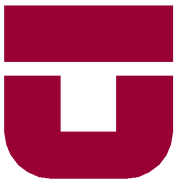


UNIVERSIDAD DE TALCA
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA CIVIL INDUSTRIAL

PROYECTO DE TITULO

PREDICCIÓN DE FUGA DE TRABAJADORES UTILIZANDO MINERIA DE DATOS

AUTOR:
Eliseo Fernando Palma Valdés

PROFESOR TUTOR:
Daniel Hormazábal Ocampo

CURICÓ - CHILE
AGOSTO DE 2021

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su encargado Biblioteca Campus Curicó certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Curicó, 2023

AGRADECIMIENTOS

En primera instancia, quiero agradecer a mi familia que tanto me ha dado en la vida y que pudo hacer posible llegar a esta instancia académica que me encuentro. Agradecer a mi papá que siempre me ha apoyado en los momentos buenos y difíciles. A mi mamá que siempre se ha preocupado que no falte nada. A mi hermana mayor que siempre me aconseja y está disponible con algo que necesite. A mi sobrina Agustina que llego hace poco a nuestras vidas y que ha sido importante en esta etapa. A mi cuñado Eduardo por cuidar y querer a mi hermana. A mi familia materna y paterna, por los lazos que se han formado con el tiempo.

Agradecer también a la tía Lidia y al tío Carlos que me acogieron en Curicó, para que yo fuera a la Universidad y pudiera estudiar mi carrera. A mis amigos Thomas B., Cristopher G., Julio R., Esteban M., Rodrigo C., Iván R., Carlos U., Bastián R., Diego R., Diego L., Fernando C. y Alejandro V. que conocí en la Universidad, los cuales se crearon lazos y se les aprecia mucho. También a los amigos de la educación básica y media, Francisco O., Javier A., Víctor H., Kevin A., Rodrigo T. y Felipe G. los cuales también fueron importantes en este proceso e igualmente se les aprecia mucho. También a los demás amigos que pude compartir en estos procesos fundamentales en mi vida.

Agradecer también a don José Antonio Akel y al equipo de Talana, que me ayudaron a realizar este proceso de proyecto de título y darme la oportunidad de trabajar con ellos. También a mi profesor guía, el profesor Daniel Hormazábal, el cual, a pesar de no conocernos anteriormente, me ayudó mucho en la confección y realización del presente proyecto. A los profesores que conocí a lo largo de la carrera los cuales aprendí de ellos.

Muchísimas gracias a todos por estar en mi vida y poder hacer esto realidad.

RESUMEN EJECUTIVO

En el presente proyecto de título se llevó a cabo la construcción de un modelo que predice la renuncia de los trabajadores de los clientes-empresas de Talana. El proyecto se llevó a cabo a través de la aplicación de la metodología LTDM, estableciendo ciclos de trabajo e iteraciones que permiten una flexibilidad al momento de construir los modelos, volviendo hacia etapas anteriores para ser modificadas y analizadas.

Debido a lo anterior, se define un marco teórico que contempla principalmente elementos referentes a la minería de datos, *machine learning* y a la tecnología de información. Por consiguiente, se realizó un diagnóstico de la situación actual referente a la fuente de los datos, es decir, a cómo se obtienen para que la empresa trabaje con ellos, con el fin de entender cómo se estructuran, por ende, conocer de su estado y naturaleza. Se procedió a ejecutar la metodología, empezando por entender el modelo de negocio, recopilar y entender los datos, para luego limpiarlos y filtrarlos. Luego se procede a la construcción de los modelos, evaluando en cada iteración, cual superaba a los demás en su precisión para predecir la clase objetivo de renuncia. Con ello, se mejora la calidad de los datos de entrada, se optimizan parámetros y se agrega información.

Finalizando con la evaluación de impacto del proyecto, principalmente a nivel económico, estableciendo el beneficio por cargo y de manera global que generaría su implementación. Además, se analiza el impacto social sobre cómo las medidas que pueden ser tomadas gracias a las predicciones optimizan el trabajo de los empleados a nivel psicológico y como afecta en sus relaciones dentro y fuera del empleo.

Eliseo Fernando Palma Valdés (epalma13@alumnos.utalca.cl)
Estudiante Ingeniería Civil Industrial - Universidad de Talca
Agosto de 2021

ÍNDICE DE CONTENIDOS

INTRODUCCIÓN 1
CAPÍTULO 1: INTRODUCCIÓN 2
1.1. Lugar de aplicación 3
1.1.1. Modelo de negocio 3
1.1.2. Visión 4
1.1.3. Misión 4
1.1.4. Valores 4
1.1.5. Organigrama 4
1.1.6. Características del servicio 5
1.2. Propuesta de mejora 6
1.3. Objetivo general 7
1.4. Objetivos específicos 7
1.5. Resultados tangibles esperados 7
CAPÍTULO 2: MARCO TEÓRICO Y METODOLOGÍA 9
2.1. Marco teórico 10
2.1.1. Big data 10
2.1.2. Ciencia de datos 10
2.1.3. Minería de datos 11
2.1.4. Clasificacion supervisada 11
2.1.5. Fuga de trabajadores 15
2.2. Metodología de solución 17
2.2.1. CRISP-DM 18
2.2.2. Design Thinking 19
2.2.3. Propuesta metodológica 21
2.3. Secuenciación de actividades 22
CAPÍTULO 3: ANÁLISIS Y DIAGNÓSTICO 23
3.1. Obtención de los datos 24
3.1.1. Servicio web 24
3.1.2. *Endpoints* 26
3.2. Calidad de los datos 28
3.2.1. Estado de los datos 31

CAPÍTULO 4: ESTRUCTURA DE MEJORA 36
4.1. Herramientas 37
4.1.1. Open Refine 37
4.1.2. KNIME 38
4.1.3. Power BI 41
CAPÍTULO 5: APLICACIÓN DE LA METODOLOGIA 43
5.1. Primera iteración 44
5.1.1. Preparación de datos 44
5.1.2. Construcción y evaluación del modelo 50
5.1.3. Evaluación 51
5.2. Segunda iteración 51
5.3. Tercera iteración 53
5.4. Cuarta iteración 55
5.5. Quinta iteración 57
5.6. Segmentación 61
5.6.1. Estratégico 62
5.6.2. Táctico 64
5.6.3. Operativo 65
5.7. *Dashboard* de validación 67
5.8. Propuesta de implementación 68
5.9. Predicción en fecha 70
CAPÍTULO 6: EVALUACION DE IMPACTO 74
6.1. Impacto económico 75
6.2. Impacto social 78
CONCLUSIONES 81
BIBLIOGRAFÍA 84
ANEXOS 86

ÍNDICE DE FIGURAS

Ilustración 1: Organigrama de la empresa 4
Ilustración 2: Grafico de K vecinos más cercanos 12
Ilustración 3: Representación del árbol de decisión 12

Ilustración 4: Representación del árbol de decisión 13
Ilustración 5: Representación del árbol de decisión 13
Ilustración 6: Representación de una red neuronal 14
Ilustración 7: Encuesta de metodologias usadas en mineria de datos 18
Ilustración 8: Fases de CRISP-DM 18
Ilustración 9: Fases de design thinking 20
Ilustración 10: Diagrama BMPN de *Web Service* 25
Ilustración 11: Diagrama BMPN de la aplicación del proyecto 25
Ilustración 12: Categorización de variables en fuga de trabajadores 29
Ilustración 13: Variables a utilizar 29
Ilustración 14: Variables representadas 30
Ilustración 15: Grafico de causa de fuga 31
Ilustración 16: Grafico de datos faltantes 32
Ilustración 17: Grafico de sueldo base vs causa de fuga 33
Ilustración 18: Diagrama de cajas y bigotes del sueldo base 33
Ilustración 19: *Dashboard* de los datos 34
Ilustración 20: *Dashboard* de los datos filtrados 35
Ilustración 21: Interfaz de Open Refine 37
Ilustración 22: Interfaz de KNIME 38
Ilustración 23: Ciclo de minería de datos en KNIME 39
Ilustración 24: Cuadrante mágico de Gartner en *softwares* de minería de datos 41
Ilustración 25: Interfaz de Power BI 41
Ilustración 26: Cuadrante mágico de Gartner en *software*s de inteligencia de negocio 42
Ilustración 27: Muestra de los valores de “Cargo” post limpieza 45
Ilustración 28: Valores del atributo "Cargo" 45
Ilustración 29: Nodo "*Column Filter*" 46
Ilustración 30: Nodo "*Missing value*" 47
Ilustración 31: Nodos "*Math Formula*" 48
Ilustración 32: Valores del atributo “Motivo término” 48
Ilustración 33: Nodo *“String Replace (Dictionary)"* 48
Ilustración 34: Grafico de fuga 49
Ilustración 35: Nodos que tratan con fechas 49
Ilustración 36: Nodos de clase y filtro 50
Ilustración 37: Nodo de partición 50

Ilustración 38: Nodo de entrenamiento 50
Ilustración 39: Nodo de predicción 50
Ilustración 40: Nodo constructor de la matriz de confusión 51
Ilustración 41: Nodo de filtrado de "Despidos" 52
Ilustración 42: Curva ROC para los modelos 52
Ilustración 43: Nodo de correlación lineal 53
Ilustración 44: Valores de tipo de contrato 55
Ilustración 45: Grafico de correlaciones 57
Ilustración 46: Nodos de validación cruzada 58
Ilustración 47: Árbol de decisión general resultante (muestra referencial) 59
Ilustración 48: Diagrama de cajas y bigotes del sueldo base post preprocesamiento 61
Ilustración 49: Árbol de decisión del conjunto estratégico (muestra referencial) 62
Ilustración 50: Diagrama de cajas y bigotes del sueldo base estratégico 63
Ilustración 51: Árbol de decisión del conjunto táctico (muestra referencial) 64
Ilustración 52: Diagrama de cajas y bigotes del sueldo base operativo 65
Ilustración 53: Árbol de decisión del conjunto operativo (muestra referencial) 66
Ilustración 54: Diagrama de cajas y bigotes del sueldo base táctico 67
Ilustración 55: *Dashboard* de *testing* 68
Ilustración 56: Icono para configurar el espacio de trabajo 68
Ilustración 57: *KNIME Explorer* 69
Ilustración 58: Mostrar código API 69
Ilustración 59: Mostrar código 70
Ilustración 60: Código POST 70
Ilustración 61: Preparación de datos para predecir fecha de renuncia 71
Ilustración 62: Nodos de AutoML 71
Ilustración 63: Nodos de *Deep Learning Keras* 72
Ilustración 64: Optimización de capas ocultas 72

ÍNDICE DE TABLAS

Tabla 1: Revisión de literatura 16
Tabla 2: Resultados de precisión de la iteración 1 51
Tabla 3: Resultados de precisión de la iteración 2 52
Tabla 4: Resultados de correlación de las variables entre si 54

Tabla 5: Resultados de precisión de la iteración 3 54
Tabla 6: Resultados de precisión de la iteración 4 56
Tabla 7: Resultados de precisión de la iteración 5 57
Tabla 8: Resultados de la validación cruzada 58
Tabla 9: Ahorro en entrenamiento de cargos operativos 76
Tabla 10: Ahorro en entrenamiento de cargos tácticos 77
Tabla 11: Ahorro en entrenamiento de cargos operativos 77

ÍNDICE DE ANEXOS

Anexo 1: *Endpoint* de "Persona" 86
Anexo 2: Objeto “Detalle” proveniente de “Persona” 86
Anexo 3: Objeto “Permiso” proveniente de “Persona” 86
Anexo 4: *Endpoint* de "Contratos" 86
Anexo 5: *Endpoint* de "Contratos v2" 88
Anexo 6: *Endpoint* de "Contratos resumido" 89
Anexo 7: *Endpoint* de "Cargos" 90
Anexo 8: *Endpoint* de "Vacaciones" 90
Anexo 9: *Endpoint* de "Vacaciones resumido" 90
Anexo 10: *Endpoint* de "Ausentismo" 90
Anexo 11: *Endpoint* de "Ausentismo resumido" 91
Anexo 12: *Endpoint* de "Prorrateo por centro de costo" 91
Anexo 13: *Endpoint* de "Centralización contable" 91
Anexo 14: *Endpoint* de "Asignación de ítems de pago" 92
Anexo 15: *Endpoint* de "Descarga de documentos de personas" 92
Anexo 16: *Endpoint* de "Creación de documentos de personas" 92
Anexo 17: *Endpoint* de "Días administrativos" 92
Anexo 18: *Endpoint* de "Días administrativos resumido" 92
Anexo 19: *Endpoint* de "Enrolamiento de firma digital" 93
Anexo 20: *Endpoint* de "Solicitud de firma digital" 93
Anexo 21: *Endpoint* de "Solicitud de firma digital" asociado al registro 93
Anexo 22: *Endpoint* de "Asignación de personas a turnos” 93
Anexo 23: *Endpoint* de "Inyección y visualización de marcas” 93
Anexo 24: *Endpoint* de “Días trabajados por contrato” 93

Anexo 25: *Endpoint* de “Turnos y horarios asignados por trabajador”.....94
Anexo 26: Objeto “Asignación” de “Turnos y horarios asignados por trabajador”.....94
Anexo 27: Actividades 94

INTRODUCCIÓN

Es recurrente que las empresas generen una gran cantidad de datos de distinto tipo, los cuales son almacenados en grandes bases de datos, con la finalidad de utilizarlos posteriormente para realizar consultas o estudios. Esto es gracias a la revolución digital y al uso de la tecnología que ha habido en los últimos tiempos, permitiendo almacenar información en distintos procesos que conlleva su utilización. Dicho progreso ha llevado a querer descubrir información en los datos que no está a simple vista, de esto se encarga la minería de datos, la cual posee distintas técnicas con el poder de realizar distintos análisis y descubrimientos dependiendo de la naturaleza de los datos, es decir, datos de tipo numérico, texto, imagen, video, entre otros.

