Ambas variables tienen relación con variables psicométricas que, como se mencionó en un comienzo, no serían consideradas debido a que no se cuenta con historia suficiente para poder ser incluidas en un modelo. Sin embargo, se utilizó una encuesta realizada por Great Place to Work[5]

[5] http://www.greatplacetowork.cl/

en la empresa durante el año 2011 para poder observar si en la empresa existe el efecto de OSD y PCO mencionado. Es importante mencionar que estas encuestas son anónimas y solo entregan un promedio por área de la compañía de los distintos ítems que contiene la encuesta. Para aproximar el OSD y PCO se utilizaron los siguientes ítems de la encuesta:

- "Se me ofrece entrenamiento para desarrollarme más profesionalmente" (OSD)
- "Los ascensos se les da a quienes más lo merecen" (PCO)

Para comprobar si existe una brecha, se utilizó un test de medias [36] para poder determinar si la diferencia de puntajes obtenidos para cada ítem son estadísticamente significativos, es decir, que existe evidencia estadística para asegurar que la diferencia no es producto del azar. El test procede de la siguiente manera, donde la hipótesis nula es que ambas medias son iguales, por lo tanto se busca rechazarla. La tabla de valores del estadístico t se encuentra en la sección de anexos.

$$H_0\colon \bar{Y}_1 = \bar{Y}_2$$

$$H_1\colon \bar{Y}_1 \neq \bar{Y}_2$$

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_{\bar{y}_1 - \bar{y}_2}},\ s_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ t_c = t^{\alpha}/_{n_1+n_2-2}$$

$$Rechazar\ H_0\ si\ t > t_c$$

Ecuación 22 - Test de Medias

La siguiente tabla muestra los resultados obtenidos al realizar este test para los cuatro turnos existentes de operarios mina.

DIMENSIÓN	n_1	n_2	$\bar{y}_1 - \bar{y}_2$	$s_{\bar{y}_1 - \bar{y}_2}$	t_i	¿Rechazo?
OSD - PCO	4	4	23	8,204	2,793	SI
$t_{95\%,6}$	2,447					

Tabla 10 - Test de medias para OSD – PCO

En consecuencia, se puede asegurar que la diferencia existente entre ambos ítems dentro del cuestionario de Great Place to Work es estadísticamente significativa al 95% de confianza. Esto se traduce en que, al menos durante el año 2011 (el cual fue el periodo con mayor cantidad de renuncias voluntarias dentro de la historia de la compañía) los operarios mina de la compañía si percibían una brecha entre OSD y PCO. Si bien, esta no es una prueba conclusiva, sirve para aproximar el conocimiento que existe en la literatura acerca de la relación de capacitaciones con

renuncias con la realidad de la empresa. Se puede intuir a través de esto que existe un problema en la percepción del plan de carrera que ofrece la compañía a sus trabajadores, el cual no cumple con sus expectativas. Considerando que el plan de carrera de un operario mina es bastante acotado, el cual depende prácticamente de la complejidad de la máquina que opera, probablemente estas personas vean que, dado que no pueden seguir subiendo en la escala en la compañía (lo que implica un aumento de la remuneración), pueden seguir aumentando sus ingresos en una faena de mayor tonelaje que puede pagar mejores sueldos. Esto implica entregar capital humano capacitado a la competencia.

5.2. Extensiones del modelo

Aprovechando la capacidad del modelo de gradiente incremental, el cual calcula probabilidades de pertenencia de cada individuo a una clase, sumado a la capacidad de los árboles de decisión de trabajar con variables continuas, es que se realizó un árbol de decisión para poder "ordenar" las variables del problema según las probabilidades calculadas. Para esto, y con el fin de sacar mayor provecho a los datos que no aparecieron en el modelo seleccionado, es que se realizó un árbol de decisión utilizando el software *SAS Enterprise Miner* donde la variable objetivo fuera la probabilidad de pertenecer a la clase "renuncia voluntaria", sin utilizar las variables respectivas a capacitaciones.

A partir del árbol, que se encuentra disponible en la sección de anexos, se pueden encontrar los siguientes patrones aproximados para la probabilidad de renuncia:

- Disminuye un 30% si la persona lleva más de 2.5 años en la empresa.
- Disminuye en un 15% cuando la persona es soltera, en comparación a los casados.
- Disminuye en un 15% si la persona es menor de 39 años.
- Aumenta en un 15% si la persona posee licencia profesional de conducir.

En primer lugar, estas probabilidades son condicionales a que ocurra lo que existe en la rama anterior, por lo que estas observaciones corresponden solo a observaciones de lo que el árbol muestra. Por ejemplo, el seguimiento de una rama se podría expresar de la siguiente manera:

"*Si la persona lleva más de 2.5 años, su probabilidad de renunciar será de 31%. Si cumple esto y además posee licencia profesional de conducir, su probabilidad aumentará 5%. Si cumple con todo lo anterior, y además tiene menos de 39 años, su probabilidad de fuga será cercana al 50%.*"

Se observa que debido a que el poder predictivo de estas variables para el modelo es menor con respecto a las analizadas en secciones anteriores, las probabilidades que muestra el árbol no superan el 60%, lo que quiere decir que en ningún caso el árbol generado es capaz de predecir de buena manera la renuncia (la cual en un modelo al azar correspondería a un $33,\overline{3}\%$ de probabilidad ya que existen 3 clases a predecir). No obstante, dada la historia de la compañía estos son puntos a considerar tanto para la contratación de nuevos empleados como también para gestionar de mejor manera los despidos, evitando despedir a quien es más propenso a dejar la compañía dado su perfil.

5.3. Recomendaciones

En base a lo encontrado utilizando tanto aprendizaje supervisado como no supervisado, donde se aprecia el efecto que tienen las capacitaciones sobre los empleados y cómo se comportaron los empleados que renunciaron voluntariamente frente a éstas es que las siguientes recomendaciones apuntan a la gestión de capacitaciones dentro de la empresa y como modificar la percepción de los empleados vistas en la sección 5.1.

En primera instancia, se podría pensar en que las capacitaciones no tienen un efecto positivo dentro de la organización, ya que parecieran ser un factor motivador a la renuncia. Por lo tanto, esto lleva a pensar en que la empresa no debería realizar capacitaciones a sus empleados. Sin embargo la ley no permite hacer esto, tal como se puede observar en la sección 1.4. donde se explica que la empresa debe mantener responsabilidad frente a este tema y que debe elegir un comité conformado por sus trabajadores para velar por ello. Además, esto afectaría la reputación de la empresa y en consecuencia, el OSD percibido por las personas, haciendo a la compañía poco atractiva para nuevos empleados.