La empresa relacionada a este trabajo ve con buenos ojos la factibilidad de utilizar este tipo de técnicas, específicamente para la predicción de los trabajadores que renuncian de sus puestos, esto les ayudaría a resolver problemas en el ámbito de recursos humanos a sus clientes, los cuales resultan en pérdidas de tiempo, esfuerzo y dinero.

Por ello, se detalla la construcción de un modelo de predicción utilizando minería de datos, que a primera impresión presenta un gran potencial por ser utilizado, generando más información para los clientes de la empresa en termino de sus empleados. No obstante, dicha elaboración conlleva una serie de etapas que deben ser definidas, además de realizar una evaluación financiera que permitirá saber cómo afectará el proyecto en la organización.

CAPÍTULO 1: INTRODUCCIÓN

En el presente capitulo se introduce el proyecto, informando sobre el lugar, la problemática, objetivo general, objetivos específicos y los resultados esperados.

1.1. Lugar de aplicación

Talana es una empresa creada a partir de una idea de negocio cubriendo una necesidad de mercado, el cual fue implementada en el año 2016 por dos ingenieros de vasta experiencia. Su propósito es gestionar y mantener un *software* que simplifica procesos administrativos, específicamente en el área de recursos humanos de las empresas, automatizándolos y digitalizando los documentos relacionados.

Por ello, el *software* tiene una interfaz simple y usable, tanto trabajadores como jefes pueden tener en su plataforma un registro con todos los datos y trámites de cada miembro de la compañía. Desde el horario de llegada, el cual queda marcado una vez que el trabajador se enrola en la aplicación mediante un sistema GPS que ubica la sucursal a la que se accede, hasta la última vez que tuvo un aumento de sueldo o la fecha en que solicitó vacaciones (Información de Mercados, 2017).

La empresa en la actualidad cuenta con alrededor de 100 colaboradores alineados por valores comunes que transmitidos permanentemente. Además, es líder en el mercado en cuanto a transformación digital, trabajando con más de 2.000 empresas (Talana, 2021) y ha sido destacada por los medios, elaborando una variedad de estudios y encuestas de sus clientes, de los cuales, las más influyentes han sido sobre la forma de trabajo remoto y virtual, producto de la pandemia que afecta actualmente al mundo.

1.1.1. Modelo de negocio

El modelo de negocio de la empresa consiste en un SaaS (*Software as a Service*), lo que significa que no se realiza un *software* a medida, es decir, que se cobra por el servicio al acceso de la aplicación que la empresa posee. La infraestructura se encuentra en la red de la empresa y no en sus clientes, garantizando la disponibilidad y seguridad de la información que estos poseen. Para ejemplificar lo mencionado, los clientes contratarían un servicio similar a como funciona Netflix, Spotify, entre otros. Algunos de las funciones sobre la plataforma que la empresa debe cumplir, en relación con el tipo de servicio que entregan, son; mantenerlo actualizado, mejorarlo continuamente, entregar soporte técnico, actualizar parámetros o funciones de acuerdo con cambios legales y garantizar el respaldo de los datos.

1.1.2. Visión

"Hacer más felices a las personas, rediseñando la forma de trabajar."

1.1.3. Misión

"Utilizar la tecnología como un catalizador del cambio, permitiendo acortar brechas y mejorar las relaciones internas de nuestros clientes."

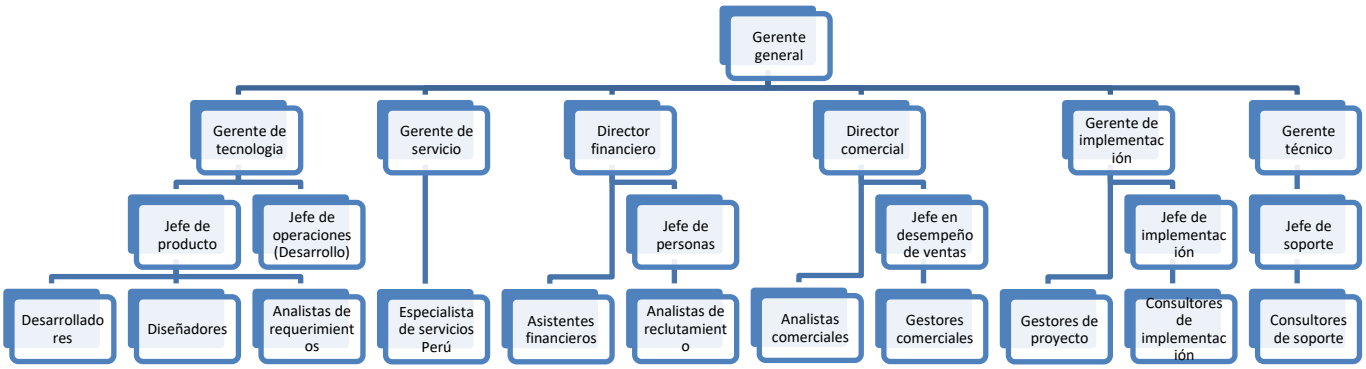
1.1.4. Valores

- *"Cuidar el karma"*
- *"Pensar y tomar la iniciativa"*
- *"Contagiar alegría"*
- *"Pensar en el equipo"*
- *"Crear impacto positivo"*

1.1.5. Organigrama

A continuación, en la Ilustración 1 se presenta el organigrama de la empresa, en donde, se puede apreciar los principales departamentos y roles que se cumplen en el trabajo, detallados jerárquicamente.

Ilustración 1: Organigrama de la empresa



Fuente: Elaboración propia en base a Talana

Cabe señalar, que, por cada gerencia, existe un área en específico, excepto con el gerente de implementación y técnico que están a cargo de subáreas. La empresa es encabezada por el gerente general, el cual tiene a cargo 6 gerentes, estos son:

- **Gerente de tecnología (CTO):** está a cargo del área de producto, el cual se preocupa del mantenimiento y codificación del *software*.
- **Director financiero (CFO):** está a cargo del área de administración, el cual se preocupa del personal y de realizar la facturación y cobranza de la empresa.
- **Director comercial (CMO):** está a cargo del área comercial, el cual se preocupa de la mercadotecnia y publicidad de la empresa.
- **Gerente de servicios:** está a cargo el área de operaciones, el cual está dividido en 3 subáreas llamadas implementación *front*, implementación *back* y soporte.
- **Gerente de implementación:** está a cargo de las subáreas de implementación *back* y *front*.
- **Gerente técnico:** está a cargo de la subárea de soporte, el cual se preocupa de resolverle los problemas a los clientes que puedan tener con la aplicación.

1.1.6. Características del servicio

La aplicación ofrece una variedad de módulos que las empresas pueden ocupar para gestionar sus áreas de recursos humanos, estos módulos son:

- **Gestión de personas:** acá, la documentación de cada trabajador es administrada, pudiendo realizar cambios y mantener su historial dentro de fichas o "ficha de empleado". Todos los módulos restantes dependen de este módulo directamente, ya que es el que permite la entrada de la información con la que los demás operan.
- **Firma digital:** en dicho modulo da la posibilidad de enviar y firmar miles de documentos de forma digital a través de distintos medios como SMS o correo electrónico. Esto permite ahorrar tiempos y reducir costos en papel, ayudando de manera directa al medio ambiente.

- **Remuneraciones:** aquí se calculan los sueldos de los trabajadores, el cual tiene en sus funcionalidades la generación de finiquitos y reliquidaciones. Además, se administran los pagos generados en pymes y corporaciones, diferenciados en cantidad de empleados.
- **Control de asistencia y turnos:** este módulo permite la creación de turnos, asignación de turnos, reemplazos, corrección de marcas, traspaso de marcas a remuneraciones y la obtención de reportes sobre la asistencia que presentan los empleados. Además, permite la administración de horas extras que se realizan, permitiendo ser adicionadas en el módulo de remuneraciones para facilitar los cálculos salariales.
- **Comunicaciones:** en el último modulo, da la posibilidad de comunicar noticias y anuncios de la empresa de forma segmentada o general. Además, admite medir el clima de la empresa por medio de encuestas enviadas automáticamente a los trabajadores.

1.2. Propuesta de mejora

La oportunidad de mejora nace de la necesidad de las empresas que utilizan el servicio de Talana, de conocer más información sobre sus trabajadores, con el fin de saber si realizan de buena forma la gestión de su personal. Una de las consecuencias de un posible mal manejo en este ámbito, son las renuncias de los trabajadores, lo cual es un problema considerable por parte de aquellos que tienen personal experimentado y de reputación, ya que, es difícil encontrar un reemplazante en tal caso. Cerca de 87.750 trabajadores en total, se registró su salida de sus empresas antes de la pandemia, de ese valor, 48.930 trabajadores renunciaron, demostrando que es un problema para considerar. Probablemente, si un empleado renuncia, se encontraría en medio de proyectos y trabajos que terminan siendo afectados a largo plazo, dejando insatisfechos a los clientes. En adición, para reclutar personal se requieren de tiempo y costos adicionales que la empresa tendrá que realizar y que no se tiene previsto con antelación, esto considerando su curva de aprendizaje, es decir, su adaptación en el puesto para lograr la productividad requerida.

Considerando lo anterior, se hace importante la confección de un modelo que pueda predecir la fuga de los trabajadores para las empresas, para así, transformar los datos en dinero o en mayor demanda del servicio. Esto es posible gracias a las técnicas de minería de datos que existen actualmente, como, por ejemplo, *machine learning* o algún tipo de regresión.

1.3. Objetivo general

Desarrollar un modelo de predicción de fuga de trabajadores, en función de los registros históricos que existen, mediante técnicas de minería de datos, para los clientes que utilizan Talana.

1.4. Objetivos específicos

Los objetivos específicos, que ayudan a cumplir con el objetivo general, son:

- Realizar un diagnóstico de la situación actual con tal de conocer de donde provienen los datos y su estado, así como de las problemáticas y necesidades del negocio, de forma tal de plantear hipótesis adecuadas.
- Aplicar la metodología LDTM que permita desarrollar un modelo para pronosticar la fuga de trabajadores, segmentados por características a definir en base al análisis y entendimiento de los datos.
- Desarrollar una propuesta para la implementación del modelo en forma sistémica para su uso a lo largo tiempo.
- Realizar una evaluación financiera y social de factibilidad para la incorporación de la propuesta realizada.

1.5. Resultados tangibles esperados

Los resultados tangibles esperados del proyecto son los siguientes:

- Un conjunto de archivos (.knwf) que ejecuten un proceso automatizado para la generación de un área de datos tratados en la empresa, donde almacenar los datos extraídos, transformados y limpiados.
- Un repositorio de datos (dirección) desde el cual desarrollar un modelo de predicción, donde los datos sean de calidad y adecuados al propósito, esto ayudara también a realizar de buena manera el entendimiento de los datos.

- Una documentación (.docx) para el entendimiento de lo realizado, ya que, se debe documentar de manera visual y analítica las hipótesis en las diferentes etapas del proyecto.
- Un modelo predictivo (.knwf) que permita pronosticar la probabilidad de que un trabajador renuncie a su empleador, el cual estará optimizado y validado.

CAPÍTULO 2: MARCO TEÓRICO Y METODOLOGÍA

En el siguiente capítulo se presenta el modo en el cual se aborda la problemática planteada, mencionando los elementos teóricos y técnicos a utilizar, junto con la metodología de solución.

Marco teórico y metodología

2.1. Marco teórico

Para un desarrollo correcto del proyecto a realizar, se deben establecer los conceptos técnicos y teóricos de la mejora a realizar. A continuación, se definen estos elementos, los cuales conforman el marco teórico.

2.1.1. Big data

Para poder mencionar los temas que están presentes en este marco, primero se debe definir el concepto de *big data* y su importancia. El *big data* se ha definido en la existencia de grandes cantidades de datos, específicamente, en un activo de información de alto volumen, alta velocidad y variedad, que exigen formas rentables e innovadoras de procesamiento de la información para una visión mejorada y la toma de decisiones (Gartner, 2020).

El *big data* es importante para muchas empresas, debido a que proporciona respuestas a muchas preguntas que estas ni siquiera sabían que tenían, en otras palabras, proporciona un punto de referencia. Con una cantidad tan grande de información, los datos pueden ser moldeados o probados de cualquier manera que la empresa considere adecuada. Al hacerlo, las organizaciones son capaces de identificar los problemas de una forma más comprensible.

2.1.2. Ciencia de datos

La ciencia de datos o también llamado *data science*, se encarga de estudiar y representar la información con tal de poder convertir los recursos en estrategias y negocios para las empresas. Por ello, se busca la extracción de grandes cantidades de datos (*big data*) para identificar relaciones entre variables y crear modelos, para así, ayudar a las organizaciones a aumentar su eficiencia, identificar nuevas oportunidades en el mercado y obtener una ventaja competitiva con el resto de las organizaciones (U. de Alcalá, 2021). Sabido lo anterior, podemos decir que la ciencia de datos es un campo de múltiples disciplinas, el cual integra el cálculo matemático, estadística, minería de datos y *machine learning.*

2.1.3. Minería de datos

El concepto de minería de datos corresponde a una de las etapas del proceso llamado *Knowledge Discovery in Databases* o *KDD*. Está conformado por un conjunto de técnicas y algoritmos que sirven para hacer análisis de conjuntos de datos, extrayendo patrones y relaciones entre ellos, convirtiéndolos en información valiosa y útil para quienes toman las decisiones. Estas técnicas y algoritmos se implementan y se comparan entre si con el fin de obtener buenos resultados. Los algoritmos que destacan son los de regresión, clustering y clasificación.

El uso de minería de datos se debe entender como un apoyo para los analistas, y no reemplaza al conocimiento que tienen los expertos del negocio, ni elimina la necesidad de entender los datos. El *Data Mining* no funciona por sí sólo, ya que los patrones que se encuentren en los datos deben ser interpretados y validados para ver si responden a las consultas del negocio, y si son aplicables en el mundo real (Martínez, 2012).

2.1.4. Clasificacion supervisada

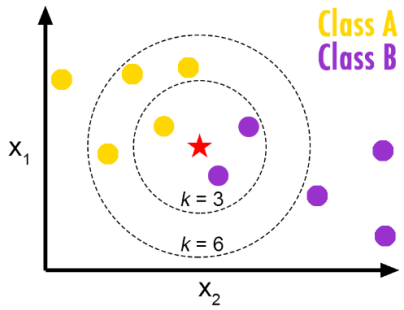
En minería de datos y el *machine learning*, los modelos de clasificacion buscan encontrar un sistema que sea capaz de identificar automáticamente para cada registro, la clase a la cual pertenece. La clase refiere a categorías arbitrarias según el tipo de problema.

Para desarrollar este tipo de predicción se debe tener un conjunto de entrenamiento, con los datos ya clasificados previamente y el modelo que se requiere implementar. Con ello se podrá obtener un modelo entrenado y ajustado para poder ser utilizado posteriormente. También de un conjunto de prueba, que permitirá probar el modelo con datos nuevos. En esta área, aparecen distintos algoritmos que realizan la construcción de los modelos de clasificacion, como lo son, los *k* vecinos más cercanos, arboles de decisión, máquina de soporte vectorial, redes neuronales y el clasificador bayesiano.

- ***K* vecinos más cercanos:** Este método, también llamado KNN, estima la función de densidad $F(x/C_j)$ de la predictora x por cada clase C_j. Es no paramétrico, estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase C_j a partir de la información

proporcionada por el conjunto de ejemplos. En el proceso de aprendizaje, no se hace ninguna suposición acerca de la distribución de las variables predictoras (Parra, 2019). En la Ilustración 2 se aprecia el grafico resultante de este método.

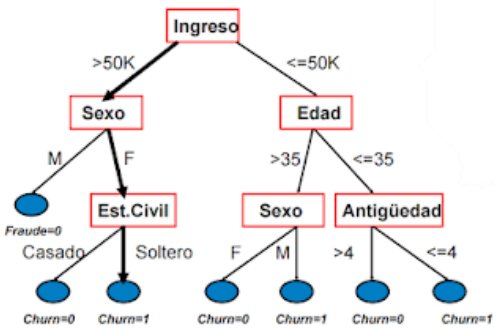
Ilustración 2: Grafico de K vecinos más cercanos



Fuente: (Salcedo, 2018)

- **Árboles de decisión:** Los árboles de decisión son una técnica explicativa y des composicional que utiliza un proceso de división secuencial, iterativo y descendente que, partiendo de una variable dependiente, forma grupos homogéneos definidos específicamente mediante combinaciones de variables independientes en las que se incluyen la totalidad de los casos recogidos en la muestra. En los árboles de decisión se encuentran los siguientes componentes: nodos, ramas y hojas. Los nodos y el nodo raíz son las variables de entrada, las ramas representan los posibles valores de las variables de entrada y las hojas son los posibles valores de la variable de salida (Parra, 2019). En la Ilustración 3 se aprecia la representación de este método.

Ilustración 3: Representación del árbol de decisión

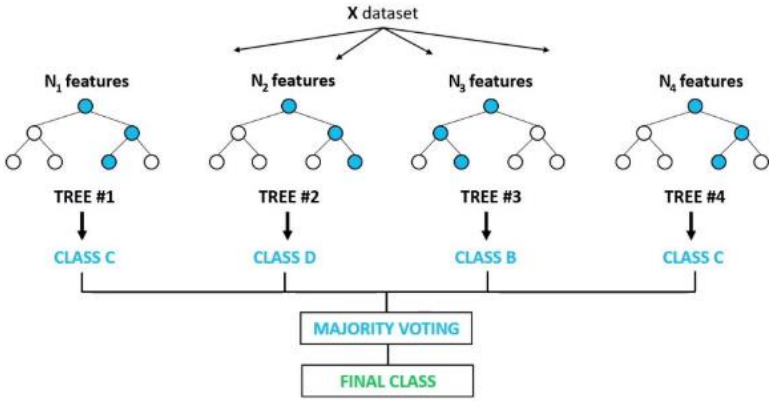


Fuente: (Aprendesas, 2015)

- **Bosques aleatorios:** este modelo es un conjunto (*ensemble*) de árboles de decisión combinados con *bagging*. Al usar *bagging*, lo que en realidad está pasando, es que

distintos árboles ven distintas porciones de los datos. Ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor (Martinez, 2020). En la Ilustración 4 se aprecia la representación de este método.

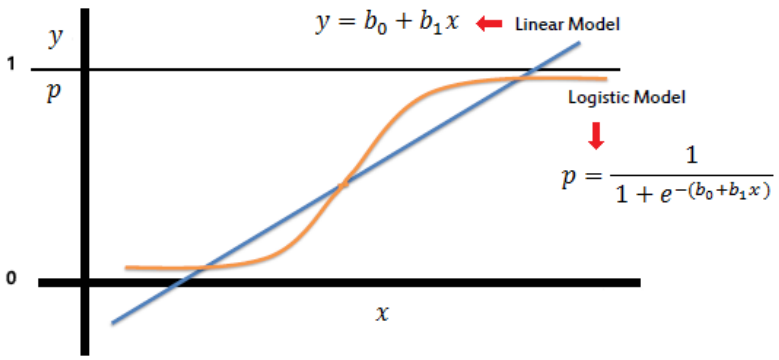
Ilustración 4: Representación del árbol de decisión



Fuente: (Cardellino, 2021)

- **Regresión logística**: La regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de predictores. Es similar a un modelo de regresión lineal, pero está adaptado para modelos en los que la variable dependiente es dicotómica. Los coeficientes de regresión logística pueden utilizarse para estimar la razón de probabilidad de cada variable independiente del modelo. La regresión logística se puede aplicar a un rango más amplio de situaciones de investigación que el análisis discriminante (IBM, s.f.). En la Ilustración 5 se aprecia la representación de este método.

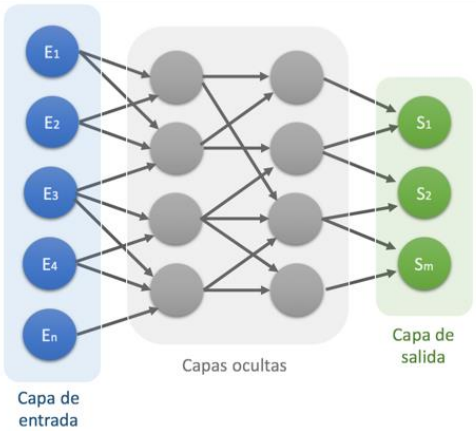
Ilustración 5: Representación del árbol de decisión



Fuente: (Rodriguez, 2018)

- **Redes neuronales artificiales:** Una Red Neuronal Artificial (RNA) es un modelo matemático inspirado en el comportamiento biológico de las neuronas y en cómo se organizan formando la estructura del cerebro. Las redes neuronales intentan aprender mediante ensayos repetidos como organizarse mejor a sí mismas para conseguir maximizar la predicción. Un modelo de red neuronal se compone de nodos, que actúan como input, output o procesadores intermedios. Cada nodo se conecta con el siguiente conjunto de nodos mediante una serie de trayectorias ponderadas. Basado en un paradigma de aprendizaje, el modelo toma el primer caso, y toma inicial basada en las ponderaciones. Se evalúa el error de predicción y modifica las ponderaciones para mejorar la predicción (Parra, 2019). En la Ilustración 6 se aprecia la representación de este método.

Ilustración 6: Representación de una red neuronal



Fuente: (Molina, 2020)

- **Clasificador bayesiano:** El clasificador bayesiano o también llamado Naïve Bayes es uno de los clasificadores más utilizados por su simplicidad y rapidez. Se trata de una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados, en base al Teorema de Bayes (Ecuación 1), también conocido como teorema de la probabilidad condicionada.

Ecuación 1: Teorema de Bayes

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Fuente: (Gascó, 2019)

2.1.5. Fuga de trabajadores

La fuga de clientes es una de las tantas temáticas que aborda la minería de datos y el *machine learning.* Tiene como objetivo el de predecir que trabajador tiene más probabilidad de irse de la empresa, ya sea, por despido o renuncia. Por lo tanto, es un beneficio para las organizaciones poder saber este tipo de información, con el fin de tomar medidas preventivas y así reducir costos y tiempo de invertir en nuevo personal. Para poder realizar este estudio de manera adecuada, se hace una revisión de cómo ha sido abordada esta problemática anteriormente, de esta manera, se puede tener una referencia en el procedimiento de este caso.

Para empezar la revisión, primero se tiene el informe llamado, *"Improving employee retention by predicting employee attrition using machine learning techniques"* (Salunkhe, 2018), el cual predice la fuga y mejora la tasa de retención de los empleados valiosos, minimizando el costo de rotación de su empresa. En este caso, para encontrar los atributos idóneos para la predicción, se realizó una revisión de literatura. Se concluye que el algoritmo de regresión logística es la que mejor entrego resultados.

El segundo informe llamado, *"Employee Resignation Prediction Model Based on Machine Learning"* (Dai & Zhu, 2020), también toma referencias para los atributos y los divide en dos grupos de categorías, una es "*personal reasons*" y la otra es "*corporate factors*". El autor define a estas últimas categorías como importantes, ya que, impactan en mayor grado al resultado. En este trabajo, también se concluye que la regresión logística es el mejor modelo para el problema.

En el tercer informe llamado, *"An Approach for Predicting Employee Churn by Using Data Mining"* (Onuralp, 2017), se recolectan los datos de recursos humanos proporcionados por IBM, estos datos tienen una variedad de categorías los cuales filtraron algunas debido a que no eran de utilidad, como, por ejemplo, si cumplían con la mayoría de edad o el ID del empleado. Acá, se obtuvo como mejor modelo a SVM o *support vector machines*.

Tabla 1: Revisión de literatura

Literatura	Atributos	Mejor predictor
(Salunkhe, 2018)	• Edad • Sexo • Estado civil • Desgaste • Aumento salarial porcentual • Ingresos mensuales • Años desde la última promoción • Distancia de casa • Rol laboral • Calificación de rendimiento • Nivel de trabajo • Satisfacción ambiental • Años en el rol actual • Satisfacción de la relación • Años con el gerente actual • Satisfacción laboral • Equilibrio de la vida laboral • Número de empresas trabajadas • Años en la empresa • Tiempo excesivo • Años de trabajo totales	Regresión logística
(Dai & Zhu, 2020)	• Edad • Genero • Estado de salud • Educación superior • Profesión • Residencia actual • Matrimonio • Salario • Departamento • Calificación del trabajo • Posición • Horarios de trabajo diario • Horas extras diarias en promedio • Años en la empresa • Años en el puesto • Número total de empresas en que trabajo	Regresión logística

(Onuralp, 2017)	• Edad • Viaje de negocios • Tarifa diaria • Departamento • Distancia desde la casa • Educación • Campo de educación, • Genero • Satisfacción del medio ambiente • Tarifa por hora, participación laboral • Nivel de empleo • Rol de trabajo • Satisfacción laboral • Estado civil • Ingresos mensuales • Tasa mensual • Número de empresas trabajadas • Tiempo extra • Porcentaje de aumento salarial • Clasificación de rendimiento • Satisfacción de la relación • Nivel de opción • Total de años de trabajo • Tiempos de formación el año pasado • Años en la empresa • Años en el papel actual • Años desde la última promoción • Años con el actual gerente	SVM

Fuente: Elaboración propia en base a (Salunkhe, 2018), (Dai & Zhu, 2020) *y* (Onuralp, 2017)

Sabido lo anterior, se puede tener una idea de cómo afrontar el problema inicialmente en los capítulos posteriores, el cual seguirá una línea en base a dicho problema a abordar. Con ello se planea usar los atributos mostrados que se consideran importantes, es decir, los que afectan al trabajo directamente, ya que, los autores concluyen que estos son utilizados mayoritariamente por los algoritmos, los cuales también dependerá su utilización, de que si la empresa tenga en sus bases de datos dicha información y de que si se pueda acceder.

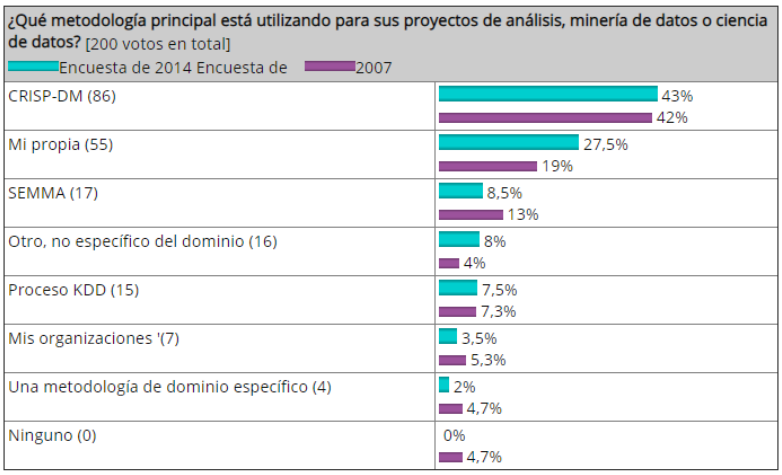
2.2. Metodología de solución

A continuación, se plantea la metodología de solución que utilizara el proyecto, para su correcto desarrollo e implementación posterior

2.2.1. CRISP-DM

CRISP-DM es una metodología en la cual se abordan proyectos de minería de datos. Se le nombra de esta forma por sus siglas en ingles "*Cross Industry Standard Process for Data Mining*" y es una de las más conocidas sobre este tema. Como se puede observar en la Ilustración 7, correspondiente a una encuesta realizada el 2014, esta metodología ha sido mayoritariamente utilizada en los proyectos con un 43% de los votantes.