Otra forma que podría pensarse para enfrentar el problema es considerar un bono por mérito relacionado a capacitaciones[6]. De esta forma se entregaría una compensación monetaria para retener a quienes demuestren buen comportamiento en las capacitaciones, es decir, alta tasa de asistencia y aprobación. No obstante, para efectuar una medida así se debe tener en cuenta al sindicato de trabajadores, quienes podrían oponerse a medidas que consideren no sean igualitarias para quienes suscriben al contrato sindical. En adición a este punto, y considerando lo expuesto en la sección anterior, un bono de éstas características podría no tener el efecto deseado si a esto no se suman oportunidades de crecimiento dentro de la empresa. Si esto último no se propicia, el bono podría tener un efecto positivo en el comportamiento de los empleados frente a capacitaciones, lo cual aumentaría tanto su productividad como su empleabilidad, pero no necesariamente actuaría como medida de retención, ya que este aumento de empleabilidad debido a los cursos realizados puede transformarse en una motivación para dejar la empresa.

[6] No se hace referencia al existente "Bono por mérito" otorgado a empleados que cumplan más de tres años en su máximo subnivel de sueldo y posean una evaluación de desempeño mayor a 80%.

Dicho lo anterior, una forma es proponer nuevas dimensiones a considerar a la hora de que se realicen ascensos para los operarios de la empresa. Generalmente en la minería se consideran los siguientes aspectos a la hora de proponer el ascenso de un empleado:

- Mínimo de años de servicio en la compañía
- Competencia certificada por experto
- Recomendación del supervisor
- Rendimiento dentro del área
- Baja accidentabilidad

Sin embargo, no siempre es considerado el esfuerzo realizado por los empleados en las capacitaciones que la empresa otorga para que se genere un ascenso. Por esta razón se propone que el comportamiento en capacitaciones sea considerado a la hora de ascender de puesto a una persona dentro de la compañía. Por un lado, se toma en cuenta lo observado en la encuesta de 2011 expuesta en la sección 5.1. , donde los operarios declaran desde su perspectiva que los ascensos no siempre los tienen quienes más lo merecen. Por otra parte evita que el empleado incurra en el riesgo que implica salir al mercado laboral en busca de un nuevo empleo, con el fin de encontrar mejores condiciones, manteniéndolo en la empresa. Tomando en cuenta los resultados obtenidos de la clusterización, se puede deducir que para el grupo de *Empleados Antiguos* quienes poseen la mayor cantidad de años de servicio en la empresa, y por ende mayor experiencia, esta medida tendría el efecto de retención deseado, puesto que presentan alta tasa de asistencia y aprobación de sus cursos, pero contrastado con las medidas de OSD y PCO, pareciera que no perciben oportunidades de crecimiento en la empresa por lo que buscan trabajo en otra compañía. Para el grupo de *Empleados Promedio* se obtendría un efecto similar, ya que también presentan buena tasa de aprobación y asistencia. En el caso de *Empleados Nuevos Senior* y *Empleados Nuevos Jóvenes*, quienes presentan tasas de aprobación y asistencia más bajas, se busca que esta medida tenga un efecto motivador en los empleados, promoviendo el buen comportamiento en capacitaciones, aumento en su productividad y obteniendo un efecto de fidelización con la empresa, lo cual se ha observado en otro tipo de industrias [37].

5.4. Esquema de Base de Datos

Si bien en la compañía existen bases de datos para manejar sus recursos humanos, se plantea la alternativa de crear un modelo de Data Warehouse, el cual está orientado a generar información. Aunque la diferencia entre datos e información puede parecer sutil, se considera que la información es lo que se obtiene mediante el análisis de los datos, con lo cual se busca tomar decisiones. Por lo tanto, el dato sería la unidad elemental de información. Un Data Warehouse, según su creador W.H Inmon se define de la siguiente manera [38]:

"…una colección de datos ***orientada al negocio***, ***integrada***, ***variante en el tiempo*** *y* ***no volátil*** *para el soporte de la toma de decisiones de la gerencia."*

- Orientada al negocio: Está orientado a satisfacer las necesidades de investigación de los datos por un usuario de alto nivel, de un área en particular de la organización.
- Integrada: integra datos desde diversas fuentes.
- Variante en el tiempo: la data histórica permanece registrada en el Data Warehouse. En contraste con los sistemas de tipo transaccional, los cuales en muchos casos registran solo los datos más nuevos (por ejemplo, en el caso de una base de datos de clientes, solo registran su última dirección).
- No volátil: una vez que los datos están en el Data Warehouse, esta no sufrirá modificaciones, por lo tanto, la data histórica no debería ser alterada.

El modelo que se utiliza para desarrollar una base de datos que cumpla con estas características recibe el nombre de modelo estrella, debido a que contiene una tabla central o *tabla de hechos* que posee indicadores calculados a partir de los datos de distintas fuentes que se reciben desde las tablas de las puntas, llamadas *dimensiones*.

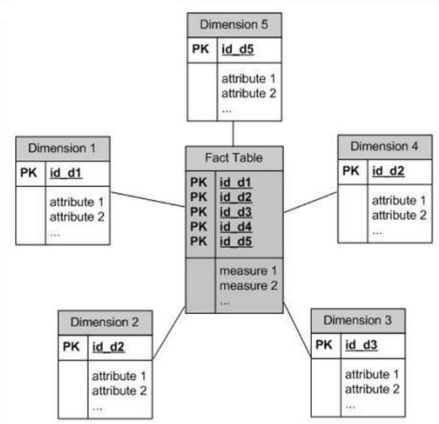


Ilustración 38 - Ejemplo de modelo estrella

En el caso particular de la empresa estudiada, existen diversas fuentes de datos que podrían conformar un esquema de base de datos de este tipo. Además, la inclusión de variables como las evaluaciones de desempeño y de remuneraciones y bonos, las cuales no fueron utilizadas en el presente trabajo por temas de confidencialidad, podrían ser incluidas para aumentar la tasa de éxito general del modelo en general, como también de cada una las clases predichas. Se propone un esquema como el siguiente, en el cual se agregan a los datos existentes variables correspondientes a remuneración y evaluaciones de desempeño[7]:

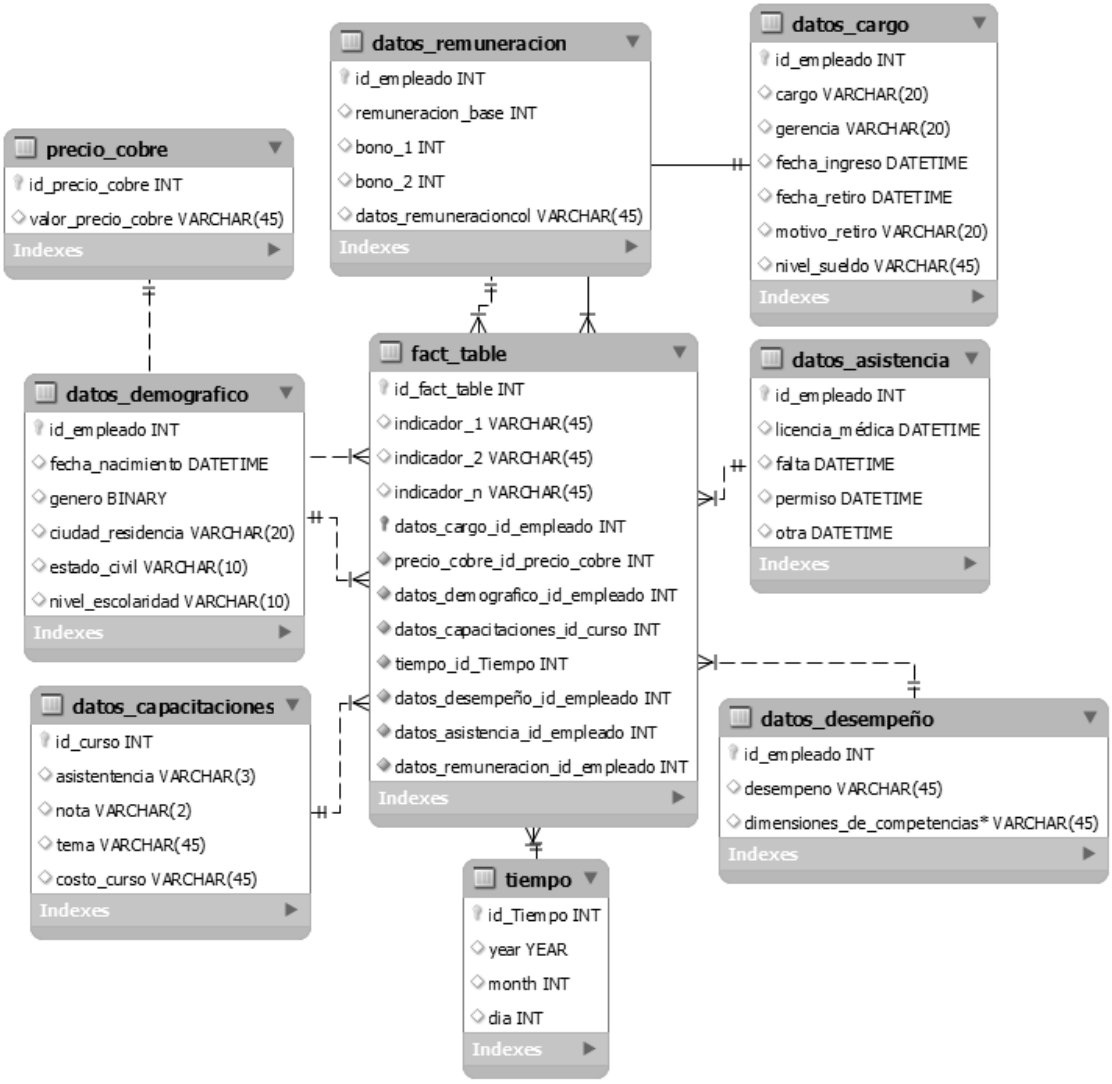


Ilustración 39 - Modelo Estrella RR.HH

[7] Dimensiones de competencias disponibles en anexos

Esta estructura permite el cálculo de indicadores de gestión, en este caso de RR.HH. de manera rápida, permitiendo controlar la *granularidad* (nivel de detalle) de cada una de las dimensiones. Con esto se puede calcular de manera rápida, por ejemplo, el dinero invertido en bonificaciones o capacitaciones en operarios mina en periodo determinado de tiempo, pudiendo generar reportes de manera rápida dentro de la faena para tener un mejor control.

6. Conclusiones y trabajos futuros

6.1. Conclusiones

Los proyectos de Minería de Datos, a modo general, pueden tener resultados que se enmarquen en los siguientes estados:

- Resultados sobre los cuales la organización tenía sospecha o que existe conocimiento popular.
- Resultados que muestren patrones nunca antes vistos y que otorguen una explicación una explicación lógica al problema.
- Resultados que muestren patrones sobre los cuales no se encuentra una explicación lógica.
- El resultado del modelo no es satisfactorio para explicar el problema.

Los resultados encontrados en este caso muestran un comportamiento que es conocido por la empresa: la rotación de personal aumenta cuando el precio del cobre se encuentran en alza. Además, relativo al tema de capacitaciones ya existen estudios de psicología laboral sobre cómo éstas afectan en la renuncia voluntaria [35], debido al alza empleabilidad de las personas, que le otorga oportunidades para postular a empresas mineras de mayor tonelaje, y sus percepciones frente a las posibilidades de ascenso y capacitación que otorga la empresa. Aunque existía la sospecha sobre el fenómeno y no se descubrió algún patrón desconocido, el uso de la metodología KDD permitió confirmar el comportamiento a través del uso empírico de datos. Además, permitió descartar variables que se creían relevantes dentro de la empresa, como lo era la distancia al hogar y los niveles de sueldo, los cuales no aparecen como buenas variables predictoras. Para el caso de la distancia al hogar, esto se puede deber principalmente a que la cantidad de personas que residen fuera de la región de Antofagasta, donde se ubica la faena, es bastante baja con respecto a quienes si lo hacen por lo que se hace difícil para los modelos discriminar a estas personas. En el caso de los niveles de sueldo, no existe evidencia en el estudio de que existan diferencias significativas entre quienes se encuentran en un nivel más bajo o más alto de sueldo. Sin embargo, este nivel indica la renta fija de los empleados, la cual varía bastante mediante bonos otorgados, información que no pudo ser utilizada por confidencialidad.