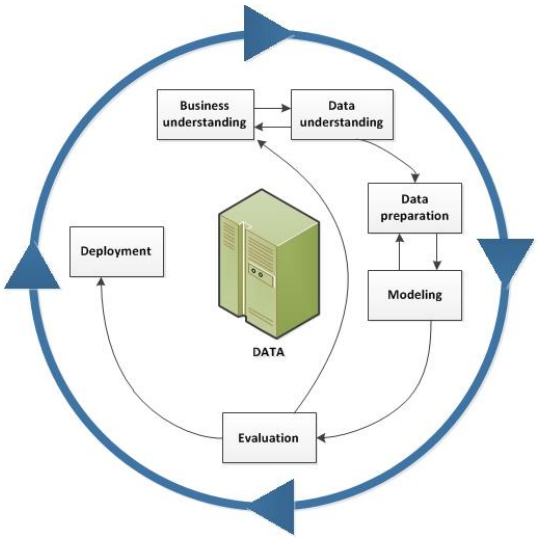
Ilustración 7: Encuesta de metodologias usadas en mineria de datos



Fuente: (KD Nuggets, 2014)

A continuación, en la Ilustración 8 se aprecian las fases de la metodología, las cuales son detalladas posteriormente.

Ilustración 8: Fases de CRISP-DM



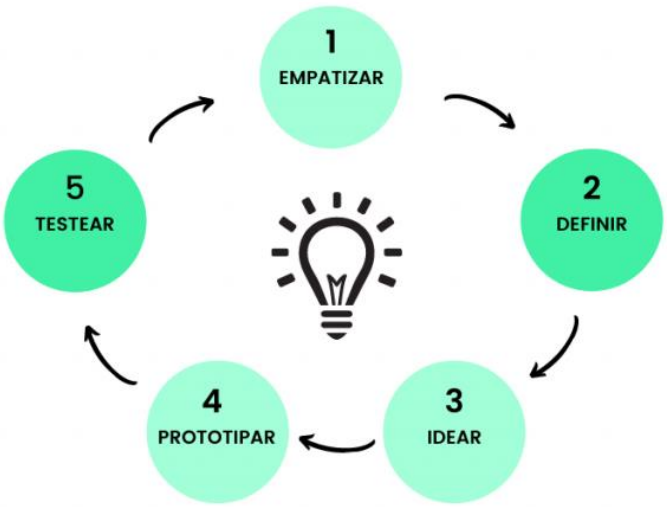
Fuente: (IBM, 2019)

- **Comprensión del negocio:** es la fase inicial, la cual se entienden los objetivos y requerimientos del proyecto. La idea es analizar la situación actual del negocio para poder transformar ese conocimiento en un problema de minería de datos.
- **Comprensión de los datos:** en esta fase se recolectan los datos necesarios provenientes del almacén de datos, verificando la calidad de estos y describiéndolos con tal de conocer su naturaleza. Se crean relaciones, gráficos y un análisis, permitiendo establecer las primeras hipótesis.
- **Preparación de los datos:** acá se preparan los datos con el fin de que se adapten al modelo que se utiliza posteriormente. Las tareas que se realizan comúnmente en dicha etapa son; limpieza de datos, generación o reducción de variables, cambios de formato, entre otros.
- **Modelado:** posteriormente se procede a seleccionar las técnicas de modelado más apropiadas, que se utilizaran para el proyecto. Estos deben disponer de los datos adecuados para su propósito, en donde, se debe procesar la información en un tiempo conveniente para la obtención del modelo. Se debe establecer de antemano, el cómo será la evaluación de la calidad del modelo para confiar en los resultados obtenidos.
- **Evaluación:** a continuación, se evalúa el modelo, verificando si se cumple el objetivo principal, el cual determinara el éxito. Es conveniente analizar el proceso completo, para verificar si existe algún error, ajustarlo y volver a realizar el procedimiento
- **Despliegue:** en la fase final, se transforma el conocimiento proporcionado de los resultados de manera satisfactoria en acciones aplicadas al negocio, esto consiste en recomendaciones por parte del analista y estrategias en su aplicación.

2.2.2. Design Thinking

El Design Thinking o pensamiento de diseño es una metodología de resolución de problemas aplicable a cualquier ámbito que requiera un enfoque creativo. Centra sus esfuerzos en empatizar con el usuario, en generar ideas creativas y en confrontarlas continuamente a través de un prototipo como instrumento de aprendizaje, pensamiento y referencia para la evaluación de soluciones (Romero, 2013). A continuación, en la Ilustración 9 se aprecian las fases de la metodología, las cuales son detalladas posteriormente.

Ilustración 9: Fases de design thinking



Fuente: (Staryfurman, s.f.)

- **Empatizar:** en la primera fase se debe entender y comprender al cliente, es decir, descubrir sus necesidades y lo que es importante para él. Realizar el compromiso de conocerlo bien para llegar a tener una buena relación en el trabajo.
- **Definir:** en esta fase se define el problema a abordar, clarificando la situación que se abordara de manera significativa, posibilitando el diseño de ambas partes. Idealmente, el cliente define el objetivo y uno la necesidad.
- **Idear:** acá se realiza una lluvia de ideas con la finalidad de idear las formas de alcanzar el objetivo definido, recopilando todas estas, desde las más ambiciosas a las más conservadoras, que permitan encontrar soluciones al problema.
- **Prototipar:** la siguiente fase es seleccionar, entre todas las ideas planteadas, con cual se trabaja. Definido lo anterior se prototipa para tener una visualización del diseño.
- **Testear:** finalizando, se testea un plan de acción para la comprobación del éxito, valorizando el aprendizaje.

Estas fases se presentan de forma consecutiva, pero al igual que CRISP-DM se trata de un proceso iterativo, el cual se puede volver hacia atrás por cualquier necesidad que se tenga.

2.2.3. Propuesta metodológica

Se propone una metodología propia en base a otra similar llamada LDTM o *Lean Design Thinking Methodology* (Ahmed, 2018), con la motivación de que la metodología CRISP-DM no está actualizada del todo, ya que, fue creado hace 21 años atrás y desde ese entonces la cantidad de datos y desarrollo han ido en aumento progresivamente, por lo que varios autores están optando de crear sus propias metodologías.

Se decidió combinar las metodologías antes mencionadas, rescatando lo mejor de cada una de ellas para la realización del proyecto. Las etapas de esta son las siguientes:

- **Descubrimiento de trabajo:** se conoce mejor al cliente y comprende sus necesidades u problemas, empatizando con él. Con ello se define la problemática y objetivos del proyecto.
- **Enfoque analítico:** en esta fase se discuten de los métodos o técnicas más adecuadas para la solución en base a literatura previa, es decir, discutir la implementación de clasificación, regresión, entre otros.
- **Recursos de datos:** acá se efectúa la obtención de los datos relevantes para desarrollar el modelo, los cuales serán descritos y representados.
- **Preparación de datos:** en dicha fase, se realiza la limpieza de los datos, combinar variables, transformación de variables, entre otros. Se utiliza estadística y visualización de estos para comprenderlos a cabalidad.
- **Construcción del modelo:** posteriormente se procede a aplicar las técnicas seleccionadas previamente para la resolución del problema y posterior evaluación de la confiabilidad de los resultados.
- **Evaluar modelo:** a continuación, se miden las técnicas utilizadas, evaluando el de mejor rendimiento. En base a los resultados, es posible que se deba volver hacia atrás al haber algún error en el proceso.
- **Aprender y actualizar:** se recopila los comentarios y apreciaciones del cliente para así aprender de lo realizado.
- **Evaluar el impacto:** este paso se realiza para estimar en cuanto afectara el proyecto de manera monetaria.

2.3. Secuenciación de actividades

Para desarrollar el proyecto con la metodología previamente explicada, se debe definir un plan de trabajo, el cual establecerá las actividades correspondientes. Estas actividades son las siguientes:

- **Diagnóstico de la situación actual:** en esta etapa se verifica la fuente de información con tal de conocer cómo se almacenan los datos y de qué forma.
- **Implementación de la propuesta metodológica:** esta etapa será estructurado por la metodología ya establecidos anteriormente.
- **Ajustes finales al proyecto:** en la etapa final del proyecto se espera realizar los ajustes al modelo e informe, realizando a su vez la evaluación de impacto, de tal manera de finalizar el trabajo.

En él Anexo 27 se puede visualizar esta información de mejor manera.

CAPÍTULO 3: ANÁLISIS Y DIAGNÓSTICO

En el presente capítulo se analiza en detalle la problemática y se realiza el diagnóstico de la situación actual de la empresa.

3.1. Obtención de los datos

A continuación, se analiza la fuente de los datos de los trabajadores y como estos se obtienen mediante el uso de la aplicación por parte de las empresas enroladas, así como también la calidad de los que son almacenados históricamente.

3.1.1. Servicio web

Para realizar un diagnóstico a la situación actual de forma adecuada, se procede a recopilar conocimiento con tal de saber cómo funciona el sistema de información de la empresa, esto ayudara a entender la estructuración de los datos que se almacenan y de sus características, ya sean cualitativas o cuantitativas.

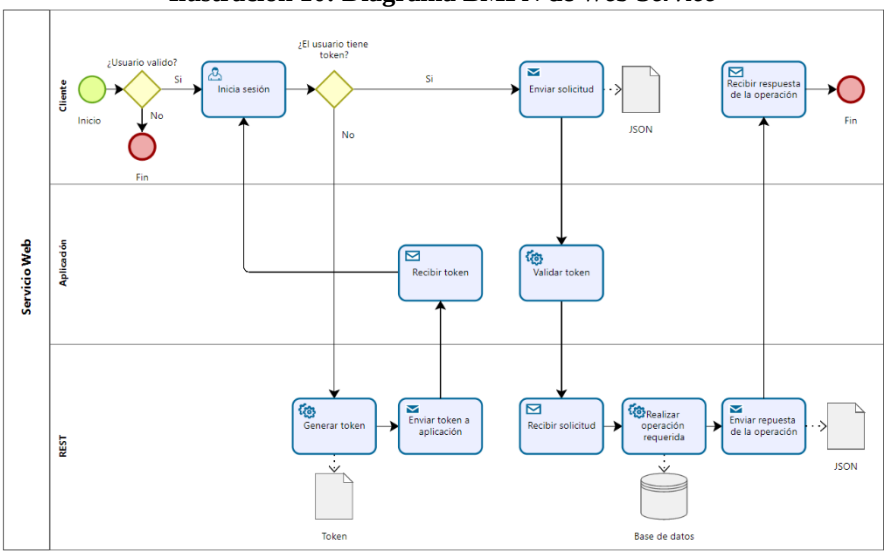
Talana ofrecía la posibilidad de conectarse directamente a la base de datos a sus clientes para interactuar con la información recaudada. Tal mecanismo fue desechado por la empresa, la cual actualmente utiliza el sistema llamado *Web Services Rest*. Con este sistema crea protocolos para la manipulación de la información de una forma estandarizada para todos los clientes. Esto se da, ya que, cuando el cliente realiza una consulta de distinto índole mediante la interfaz de la aplicación, esta se comunica con otra que está alojado en la empresa, la cual se conecta a la base de datos. Tal comunicación se realiza a través de mensaje serializados en formatos específicos de programación como *JSON*, permitiendo la ejecución de la operación requerida sobre los datos.

Las consultas u operaciones que los clientes pueden realizar son cuatro, las cuales son *post, get, put* y *delete,* con ellas es posible obtener, añadir, modificar y eliminar datos respectivamente. Para que cada cliente pueda realizar este tipo de solicitudes, es necesario la generación de un *token* para autenticar cada uno de ellos y poder tener el permiso de realizarlos.

El proceso comienza con el cliente iniciando sesión con un usuario y contraseña, los cuales ya están registrados en el sistema, esto hace que se genere el token automáticamente si este no lo tiene aún, siendo validados previamente por la aplicación. El sistema envía dicho *token* a la aplicación para que se guarde y pueda ser validado automáticamente por cada solicitud del cliente que requiera. Por ello el proceso sigue con enviar la respectiva solicitud al

sistema, la cual la aplicación valida el *token* y el sistema recibe la solicitud, realizando posteriormente la operación requerida, para luego enviar la respuesta a esta solicitud al usuario. En la Ilustración 10 se puede apreciar el diagrama que muestra el proceso que realiza en el sistema descrito.

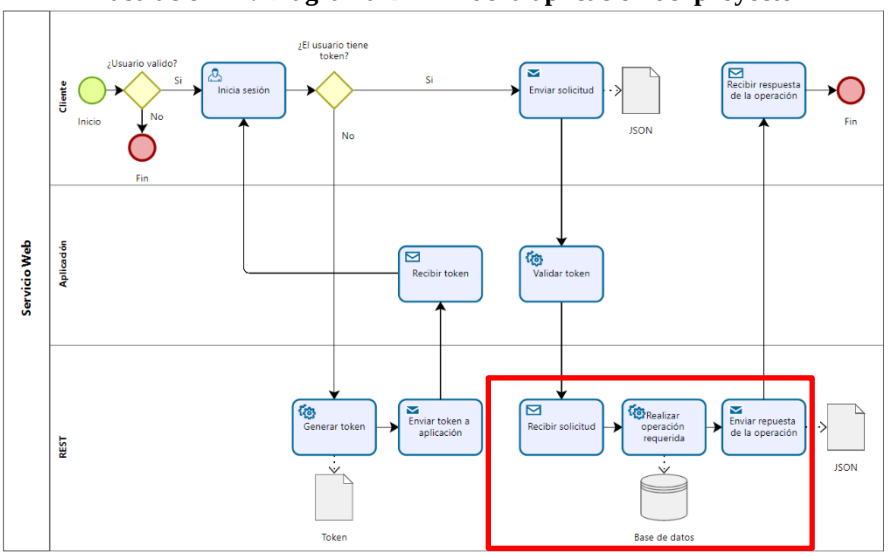
Ilustración 10: Diagrama BMPN de *Web Service*



Fuente: Elaboración propia

Para la elaboración del estudio y aplicabilidad del proyecto en relación con los procedimientos mostrados anteriormente, en la Ilustración 11 se muestra en donde tendrá lugar el estudio y su alcance respectivo.

Ilustración 11: Diagrama BMPN de la aplicación del proyecto



Fuente: Elaboración propia

Mencionado lo anterior, para hacer referencia a los datos hay que detallar la existencia de los *endpoints*, los cuales son las *URLs* de una base de datos o API que responden a una solicitud. Por cada solicitud puede haber más de un *endpoint* sirviendo los datos para llenar cada respuesta que se requiera por parte del cliente en la interfaz.