Se puede ver que el uso de estructuras de datos que propicien la captura de cada individuo particular, en contraposición a la estructura clásica que tienen las bases de Recursos Humanos que asimila a los sistemas transaccionales, son de gran utilidad para realizar análisis a nivel de la dotación completa del personal. Para esto, existen estructuras como el modelo estrella antes

mencionado, que permiten regular la granularidad de los datos, pudiendo observar comportamiento tanto de conjuntos de personas como de individuos particulares, pudiendo ser regulado además por las otras dimensiones del modelo.

Por otra parte, si bien el objetivo final del proyecto no es realizar predicción de despidos, se puede evidenciar que en general para los modelos es difícil diferenciar esta clase utilizando los datos de la empresa. Una explicación sensata es que, dado que esta es una decisión de la empresa y no del individuo, difícilmente este estado puede ser determinado utilizando indicadores para la persona. Sin embargo, se puede observar qué dado los datos disponibles, pareciera que el comportamiento de despidos parece ser relativamente aleatorio por parte de la empresa, lo cual no debería ocurrir. Esta predicción posiblemente mejoraría incorporando las evaluaciones de desempeño de los empleados, las cuales son un factor crítico a la hora de que la empresa decide reducir su dotación de personal.

Finalmente, la metodología plateada permite, además de obtener información desde los datos para conocer lo que ha acontecido, predecir para los empleados actuales de la empresa su probabilidad de fuga dada sus características. Esta es una poderosa herramienta en la empresa considerando el panorama actual de la minería donde sería bastante costoso para la empresa reemplazar a un operario experimentado. En consecuencia, el modelo puede decir quiénes son las personas que son más propensas a renunciar lo que permitirá a la empresa tomar acciones de retención sobre estas personas, siempre y cuando esto sea considerado necesario, como en el caso de los operarios de pala quienes tienen asociada una vasta experiencia debido a la complejidad de la máquina.

6.2. Trabajos futuros

La inclusión de variables más específicas de remuneraciones, en particular el caso de los bonos, podría incrementar tanto el poder predictivo de los modelos como también dar importantes conclusiones acerca de la política de remuneraciones de la empresa. Esto, en conjunto con las evaluaciones de desempeño, podría determinar efectos de "justicia" percibida por los empleados. Esta justicia se define como la remuneración que recibe la persona en comparación a sus pares que realizan actividades similares dentro de la empresa. Así, utilizando las evaluaciones de desempeño se podría observar si efectivamente existen estas diferencias, las cuales también son una fuerte causa de las renuncias voluntarias dentro de una empresa [11].

Aunque este proyecto plantea el estudio del comportamiento de las personas de la compañía sin la utilización de fuentes externas de datos (excepto si estas son gratuitas), una extensión de este trabajo podría considerar la inclusión de estudios de benchmark de sueldos de otras compañías mineras, lo que ayudaría a determinar si efectivamente la gente deja su puesto por temas de remuneraciones o existen otros motivos relacionados. Estos datos, sumados a variables de tipo

psicométrico, podrían aclarar si los operarios de esta compañía consideran más importante aspectos del ambiente laboral o su remuneración a la hora de tomar la decisión de renunciar voluntariamente a la empresa.

Se plantea además un modelamiento más exhaustivo del esquema de Data Warehouse planteado, el cual podría ayudar a mejorar la gestión de personas tanto dentro de la empresa estudiada, como también en otras industrias. Con esto, se podría desarrollar de mejor manera tanto la política de despidos, comportamiento el cual no pudo ser bien discriminado por los modelos, ya que permite generar rápidos reportes sobre el comportamiento de los trabajadores. También aplica para llevar un control más automatizado de los gastos que se incurren tanto en capacitaciones como bonificaciones.

Considerando datos de desempeño de la persona en la faena, con los cuales no se contaba para el presente estudio, se puede desarrollar la valoración económica de la retención de empleados. Proyectando el rendimiento futuro de una persona propensa a la fuga y teniendo en cuenta su remuneración fija y variable, se puede calcular el valor económico de la persona para la faena y ver como su ausencia afecta en la unidad. Realizando una adaptación del *Customer Lifetime Value* [39] para este caso, se puede resolver el siguiente problema de optimización.

$$\max \left\{ \sum_{j=1}^{J} \sum_{i=1}^{I} \left(\sum_{t=1}^{T} (1 - p_{tij}) * \left(\frac{Productividad_{tij} + sueldo_{tij}}{(1+r)^t} \right) \right) * x_{ij} \right\}$$

$$s.a. \quad \sum_{i=1}^{I} x_{ij} \leq N_j, \qquad \forall j$$

Ecuación 23 - Problema de optimización propuesto

Donde p_{tij} es la probabilidad de renuncia calculada por el modelo de la persona i del turno j en el tiempo t, la $productividad_{tij}$ es el valor monetario de la producción del individuo i perteneciente al turno j en el tiempo t, teniendo un parámetro análogo $sueldo_{tij}$ que considera el sueldo fijo más el variable. Se tiene una tasa de descuento r para traer esto a valor presente. La variable de decisión x_{ij} es de tipo binaria, la cual indica 1 si la persona es ascendida y 0 si no, la cual está sujeta a que no pueden existir más de N_j ascensos en el turno j. Es importante considerar para este problema que si una persona k deja la empresa, la productividad del conjunto de personas en el turno j será menor que la simple resta de su productividad al conjunto entero, debido a las ineficiencias relacionadas a la renuncia expuestas en este informe.

$$\sum_{i=1}^{I} productividad_{ij} - productividad_{kj} \geq \sum_{i \in I/\{k\}} productividad_{ij} \, , \forall j$$

Ecuación 24 - Efecto de la renuncia en productividad

Finalmente, la metodología utilizada en este proyecto es extrapolable a otros cargos dentro de la compañía, ya que las conclusiones de este estudio son solo aplicables para los operarios mina, pudiendo presentarse comportamientos distintos en perfiles administrativos y/o profesionales.

7. Bibliografía

[1] SOFOFA, «Sociedad de Fomento Fabril Chile,» 2014. [En línea]. Available: http://web.sofofa.cl/informacion-economica/indicadores-economicos/estructura-de-la-industria/.

[2] U.S. Geological Survey, «Mineral Commodity Summaries,» 2012. [En línea]. [Último acceso: Junio 2015].

[3] El Mercurio, «Economía y Negocios,» 18 Octubre 2011. [En línea]. Available: http://www.economiaynegocios.cl/noticias/noticias.asp?id=89607. [Último acceso: Mayo 2015].

[4] La Tercera, «La Tercera Negocios,» Julio 2015. [En línea]. Available: http://www.latercera.com/noticia/negocios/2015/07/655-639487-9-cochilco-advierte-que-produccion-de-cobre-podria-caer-a-contar-del-2025.shtml.

[5] COCHILCO, «www.cochilco.cl,» 1 Julio 2015. [En línea]. Available: http://www.cochilco.cl/estadisticas/grafico.asp?tipo_metal=1. [Último acceso: Julio 2015].

[6] Economía y Negocios, El Mercurio, «El 85% de los trabajadores de proyectos de Codelco cambió de empleo en 15 meses,» 5 Marzo 2013. [En línea]. Available: http://www.economiaynegocios.cl/noticias/noticias.asp?id=106476. [Último acceso: Junio 2015].

[7] Fundación Chile, Fuerza laboral en la gran minería chilena, Santiago, 2011.

[8] Consejo de Competencias Mineras, Fuerza Laboral de la gran minería en Chile 2013 - 2022, Santiago: Innovum - Fundacion Chile, 2013.

[9] Centro de Competencias Mineras, «www.ccm.cl,» 2014. [En línea]. Available: http://www.ccm.cl/app_ccm/frontend/modules/index.php?modulo=datamart&accion=GenerarGrafico&idSeccion=analisis_sectorial&idTipoFuenteDatos=perfiles_cargo_proy&idFuenteDatos=demanda_total&idGrafico=column&idFiltro=sin_filtro. [Último acceso: Julio 2015].

[10] D. Goyal y S. Ahson, «Data Mining Techniques for better decisions in human resource management systems,» *Int. J. Business Information Systems, Vol. 3, No. 5*, pp. 464-481, 2008.

[11] M. Kane-Sellers, Predictive Models of employee voluntary turnover in a North American professional sales force using data mining analysis, Ph.D. dissertation, Texas: A&M University, 2007.

[12] W. H. Mobley, «Intermediate Linkages in the Relationship Between Job Satisfaction and Employee Turnover,» *Journal of Applied Psychology,* vol. 62, nº 2, pp. 237-240, 1977.

[13] B. Becker y B. Gerhart, «The Impact of Human Resource Management on Organizational Performance: Progress and Prospects,» *The Academy of Management Journal* , vol. 39, nº 4, pp. 779-801, 1996.

[14] M. Armstrong, «Learning and development,» de *A Handbook of Human Resource Management Practice,* Cambridge University Press, 2006, pp. 559-583.

[15] R. A. Feinberg y N. Jeppeson, «Validity of exit interviews in retailing,» *Journal of Retailing and Consumer Services,* vol. 7, pp. 123-127, 2000.

[16] J. Ranjan, D. Goyal y S. Ahson, «Data Mining Techniques for better decisions in human resource management systems,» *Int. J. Business Information Systems, Vol 3, No. 5,* pp. 464-480, 2008.

[17] C.-Y. Fan, «Using Hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals,» *Expert System with Applications,* pp. 8844-8851, 2012.

[18] C.-F. Chien y L.-F. Chen, «Using Rough Set Theory to Recruit and Retain High-Potential Talents for Semiconductor Manifacturing,» *IEEE Transactions on Semiconductor Manufacturing,* pp. 528-541, 2007.

[19] D. Waintrub y J. Miranda, "Prediccion de Fuga de Empleados con Modelos Hibridos en Data Mining para una empresa de Servicios Mineros", Santiago: Facultad de Economia y Negocios, Universidad de Chile, 2013.

[20] Ministerio del trabajo y previsión social, «www.leychile.cl,» Biblioteca del Congreso Nacional de Chile, 10 Septiembre 1997. [En línea]. Available: http://www.leychile.cl/Navegar?idNorma=76201. [Último acceso: Junio 2015].

[21] Dirección del Trabajo, «www.dt.gob.cl/,» Dirección del Trabajo, 19 Agosto 1997. [En línea]. Available: http://www.dt.gob.cl/legislacion/1611/w3-article-87630.html. [Último acceso: Junio 2015].

[22] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «The KDD Process for Extracting Useful Knowledge from Volumes of Data,» *COMMUNICATIONS OF THE ACM,* vol. 39, nº 11, pp. 27-34, 1996.

[23] V. Labatut y H. Cherifi, «Accuracy Measures for the Comparison for Classifiers,» *The 5th International Conference on Information Technology,* 2011.

[24] M. Sokolova y G. Lapalme, «A systematic analyisis of performance measures for classification tasks,» *Information Processing and Management* , vol. 45, pp. 427-437, 2009.

[25] J. Weston y C. Watkins, «Support Vector Machines for Multi-Class Pattern Recognition,» *ESANN*, vol. 99, pp. 219-224, 1999.

[26] C. Gershenson, «Cornell University Library,» 20 Agosto 2003. [En línea]. Available: http://arxiv.org/abs/cs/0308031. [Último acceso: Julio 2015].

[27] Y. Croissant, *Estimation of multinomial logit models in R: The mlogit Packages*, Université de la Réunion.

[28] G. Williams, «Chapter 12: Random Forest,» de *Data Mining with Rattle and R: The art of Excavating Data for Knowledge Discovery*, 2011, pp. 245-260.

[29] J. H. Friedman, «Greedy Function aproximation: A Gradient Boosting Machine,» *The Annals of Statistics*, vol. 29, nº 5, pp. 1189-1232, 2001.

[30] A. K. Jain, «Data clustering: 50 years beyond K-means,» *Pattern recognition letters*, vol. 31, nº 8, pp. 651-666, 2010.

[31] K.-A. Yoon, O.-S. Kwon y D.-H. Bae, «An Approach to Outlier Detection of Software Measurement Data using the K-means Clustering Method,» de *Empirical Software Engineering and Measurement, 2007. ESEM 2007.*, Madrid, 2007.

[32] F. Massey, «The Kolmogorov-Smirnov Test for Goodness of Fit,» *Journal of the American Statistical Association*, pp. 68-78, 1951.

[33] Revista Minería Chilena, «http://www.mch.cl/,» 12 Enero 2015. [En línea]. Available: http://www.mch.cl/2015/01/12/baja-del-cobre-lleva-grandes-mineras-realizar-ajustes-y-hay-mas-de-10-000-empleos-en-riesgo-en-2015/.

[34] J. Jamrog, «The perfect storm: The future of retention and engagement.,» *Human Resource Planning*, vol. 27, nº 3, pp. 26-33, 2004.