Cabe destacar que cada solicitud está delimitada por el tipo de información a operar, es decir, computacionalmente, cada objeto tiene sus propios métodos disponibles en los datos. Además, en cada interacción o transferencia de JSON, los objetos se representarán con su estado actual y sus atributos asociados, los cuales describirán sus características, siendo un símil del registro en la base de datos.

3.1.2. *Endpoints*

Los objetivos de este trabajo tienen relación con los trabajadores registrados, por ende, se analizarán los *endpoints* que contengan los datos asociados a ellos de forma directa. El sistema de información actual de la empresa presenta los siguientes datos:

- **Persona:** permite obtener la información de las personas, cada persona puede tener uno o más contratos a lo largo de la relación laboral con una empresa, pero sólo tendrá una ficha laboral en el sistema. Los métodos que se pueden realizar en dicho *endopoint* son *get, post, put*. Además, tiene relacionados dos objetos los cuales se llaman "detalle" y "permiso" para su creación. En el Anexo 1, Anexo 2 y Anexo 3 se puede apreciar los atributos relacionados y su descripción.
- **Contratos**: permite obtener los contratos de los trabajadores y sus condiciones contractuales hoy. Los métodos que se pueden realizar en dicho *endopoint* son *get, post, put* y *patch.* Además, están su símil "contratos v2" y "contratos resumidos", los cuales solo pueden ser usados por el método *get.* En el Anexo 4, Anexo 5 y Anexo 6 se puede apreciar los atributos relacionados y su descripción.
- **Cargos:** permite obtener el cargo de los trabajadores. El método que se puede realizar en dicho *endopoint* es *get*. En el Anexo 7 se puede apreciar los atributos relacionados y su descripción.
- **Vacaciones:** permite obtener las solicitudes de vacaciones de los trabajadores, aprobadas o pendientes de aprobación. Los métodos que se pueden realizar en dicho

endopoint son *get* y *post*. Además, esta su símil "vacaciones resumido" el cual solo puede ser usado por el método *get*. En el Anexo 8 y Anexo 9 se puede apreciar los atributos relacionados y su descripción.

- **Ausentismo:** permite obtener las ausencias por licencias médicas, permisos con o sin goce, entre otros. Los métodos que se pueden realizar en dicho *endopoint* son *get, post* y *put*. Además, esta su símil "Ausentismo resumido" el cual solo puede ser usado por el método *get*. En el Anexo 10 y Anexo 11 se puede apreciar los atributos relacionados y su descripción.
- **Prorrateo por centro de costo:** permite obtener los contratos activos en cierto periodo, junto con la distribución porcentual por centro de costo durante ese periodo. El método que se puede realizar en dicho *endopoint* es *get*. En el Anexo 12 se puede apreciar los atributos relacionados y su descripción.
- **Centralización contable:** permite obtener la centralización contable, pero desagregado a nivel de datos. El método que se puede realizar en dicho *endopoint* es *get*. En el Anexo 13 se puede apreciar los atributos relacionados y su descripción.
- **Asignación de ítems de pago:** permite agregar o actualizar el valor de un ítem de pago para un contrato en el periodo actual, basándose en el rut del empleado o el id del contrato. El método que se puede realizar en dicho *endopoint* es *post*. En el Anexo 14 se puede apreciar los atributos relacionados y su descripción.
- **Descarga de documentos de personas:** permite obtener los documentos de la carpeta de un trabajador. El método que se puede realizar en dicho *endopoint* es *get*. En el Anexo 15 se puede apreciar los atributos relacionados y su descripción.
- **Creación de documentos por personas:** permite agregar documentos a la carpeta de un trabajador. El método que se puede realizar en dicho *endopoint* es *post*. En el Anexo 16 se puede apreciar los atributos relacionados y su descripción.
- **Días administrativos:** permite obtener los días administrativos solicitados por el trabajador. Los métodos que se pueden realizar en dicho *endopoint* es *get* y *post*. Además, esta su símil "días administrativos resumido" el cual solo puede ser usado por el método *get*. En el Anexo 17 y Anexo 18 se puede apreciar los atributos relacionados y su descripción.

- **Enrolamiento a firma digital:** permite obtener la lista de enrolamientos de la empresa. El método que se puede realizar en dicho *endopoint* es *get*. además, esta su símil "solicitudes de firma digital" el cual solo puede ser usado por el método *get*. En el Anexo 19, Anexo 20 y Anexo 21 se puede apreciar los atributos relacionados y su descripción.
- **Asignación de personas a turnos:** permite obtener la asignación de personas a turnos por rango de fechas. Los métodos que se pueden realizar en dicho *endopoint* es *get* y *post*. En el Anexo 22 se puede apreciar los atributos relacionados y su descripción.
- **Inyección y visualización de marcas:** permite obtener las marcas realizadas por los trabajadores. Los métodos que se pueden realizar en dicho *endopoint* es *get* y *post*. En el Anexo 23 se puede apreciar los atributos relacionados y su descripción.
- **Días trabajados por contrato:** permite obtener los días trabajados por contrato, durante un mes específico. El método que se puede realizar en dicho *endopoint* es *get*. En el Anexo 24 se puede apreciar los atributos relacionados y su descripción.
- **Turnos y horarios asignados por trabajador:** permite obtener tiempo en que el trabajador tiene la obligación de marcar y los horarios asociados a cada trabajador en el rango de fecha indicados. El método que se puede realizar en dicho *endopoint* es *get*. En el Anexo 25 y Anexo 26 se puede apreciar los atributos relacionados y su descripción.

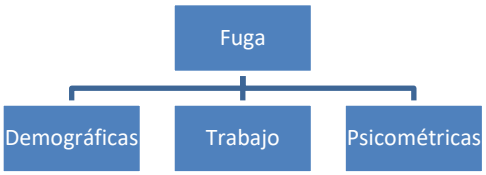
3.2. Calidad de los datos

Antes de precisar la calidad de los datos, es necesario poner en contexto sobre las variables que se utilizan en este tipo de casos, independiente si es que se tienen a disposición. Estas variables se categorizan según el tipo de relación que tiene con el trabajo, el individuo. Se pueden categorizar de la siguiente manera:

- **Demográficos:** se entiende a estos datos como todos aquellos que caracterizan a un individuo en relación con el grupo humano al que pertenece, en un momento determinado.
- **Trabajo:** estos datos refieren a los que tiene relación con el trabajo de la persona y de cómo esta caracterizado.

- **Psicométricos:** se define como los datos pertenecientes a encuestas y estudios psicológicos en relación con el trabajo y de su ambiente.

Ilustración 12: Categorización de variables en fuga de trabajadores



Fuente: Elaboración propia en base a (Jara, 2015)

Para considerar las variables psicométricas, se deben aplicar diversos tipos de encuestas al personal acerca del clima laboral, satisfacción, seguridad, entre otros. En el ámbito laboral y empresarial, estas encuestas son anónimas, por ende, no se tiene un registro histórico de este tipo de variables, los cuales no se consideran en el presente estudio.

Mencionado lo anterior, se detallarán las variables demográficas y de trabajo, para el caso actual, como un apronte inicial a los análisis posteriores que se harán con respecto a estos registros. Cabe destacar que, en algunos casos, habrá que realizar una conversión de los datos existentes para poder llegar a las variables mencionadas. A continuación, se muestra en la Ilustración 13, las variables.

Ilustración 13: Variables a utilizar

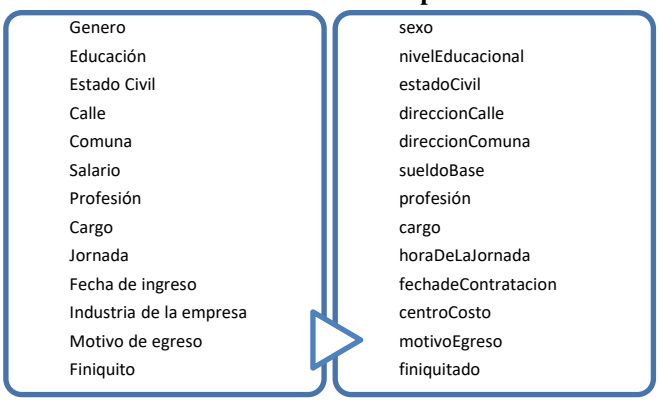
Demograficos	Trabajo
•Edad •Genero •Educación •Estado Civil •Calle •Comuna	•Salario •Profesión •Cargo •Jornada •Distancia •Fecha de ingreso •Industria de la empresa •Motivo de egreso •Finiquito

Fuente: Elaboración propia

Ahora se procede a relacionar estas variables, con las categorías existentes contenidas en los *endpoints*. Estos se representan con un nombre determinado para referenciarlos en los códigos

que utiliza el sistema de servicio para su comunicación. A continuación, en la Ilustración 14 se muestra la representación de estos contenidos en los *endpoints*.

Ilustración 14: Variables representadas



Fuente: Elaboración propia

Las variables tales como edad y distancia, se deben calcular para llegar a ellas. Ahora, se describen las categorías de forma cualitativa y cuantitativa.

- **Género:** corresponde al sexo de la persona, la que puede ser hombre o mujer. Esta variable se relaciona con "sexo" y es de tipo *string*.

- **Educación:** corresponde al nivel de estudios de la persona. Esta variable se relaciona con "nivelEducacional" y es de tipo *string.*

- **Estado Civil:** corresponde a la situación sentimental de la persona, la que puede ser soltero o casado. Esta variable se relaciona con "estadoCivil" y es de tipo *string*.

- **Calle:** corresponde a la ubicación especifica de la persona. Esta variable se relaciona con "direccionCalle" y es de tipo *string*.

- **Comuna:** corresponde a la ubicación general de la persona. Esta variable se relaciona con "direccionComuna" y es de tipo *string*.

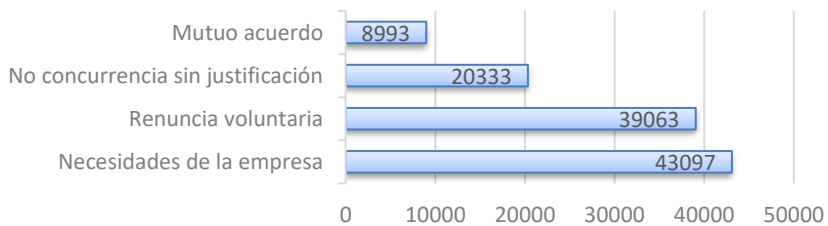
- **Salario:** corresponde al pago por el trabajo que realiza la persona. Esta variable se relaciona con "sueldoBase" y es de tipo *float.*

- **Profesión:** corresponde a la dedicación de la persona. Esta variable se relaciona con "profesion" y es de tipo *string*.
- **Cargo:** corresponde a al puesto que tiene la persona en la empresa. Esta variable se relaciona con "cargo" y es de tipo *string*.
- **Jornada:** corresponde a las horas que tiene que trabajar la persona. Esta variable se relaciona con "horaDeLaJornada" y es de tipo *integer*.
- **Fecha de ingreso:** corresponde al día en que la persona ingreso a la empresa. Esta variable se relaciona con "fechadeContratacion" y es de tipo *datetime*.
- **Industria de la empresa:** corresponde al área que pertenece la empresa en que trabaja la persona. Esta variable se relaciona con "centroCosto" y es de tipo *string*.
- **Motivo de egreso:** corresponde a la descripción de porque renuncio la persona. Esta variable se relaciona con "motivoEgreso" y es de tipo *string*.
- **Finiquito:** corresponde si la persona esta finiquitada o no. Esta variable se relaciona con "finiquito" y es de tipo binaria.

3.2.1. Estado de los datos

Los datos por utilizar provienen en su mayoría de los contratos de los trabajadores, los cuales son registrados por sus empleadores en la interfaz de la aplicación. El tamaño de la muestra de datos registra un total de 225.352 contratos históricos, los que se distribuyen en 606 empresas diferentes. En la Ilustración 15 se muestra las causas de fuga más recurrentes de los trabajadores.

Ilustración 15: Grafico de causa de fuga

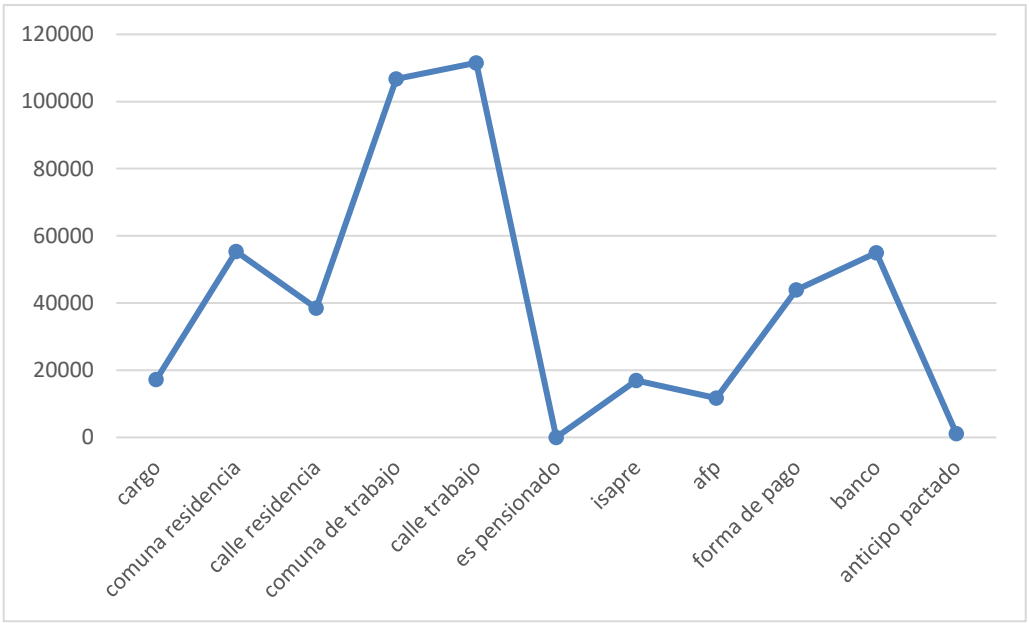


Fuente: Elaboración propia

Como se puede apreciar, el mayor porcentaje de fuga de los trabajadores consiste en un fin de contrato por necesidades de la empresa, es decir, fueron despedidos de su cargo. En modo de hipótesis, esto se puede explicar por los recortes de las empresas causados por el estallido social y la posterior pandemia.