[35] M. L. Kraimer, S. E. Seibert, S. J. Wayne, R. C. Liden y Bravo, «Antecedents and outcomes of Organizational Support for Development: The critical role of career opportunities,» *Journal of Applied Psychology*, vol. 96, nº 3, pp. 485-500, 2011.

[36] N. Malhotra, «PART III: DATA COLLECTION, PREPARATION, ANALYSIS AND REPORTING,» de *Marketing Research: An Applied Orientation*, Pearson Education India, 2008.

[37] P. AlKhoury, M. AlKotob, C. Iskandar, F. ElAmad, T. Mezher, T. Saidi, W. Ghazzawi y Z. AlBaba, «Employees' Perception About the Effect of Training on

Promotion: Evidence from Lebanon,» *Global Journal of Business Research,* vol. 8, nº 2, pp. 23-31, 2014.

[38] W. H. Inmon, «What is a Data Warehouse,» *Prism Tech Topic,* vol. 1, nº 1, 1995.

[39] Consejo de Competencias Mineras, Marco de cualificaciones para Minería, Santiago: Fundación Chile, 2012.

8. Anexos

8.1. Matrices de confusión

Modelo	Árbol de Decisión		
Real\Predicha	Permanencia	Despido	Renuncia
Permanencia	25,32%	0,00%	0,00%
Despido	7,59%	10,13%	16,46%
Renuncia	1,27%	0,00%	39,24%

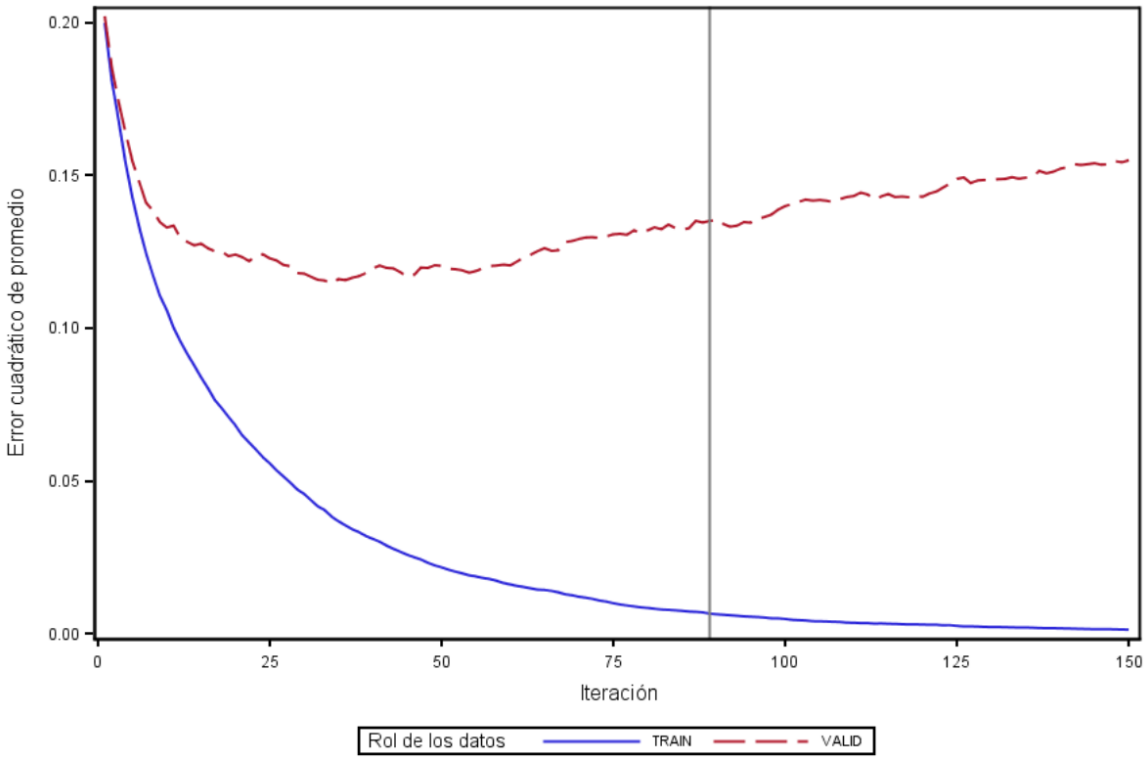
Modelo	Random Forest		
Real\Predicha	Permanencia	Despido	Renuncia
Permanencia	16,36%	11,82%	0,00%
Despido	8,18%	18,18%	9,09%
Renuncia	0,00%	3,64%	32,73%

Modelo	Logit Multinomial		
Real\Predicha	Permanencia	Despido	Renuncia
Permanencia	12,66%	10,13%	2,53%
Despido	5,06%	15,19%	13,92%
Renuncia	2,53%	6,33%	31,65%

Modelo	Red Neuronal		
Real\Predicha	Permanencia	Despido	Renuncia
Permanencia	13,92%	8,86%	2,53%
Despido	2,53%	20,25%	11,39%
Renuncia	3,80%	5,06%	31,65%

Modelo	Incremento de Gradiente		
Real\Predicha	Permanencia	Despido	Renuncia
Permanencia	24,79%	1,71%	0,00%
Despido	7,69%	17,95%	10,26%
Renuncia	0,00%	1,71%	35,90%

8.2. Gráfico de error cuadrático promedio Incremento de Gradiente



8.3. Importancia de Variables para Incremento de Gradiente

Variables	Importancia Relativa
variacion_cobre_6meses	1,00
tiempo_empresa	0,62
porcentaje_horas_aprobadas_capacitacion	0,44
AVG_of_Asistencia	0,36
meses_desde_capacitacion	0,27
numero_vacaciones_anual	0,24
edad_empresa	0,21
numero_licencias_anual	0,16
inversion_anual_capacitacion	0,16
dias_licencia_anual	0,14
duracion_ultima_licencia	0,14
meses_desde_licencia	0,13
dias_vacaciones_anual	0,13
duracion_ultima_vacacion	0,10
SUM_of_vacaciones_cortas	0,08
SUM_of_licencia_medica	0,07
distancia_hogar_lomas	0,05
SUM_of_vacaciones_completas	0,03
capacitacion_anual	0,02
meses_desde_vacaciones	0,00