Cabe destacar que se consideran más variables como Isapre, banco, AFP y forma de pago, para verificar su relevancia en el futuro modelo. Por ello, se debe mencionar en cuanto a datos faltantes (*Na's*) hay más de 450 mil aproximadamente, los cuales se detallan a continuación según variables en la Ilustración 16.

Ilustración 16: Grafico de datos faltantes

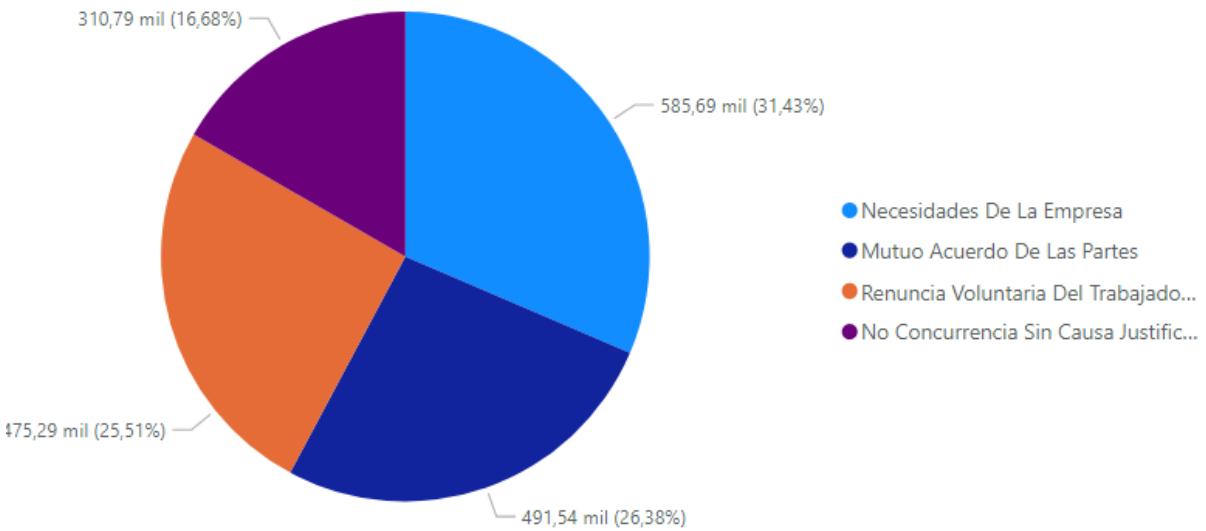


Fuente: Elaboración propia

Las variables "comuna de trabajo" y "calle trabajo" son las que más *Na´s* tienen, estas servirán para calcular la distancia de la casa al trabajo del empleado, pero al tener demasiados datos faltantes, probablemente se deberá descartar esta opción.

También se analiza el sueldo promedio por trabajador y la causa de fuga en la Ilustración 17, el cual se puede inferir que los que ganan menos o tienen un sueldo promedio más bajo, son los que más renuncian o derechamente no van a su lugar de trabajo, por lo que se cumple la lógica en cuanto a renuncias por bajo sueldo.

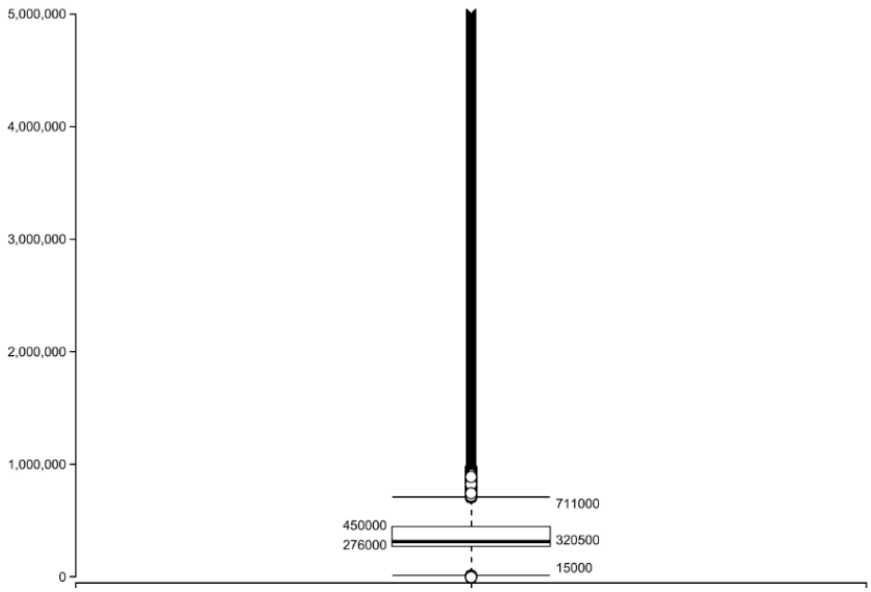
Ilustración 17: Grafico de sueldo base vs causa de fuga



Fuente: Elaboración propia

Se analiza la categoría de sueldo base por sí solo en la Ilustración 18, mediante un diagrama de cajas y bigotes, lo cual nos indica en que valores está concentrado la mayor cantidad de los datos en dicha variable. Se considera los sueldos bajo los 5 millones para el análisis, debido a que los sueldos superiores a ello representan una cantidad ínfima en los datos y distorsionaban el grafico, haciéndolo inentendible.

Ilustración 18: Diagrama de cajas y bigotes del sueldo base

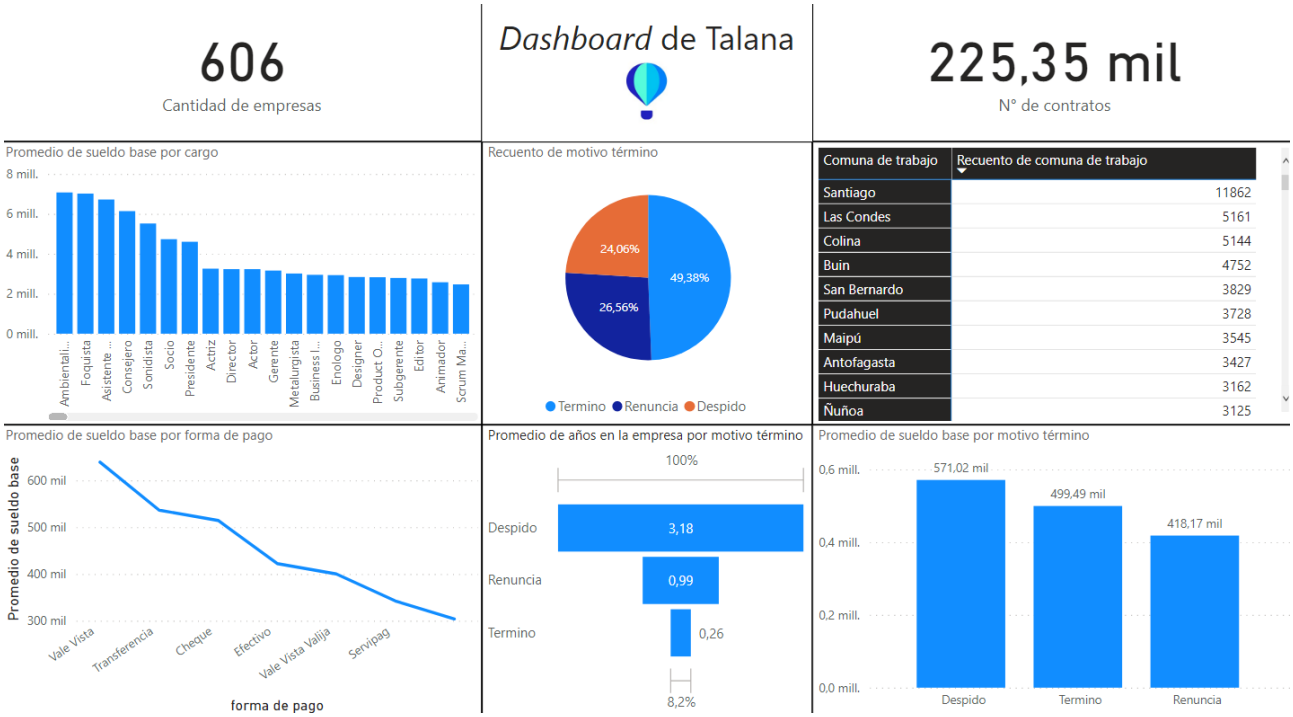


Fuente: Elaboración propia

Como podemos apreciar, es que el diagrama considero como el primer cuartil al número 276.000, al segundo cuartil o mediana a 320.500 y al tercer cuartil a 450.000. Los siguientes valores de mínimo y máximo se calculan mediante el rango intercuartílico y establecen que los números que estén sobre 711.000 y bajo los 15.000 serán considerados atípicos. Se puede destacar que los datos se concentran en gran parte bajo los 450.000, ya que, estos representan el 75% de estos en general.

Para visualizar de manera más general los datos iniciales, se crea un *dashboard* en Power BI que muestre los datos a través de tablas y gráficos interactivos, para así ver la información resumida y como se relacionan las variables entre sí. Algunos de los gráficos mostrados en la Ilustración 19 son: promedio de sueldo base v/s cargo, promedio de años en la empresa v/s motivo termino, recuento de motivo de termino, entre otros.

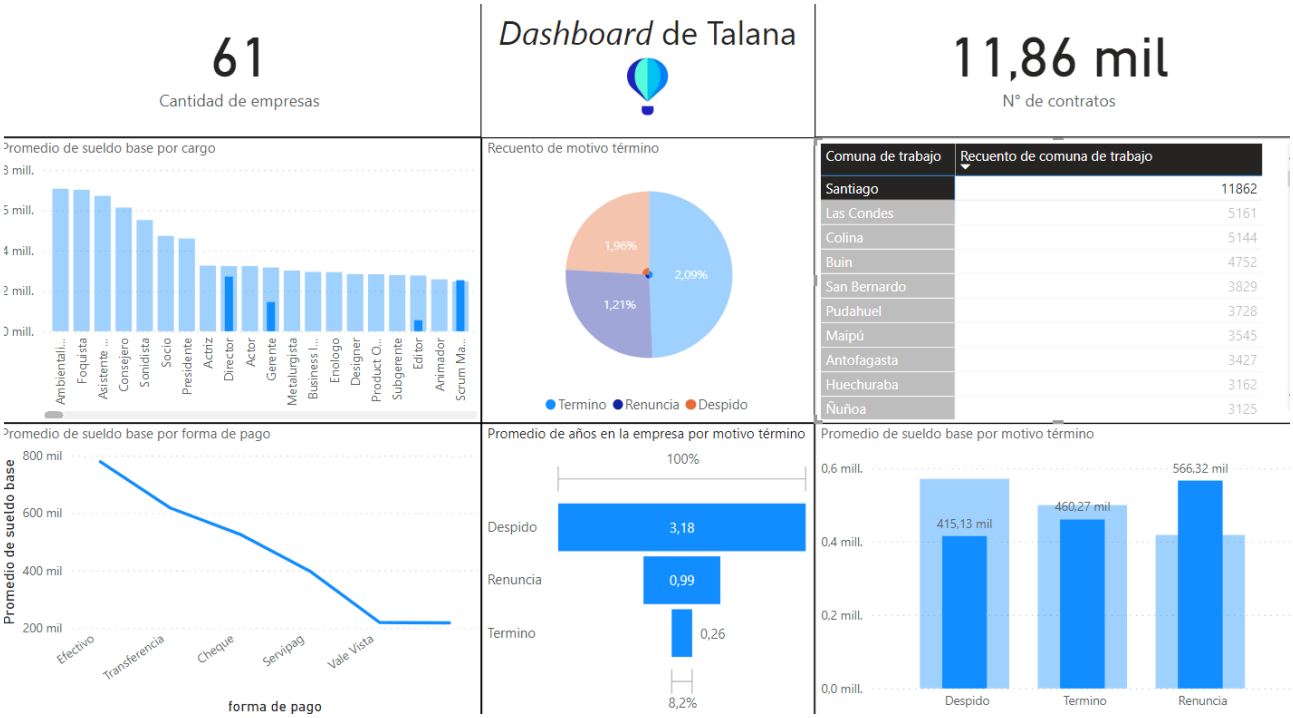
Ilustración 19: *Dashboard* de los datos



Fuente: Elaboración propia

En dicho tablero se puede realizar filtrado a alguna variable en específico, ya que, por ejemplo, en la Ilustración 20 al seleccionar una comuna de trabajo dentro de la matriz como lo es Santiago, se mostrará solo esa comuna y en la tabla general mostrará el desglose de todas las variables relacionadas a los gráficos referentes a la comuna de Santiago.

Ilustración 20: *Dashboard* de los datos filtrados



Fuente: Elaboración propia

Se puede decir entonces, que es posible realizar una predicción de los datos con las variables que se tienen, ya que, se cuenta con más de la mitad de las variables que se requieren y que se debe realizar un proceso de limpieza profundo para poder entrenar los modelos posteriores de la forma más adecuada posible.

CAPÍTULO 4: ESTRUCTURA DE MEJORA

En este capítulo se menciona el cómo será la estructura de trabajo que conllevara a la creación del modelo y de las herramientas que se utilizan para llegar a ello.

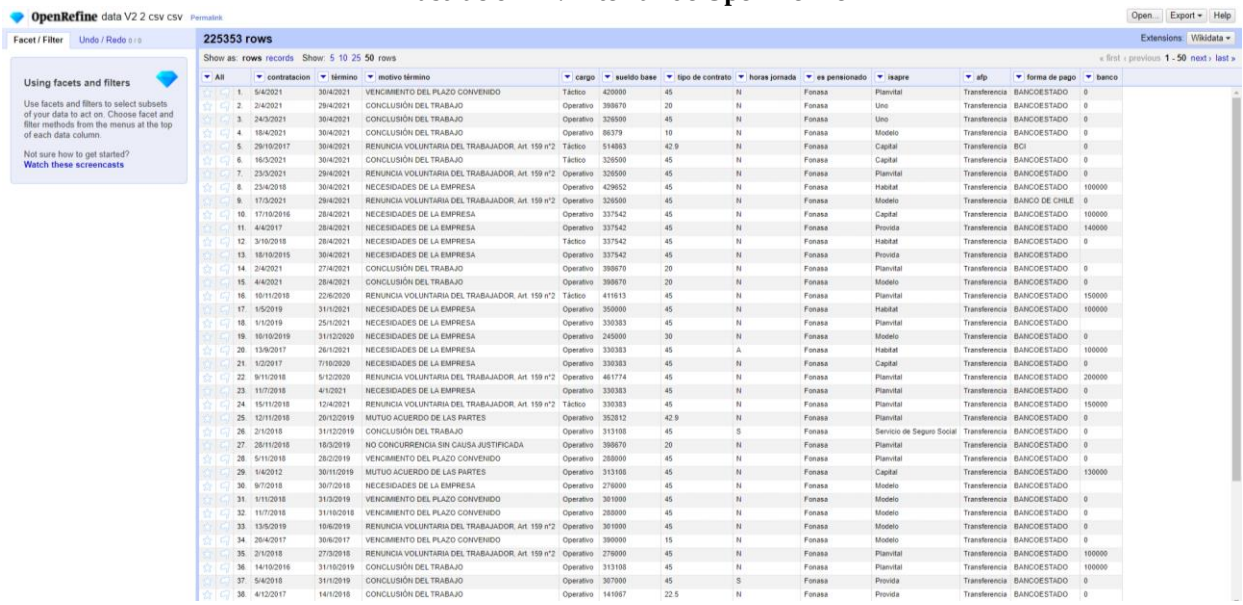
4.1. Herramientas

Antes de empezar a explicar el seguimiento de la metodología propuesta, es importante tener claro que herramientas se utilizaran en el desarrollo del proyecto, para así explicar el contexto de este y entender para que sirven.

4.1.1. Open Refine

La mayoría de las veces, los datos no siempre vienen de la forma más optima que se requiere. En la realidad, estos datos suelen estar desordenados debido a diferentes factores en su generación y estructuración. Por ello, existen herramientas que solucionan este tipo de conflictos, una de ellas es Open Refine.

Ilustración 21: Interfaz de Open Refine



Fuente: Elaboración propia en base a Open Refine

Este *software* consiste en ser una herramienta multiplataforma especializado para trabajar con datos desordenados, permitiendo limpiarlos y transformarlos (Vallés, 2013). El mencionado programa utiliza lenguaje GREL, parecido a la sintaxis de JavaScript, el cual posibilita usar expresiones regulares, es decir, podemos llevar a cabo una variedad de funciones que permiten una depuración avanzada de los datos. También permite inspeccionar los diferentes valores de datos mostrándolos en facetas. Se podría considerar una faceta como una lente a través de la cual se visualiza un subconjunto específico de los datos, basado en un criterio a elección.

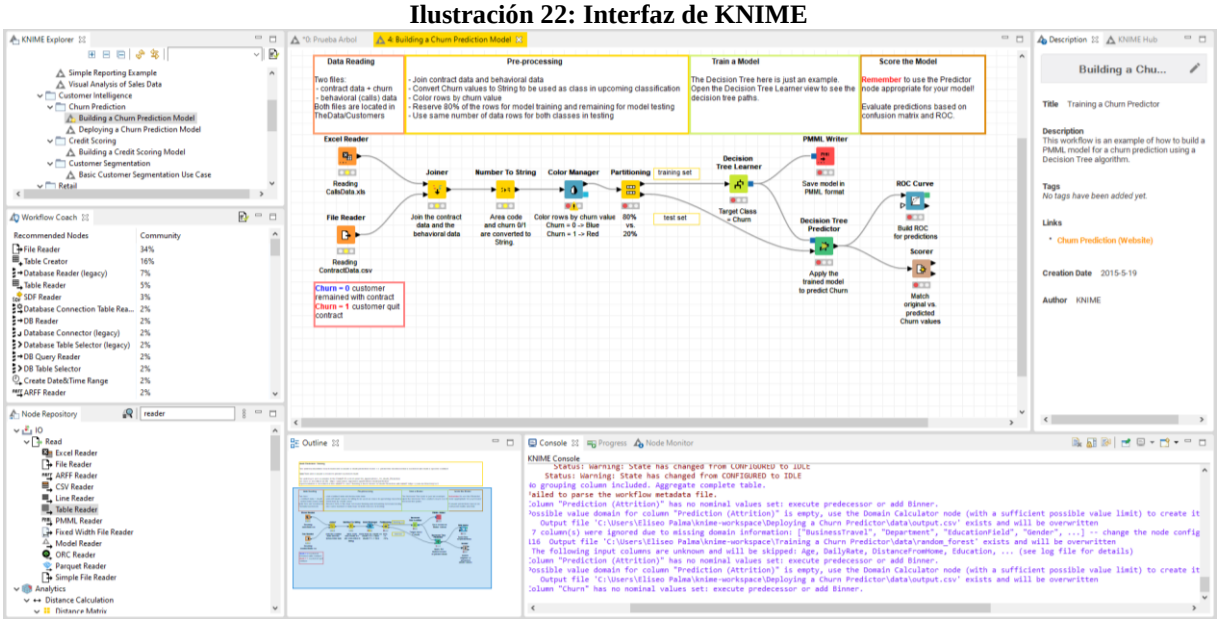
Existen las facetas de texto y facetas numéricas, dependiendo de la naturaleza de los valores contenidos en los campos, se elige uno de ellos.

En el proyecto en cuestión, dicho *software* se utilizará para la categoría, llamado "cargos", ya que, su contenido está muy desordenado, nombrando a un cargo del trabajador que es igual a otro de diferentes formas, variando en el más mínimo cambio, como lo puede ser una letra, espacio, signo, entre otros. Las demás categorías también demuestran que deben ser tratadas, pero se pueden aplicar en el *software* que se menciona a continuación, permitiendo automatizar dichos procesos.

4.1.2. KNIME

En la minería de datos y en los procesos de la información, hay varias herramientas que pueden interactuar con los registros de una base de datos, así también como de una simple tabla almacenada en un repositorio. Una de estas herramientas se denomina KNIME.

Ilustración 22: Interfaz de KNIME



Fuente: Elaboración propia en base a KNIME

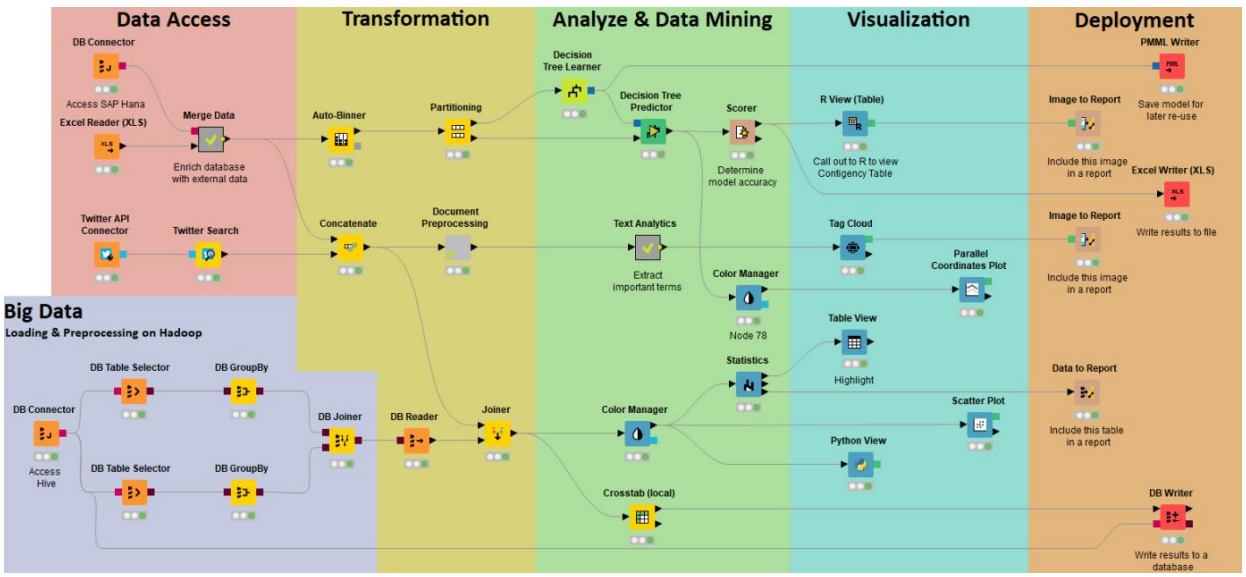
KNIME (*Konstanz Information Miner*), es un *software Open Source* que permite el análisis y manipulación de datos, de una manera fácil e intuitiva. Esto último, se debe a que, se basa en un paradigma de programación de tipo gráfico, es decir, que no es necesario saber programar

en profundidad para poder llevar a cabo la construcción de los flujos y los análisis requerido, ya que los pasos algorítmicos se van construyendo mediante nodos que son parametrizables.

Este es un entorno enfocado a la minería de datos, está construido bajo la plataforma Eclipse y programado esencialmente en Java, que fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania. Además, se puede mencionar que proporciona una alta variedad de extensiones, lo cual enriquece al programa de más funciones y elementos que se pueden ejecutar para su uso, así como también tiene múltiples integraciones con otros programas como Python, R, Java, entre otros.

Dicho *software* apoya al usuario en todo el ciclo de la minería de datos, como, por ejemplo, en la metodología propuesta en este proyecto, ya que, se distingue su extrema flexibilidad y potencia, a través de las distintas facilidades que entrega como: la integración y manipulación de datos, visualización de datos, creación de modelos de data mining, validación de modelos, creación de informes y escritura de datos.

Ilustración 23: Ciclo de minería de datos en KNIME



Fuente: Elaboración propia en base a KNIME

La empresa consultora y de investigación de IT, *Gartner* ha analizado el mercado de las herramientas basados en minería de datos, incluyendo las más importantes en el cuadrante mágico, proporcionando información útil con respecto a las principales características de cada

uno de ellos. En este sentido, dicho *software* figura como uno de los líderes en el estudio y es una de las más prometedores en este sector.

En este punto, surge la pregunta de porque utilizar KNIME por sobre otros *software*s que permiten la aplicación de las técnicas de minería de datos y que se encuentran actualmente en el mercado. Quizás el aspecto más importante en esta elección es que en la realización del proyecto, será posible agilizar la construcción de los modelos, ya que, como se dijo anteriormente, el *software* permite la ejecución de los algoritmos sin la necesidad de usar la programación y codificación algorítmica, pudiendo realizar análisis rápidamente. De todas maneras, si es que la empresa requiere que algún aspecto en la construcción del flujo sea codificado, este se puede hacer por las integraciones que tiene el programa con las demás aplicaciones.

Siguiendo con el estudio antes mencionado, se tiene en cuenta como KNIME se compara con las demás plataformas que realizan minería de datos. En la Ilustración 24, se puede notar esta apreciación, teniendo en el eje de las abscisas, la integridad de la visión y en el eje de las coordenadas, la capacidad de ejecución. Además, se generan cuatro cuadrantes, los cuales definen en que posición se encuentran las empresas TI analizadas, estos aspectos son:

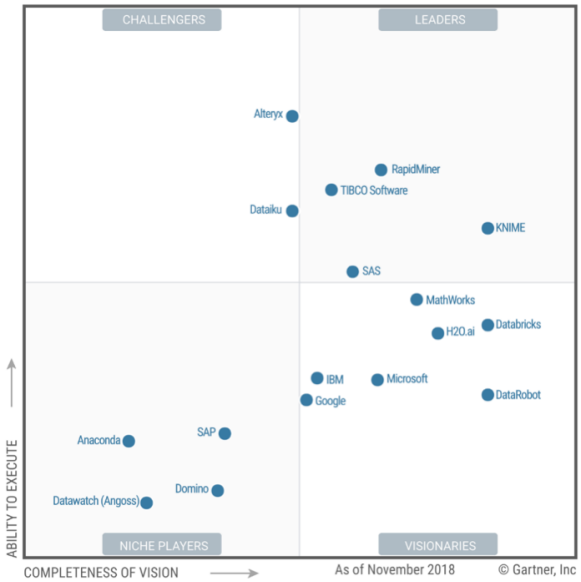
- ***Challengers:*** tienen mayor capacidad de ejecución, pero menor integridad de visión, es decir, corresponden a los que son capaces de dominar una gran parte del mercado, pero no muestran una dirección o visión del mercado en general.

- ***Niche Players:*** tienen una baja capacidad de ejecución y baja integridad de visión, es decir, corresponden a los que son exitosos en su respectivo segmento, pero no muestran una visión de mercado y no son innovadores para crear una ventaja competitiva.

- ***Visionaries:*** tienen baja capacidad de ejecución, pero alta integridad de visión, es decir, corresponden a los que tienen una buena visión de mercado, pero no son capaces de ejecutar de forma correcta las ideas para crear mayor ventaja.

- ***Leaders:*** tienen alta capacidad de ejecución y alta integridad de visión, es decir, corresponden a los que tienen un buen desempeño en los dos frentes, ya que, desarrollan bien sus capacidades con una adecuada visión del futuro del mercado.