8.4. Datos normalizados para el método de kmeans.

Variable	Promedio Población
edad_empresa	0,37782
reingreso	0,02632
sindicalizado	0,55263
IMP_SUM_of_Falta	0,28947
IMP_SUM_of_licencia_medica	0,19130
IMP_SUM_of_vacaciones_cortas	0,15213
IMP_dias_licencia_anual	0,28575
IMP_distancia_hogar_lomas	0,21716
IMP_duracion_ultima_vacacion	0,45523
IMP_licencia_profesional	0,65789
IMP_meses_desde_licencia	0,15109
IMP_numero_licencias_anual	0,31261
IMP_porcentaje_horas_aprobadas_c	0,44734
TIN_estado_civil_mayus_SOLTERO.A.	0,39474
tiempo_empresa	0,25439
estudio_en_empresa	0,05263
IMP_AVG_of_Asistencia	0,41334
IMP_SUM_of_Permiso	0,11842
IMP_SUM_of_vacaciones_completas	0,21053
IMP_capacitacion_anual	0,59478
IMP_dias_vacaciones_anual	0,57010
IMP_duracion_ultima_licencia	0,36992
IMP_inversion_anual_capacitacion	0,36810
IMP_meses_desde_capacitacion	0,52658
IMP_meses_desde_vacaciones	0,20991
IMP_numero_vacaciones_anual	0,14230
TIN_estado_civil_mayus_CONVIVENCIA	0,02632
TIN_IMP_causa_ultima_licencia_Permiso	0,05263

8.5. Centroides para método de kmedias

Variable / Cluster	1	2	3	4
edad_empresa	0,57143	0,42857	0,35714	0,24405
tiempo_empresa	0,60317	0,01235	0,30000	0,19444
reingreso	0,00000	0,11111	0,00000	0,00000
estudio_en_empresa	0,28571	0,00000	0,00000	0,00000
sindicalizado	1,00000	0,00000	0,80000	0,50000
AVG_of_Asistencia	0,55177	0,23399	0,76173	0,17679
SUM_of_Falta	0,00000	1,00000	0,10000	0,08333
SUM_of_Permiso	0,14286	0,00000	0,25000	0,08333
SUM_of_licencia_medica	0,41176	0,10471	0,14576	0,16559
SUM_of_vacaciones_completas	0,45238	0,14815	0,23333	0,09722
SUM_of_vacaciones_cortas	0,52857	0,09223	0,06932	0,04647
capacitacion_anual	0,37742	0,84524	0,18146	0,87815
dias_licencia_anual	0,33235	0,16474	0,28760	0,34777
dias_vacaciones_anual	0,70385	0,65691	0,52233	0,46679
distancia_hogar_lomas	0,10304	0,30895	0,06290	0,34343
duracion_ultima_licencia	0,39286	0,41429	0,41143	0,28869
duracion_ultima_vacacion	0,46154	0,22682	0,62751	0,47929
inversion_anual_capacitacion	0,30553	0,42500	0,31175	0,40888
licencia_profesional	0,85714	0,55556	0,20000	1,00000
meses_desde_capacitacion	0,44286	0,49075	0,42014	0,69099
meses_desde_licencia	0,22272	0,29633	0,07039	0,06762
meses_desde_vacaciones	0,03759	0,19753	0,29942	0,24513
numero_licencias_anual	0,30612	0,25497	0,25617	0,40666
numero_vacaciones_anual	0,52907	0,00862	0,11015	0,04374
porcentaje_horas_aprobadas_c	0,73137	0,22416	0,83876	0,12287
estado_civil_mayus_CONVIVENCIA	0,00000	0,11111	0,00000	0,00000
estado_civil_mayus_SOLTERO.A.	0,00000	0,22222	0,80000	0,41667
causa_ultima_licencia_Permiso	0,00000	0,00000	0,10000	0,08333
Tamaño Cluster	**18,42%**	**23,68%**	**26,32%**	**31,58%**

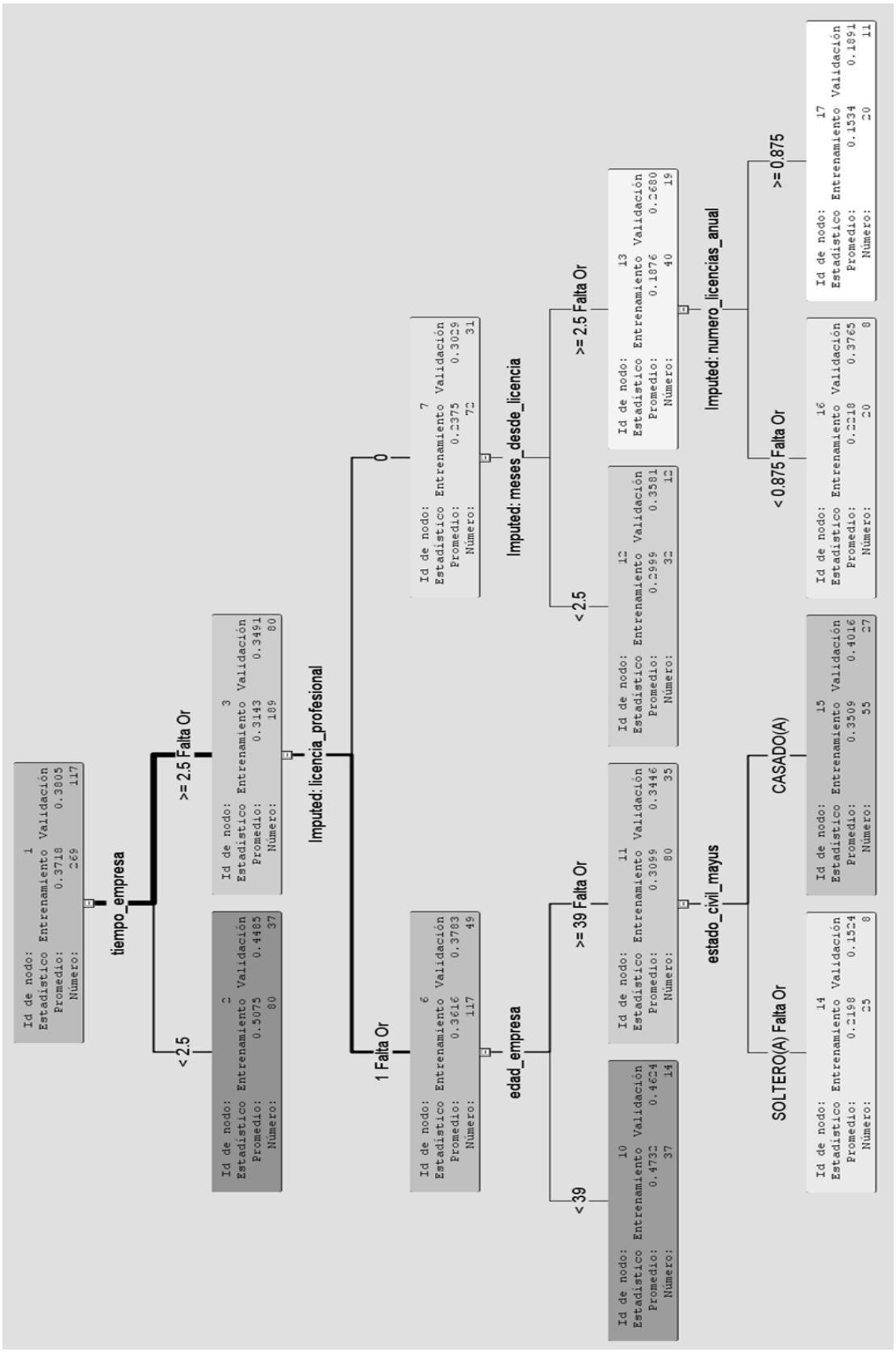
8.6. Interpretación de Clusters

Segmento 1 (18,4%) Empleados Antiguos	Edad mayor al promedio Mayor cantidad de tiempo en la empresa Siempre trabajaron en la empresa Realizaron estudios estando dentro de la empresa Alta tasa de sindicalización Buena asistencia a capacitaciones Prácticamente sin faltas dentro de su historial La mayor tasa de licencias médicas Menor número de capacitaciones por año promedio Cuentan con licencia de conducir profesional Alta tasa de aprobación de horas en capacitación Casados
Segmento 2 (23,6%) Empleados nuevos senior	Edad mayor al promedio Poco tiempo trabajado en la empresa Presentan reingreso a la empresa No realizaron estudios en la empresa Nula sindicalización Baja asistencia a capacitaciones Numero de faltas superior al promedio Bajo número de licencias medicas Alto número de capacitaciones por año No cuentan con licencia profesional Baja tasa de aprobación de horas Casados o convivencia
Segmento 3 (26,3%) Empleado promedio	Edad cercana al promedio Tiempo promedio en la empresa Siempre trabajaron en la empresa No realizaron estudios en la empresa Alta sindicalización Alta asistencia a capacitaciones Prácticamente sin faltas dentro de su historial Bajo número de licencias médicas Bajo número de capacitaciones No cuenta con licencia de conducir profesional Alta tasa de aprobación de horas Soltero
Segmento 4 (31,6%) Empleado nuevo joven	Edad menor al promedio Poco tiempo en la empresa Siempre trabajaron en la empresa No realizaron estudios en la empresa Tasa de sindicalización promedio Baja asistencia a capacitaciones Prácticamente sin faltas dentro de su historial Bajo número de licencias médicas Alto número de capacitaciones Cuenta con licencia de conducir profesional Baja tasa de aprobación de horas Soltero

8.7. Tabla estadístico *t*

Two Sided	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	12.71	31.82	63.66	127.3	318.3	636.6
2	4.303	6.965	9.925	14.09	22.33	31.60
3	3.182	4.541	5.841	7.453	10.21	12.92
4	2.776	3.747	4.604	5.598	7.173	8.610
5	2.571	3.365	4.032	4.773	5.893	6.869
6	2.447	3.143	3.707	4.317	5.208	5.959
7	2.365	2.998	3.499	4.029	4.785	5.408
8	2.306	2.896	3.355	3.833	4.501	5.041
9	2.262	2.821	3.250	3.690	4.297	4.781
10	2.228	2.764	3.169	3.581	4.144	4.587
11	2.201	2.718	3.106	3.497	4.025	4.437
12	2.179	2.681	3.055	3.428	3.930	4.318
13	2.160	2.650	3.012	3.372	3.852	4.221
14	2.145	2.624	2.977	3.326	3.787	4.140
15	2.131	2.602	2.947	3.286	3.733	4.073
16	2.120	2.583	2.921	3.252	3.686	4.015
17	2.110	2.567	2.898	3.222	3.646	3.965
18	2.101	2.552	2.878	3.197	3.610	3.922
19	2.093	2.539	2.861	3.174	3.579	3.883
20	2.086	2.528	2.845	3.153	3.552	3.850
21	2.080	2.518	2.831	3.135	3.527	3.819
22	2.074	2.508	2.819	3.119	3.505	3.792
23	2.069	2.500	2.807	3.104	3.485	3.767
24	2.064	2.492	2.797	3.091	3.467	3.745
25	2.060	2.485	2.787	3.078	3.450	3.725
26	2.056	2.479	2.779	3.067	3.435	3.707
27	2.052	2.473	2.771	3.057	3.421	3.690
28	2.048	2.467	2.763	3.047	3.408	3.674
29	2.045	2.462	2.756	3.038	3.396	3.659
30	2.042	2.457	2.750	3.030	3.385	3.646

8.8. Árbol de Decisión: Probabilidad de Renuncia



8.9. Dimensiones de competencias [39]

Competencias	Actividades Clave
Trabajar con seguridad	Identificar condiciones de seguridad.
	Resguardar condiciones de seguridad.
Medir y calcular con exactitud	Planificar y preparar el trabajo de medición.
	Realizar diferentes tipos de mediciones para determinar la magnitud de un objeto.
Identificar el uso adecuado de herramientas y equipos	Identificar las herramientas y equipos que se requieren para llevar a cabo el trabajo.
	Utilizar las herramientas y equipos de manera segura, según se indica en los procedimientos de seguridad establecidos.
Aplicar buenas prácticas medioambientales	Distinguir las operaciones mineras y sus efectos ambientales.
	Reconocer el marco legal vigente medioambiental de la actividad Minera.
Identificar procedimientos de emergencias y primeros auxilios	Identificar procedimientos básicos de toda emergencia.
	Identificar procedimientos de primeros auxilios básicos.
Organizar el trabajo diario	Preparar el trabajo para la realización de una tarea.
	Realizar las tareas encomendadas, considerando la contingencia.
Actuar con compromiso y responsabilidad	Identificar características personales para el trabajo.
	Enfrentar desafíos en el mundo del trabajo.
Comunicarse efectivamente en el lugar de trabajo	Comunicar adecuadamente en forma oral y escrita.
	Ejecutar tareas utilizando medios de comunicación establecidos por la industria de la minería.
Trabajar colaborativamente con los miembros del equipo	Contribuir al logro de las actividades del equipo de trabajo.
	Tratar en forma efectiva los problemas y conflictos del equipo de trabajo.