Ilustración 24: Cuadrante mágico de Gartner en *softwares* de minería de datos

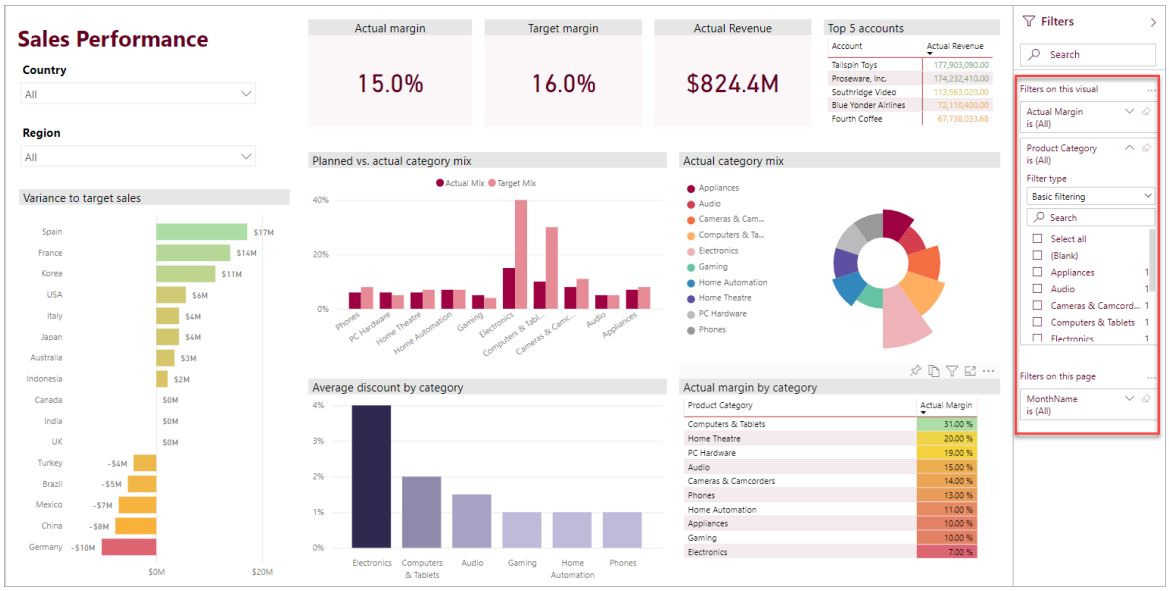


Fuente: (Gartner, 2018)

4.1.3. Power BI

Power BI, es una herramienta de *Business Intelligence* (BI), creada por Microsoft, que permite unificar diferentes fuentes de información diferentes en la nube y que arroja una vista de los datos más trascendentales de una empresa para así poder mejorarlos desde un análisis de negocio.

Ilustración 25: Interfaz de Power BI



Fuente: (Microsoft, s.f.)

Esta solución de inteligencia empresarial permite supervisar el estado de los datos a través paneles dinámicos e informes interactivos en tiempo real, disponibles de forma inmediata para todos los trabajadores de la empresa. Las fuentes de información que unifica Power Bi, permiten evaluar las debilidades, oportunidades y fortalezas del negocio, para darte una idea de dónde y cómo se deben enfocar sus esfuerzos para crecer y mejorar (XMS, s.f.).

Así como con KNIME, también se debe preguntar el porqué de la selección de Power BI por sobre otros *software*s. Esto está fundamentado principalmente en que se puede mostrar y visualizar la información de una manera dinámica y en tiempo real. También se eligió por sobre otros, ya que, es el de menor costo comparado con Tableu u otro tipo de *software*s similares que realizan las mismas funcionalidades.

Cabe destacar, que la empresa consultora Gartner, también realizo una comparación con otras plataformas de este estilo. En la Ilustración 26 se puede apreciar los mismos ejes y cuadrantes los cuales clasifican a los distintos *software*s de inteligencia de negocio.

Ilustración 26: Cuadrante mágico de Gartner en *software*s de inteligencia de negocio

Fuente: (Gartner, 2021)

CAPÍTULO 5: APLICACIÓN DE LA METODOLOGIA

En el presente capitulo se muestran los pasos de la metodología y como se fueron aplicando a lo largo del desarrollo del proyecto para llegar al resultado requerido

5.1. Primera iteración

5.1.1. Preparación de datos

Antes de empezar el capítulo se debe considerar que los primeros pasos de la metodología ya se realizaron, siendo documentados en los capítulos anteriores. Estos dan los lineamientos y bases para los siguientes procedimientos. Además, de que en algún paso de los que se mencionaran pueden resultar modificados a lo largo del proyecto, debido a que la metodología es circular, es decir, permite volver hacia atrás. Teniendo claro lo anterior, se procede a explicar cómo se preparan los datos para su posterior y debido uso en la construcción de los modelos de minería de datos.

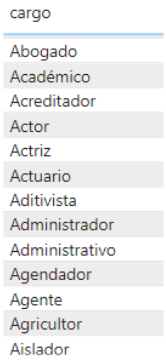
Para clarificar y recordar, se tiene una diversidad de atributos de una tabla que corresponden a los datos recopilados anteriormente. Cada uno de estos atributos presenta una naturaleza distinta, es decir, pueden ser de tipo categórico o numéricos. Si es de tipo categórico, estos pueden tener forma de texto o de fecha y si es de tipo numérico, pueden ser discretos, continuos o booleanos.

Primero, se les dio mayor importancia a los atributos más complejos, en este caso el atributo "Cargo" que es de tipo categórico en forma de texto, es uno de los más complejos de la base de datos, debido a que los registros que contiene están desordenados, es decir, dentro de las cadenas de texto contienen caracteres de más donde no debería. Además de tener una gran cantidad de valores nominales debido a que cada trabajador tiene un cargo distinto y al tener varias empresas dentro de la información contenida, estos valores aumentan considerablemente. Para limpiar esta variable, se utiliza el *software* mencionado previamente Open Refine, ya que, se especializa en la ejecución de estas prácticas.

Estando en la interfaz de la aplicación, se eliminan espacios sobrantes, signos y números. Teniendo eso procesado, se exportan los datos para utilizar el algoritmo "*Text Distance",* en función de la métrica de Jaro normalizada, e hicieran match con un umbral de 90%, de esta forma reducimos la carga de trabajo posterior, cargándolo nuevamente al *software*.

Ya se ha limpiado cierta parte de los datos, pero aún existe un gran porcentaje de registros sucios. En este punto, se utiliza la clusterización de textos y facetas para limpiar los datos que faltaban. Como hipótesis, se juntaron registros de cargos que eran similares, como, por ejemplo: “Jefe de bodega” con “Jefe de finanzas” se convirtieron en la cadena “Jefe”. Con ello se redujo los valores nominales de más de 10 mil a 400 para la variable. En la Ilustración 27, se muestra un ejemplo de ello.

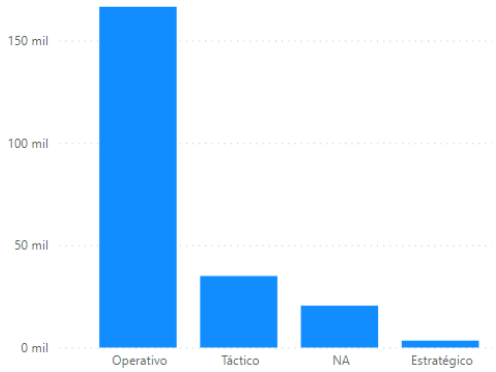
Ilustración 27: Muestra de los valores de “Cargo” post limpieza



Fuente: Elaboración propia

Considerando las etapas posteriores, hipotéticamente esas cantidades de valores pueden ser muy altas, el cual puede ocasionar ruido en el modelo que se quiere llegar, por ello se analizó por cada registro. la jerarquía del cargo normalizándolo por “Operativo”, “Táctico” y “Estratégico”. Esto representa de alguna medida la importancia del cargo en su respectiva empresa y facilita los análisis posteriores que se deben realizar.

Ilustración 28: Valores del atributo "Cargo"



Fuente: Elaboración propia

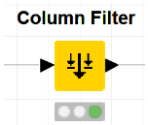
Para entender cómo se da la Ilustración 28, hay que explicar cada nivel organizacional que componen las empresas.

- **Nivel estratégico:** Se trata de la visión que mueve las acciones de la empresa. Establecen los objetivos a cumplir y las líneas maestras para alcanzarlos. La dirección juega un rol principal a la hora de definir la estrategia, el cual lo componen los gerentes, directores administradores, entre otros.
- **Nivel táctico:** Solo los departamentos se encargan de desarrollar este nivel. Se crean las acciones a realizar para hacer realidad la estrategia de la empresa. Es un tipo de planificación específica y que atiende en profundidad a los detalles. Dicho nivel, está integrado por jefes, ingenieros, entre otros.
- **Nivel operativo:** En este último nivel aparecen los agentes encargados de ejecutar las acciones desarrolladas en el nivel táctico. Realizan acciones de corta duración y todos en la empresa tienen un rol que desempeñar en este nivel. Acá, lo integran los operarios, administrativos, cocineros, entre otros.

Por ende, en las empresas en general, en su nivel operativo habrá mayor personal que en el táctico, y en este último, más que en el estratégico, formando la pirámide organizacional que cuentan las empresas. Habiendo normalizado el atributo "Cargos", se empieza a utilizar el *software* KNIME y sus funcionalidades para el tratamiento de los datos de las demás categorías que disminuyen en complejidad en cuanto al atributo anteriormente tratado, para ello, KNIME nos ofrece una variedad de nodos con los que poder realizar tratamiento de datos.

Primero que todo, para utilizar la plataforma y el marco de trabajo del programa se deben conectar y leer la última versión de los datos, con el nodo "File Reader", en el cual se debe confirmar la naturaleza de las variables y la codificación de los caracteres, los cuales son en este caso "UTF-8". Ahora, nos centraremos en el atributo "Anticipo pactado", la cual cuenta con una cantidad de valores perdidos de 1.150 pero que en más de 200 mil registros cuenta con el valor de 0, lo cual no proporciona información y se procede a utilizar *"Column Filter"* de la Ilustración 29 para prescindir de esta variable.

Ilustración 29: Nodo "*Column Filter*"



Fuente: Elaboración propia en base a KNIME

En el siguiente nodo que se conecta al flujo, tiene la funcionalidad de que puede rellenar los valores perdidos (*NA´s*) con algún elemento relacionado a la columna en sí, la cual en los casos de que una de estas fuera categórica, la puede reemplazar con el valor más frecuente, el valor próximo, el valor previo, rellenar con algún valor a elección o simplemente eliminar el registro por completo. En el caso de que fuera numérica, la puede reemplazar con una interpolación linear, el máximo, el mínimo, la media, la mediana, la media aproximada, el valor próximo y previo o de la misma forma, eliminar el registro. Cabe destacar que esas imputaciones corresponden a la presente iteración y puede cambiar más adelante. A continuación, se menciona las variables que presentaron valores perdidos y la técnica que se utilizó para ser reemplazadas dentro del nodo *"Missing Value"* de la Ilustración 30:

- **"Banco":** con 54,994 valores perdidos, se rellenan por el valor más frecuente.
- **"Isapre":** con 16.964 valores perdidos, se rellenan por el valor más frecuente.
- **"AFP":** con 11.722 valores perdidos, se rellenan por el valor más frecuente.
- **"Forma de pago":** con 43.933 valores perdidos, se rellenan por el valor más frecuente.
- **"Es pensionado":** con 29 valores perdidos, se rellenan por el valor más frecuente.
- **"Cargo":** con 20.480 valores perdidos, se rellenan por un valor a elección, la cual es "Otros".

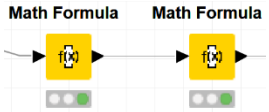
Ilustración 30: Nodo "*Missing value*"

Fuente: Elaboración propia en base a KNIME

Después, nos centramos en los atributos de "Sueldo base" y "Jornadas", las cuales presentan registros con el valor 0, los cuales no presentan mucha lógica si se mira del punto de vista de que cada uno es un trabajador que recibe su sueldo y con un horario fijo. Por lo tanto, se reemplaza estos valores por la mediana en el caso del "Sueldo base" y la media en el caso de las "Horas jornadas", gracias al nodo *"Math Formula"* de la Ilustración 31. Esto se realiza así, ya que, en el caso del "Sueldo base", existen sueldos muy altos, el cual colocar una media de los datos no sería lo más adecuado, ya que, su valor estaría muy por encima de la moda, por

ello se colocó la media. En “Jornadas” se utiliza la media, debido a que las jornadas no varían mucho en cuanto los horarios que realiza cada trabajador y se optó por esa opción.

Ilustración 31: Nodos "*Math Formula*"



Fuente: Elaboración propia en base a KNIME

Siguiendo con el flujo, para cada trabajador existe el motivo de término de su contrato la cual se encuentra en el atributo “Motivo término”. Acá, se busca poder reducir los valores a algo más genérico. En la Ilustración 32, se muestra una parte de los datos.

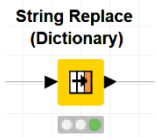
Ilustración 32: Valores del atributo “Motivo término”

motivo término
ABANDONO DEL TRABAJO
ACTOS,OMISIONES O IMPRUDENCIA
Autodespido
CASO FORTUITO O FUERZA MAYOR
CONCLUSIÓN DEL TRABAJO
Conducta de acoso Sexual
CONDUCTAS DE ACOSO LABORAL
FALTA A LA PROBIDAD
INCUMPLIMIENTO GRAVE OBLIGACIONES
MUERTE DEL TRABAJADOR
MUTUO ACUERDO DE LAS PARTES
NECESIDADES DE LA EMPRESA
NECESIDADES DE LA EMPRESA inciso 2
NEGATIVA A TRABAJAR S/CAUSA
NEGOCIOS EJECUTE TRABAJADOR
NO CONCURRENCIA SIN CAUSA JUSTIFICADA
no registrado
PERJUICIO MATERIAL CAUSADO EN LAS INSTALACIONES
RENUNCIA VOLUNTARIA DEL TRABAJADOR, Art. 159 n°2
VENCIMIENTO DEL PLAZO CONVENIDO
VIAS DE HECHO

Fuente: Elaboración propia

Estos se normalizaron en tres valores, los cuales son “Renuncia”, “Despido” y “Termino”, con el nodo “String Replace (Dictionary)” de la Ilustración 33.

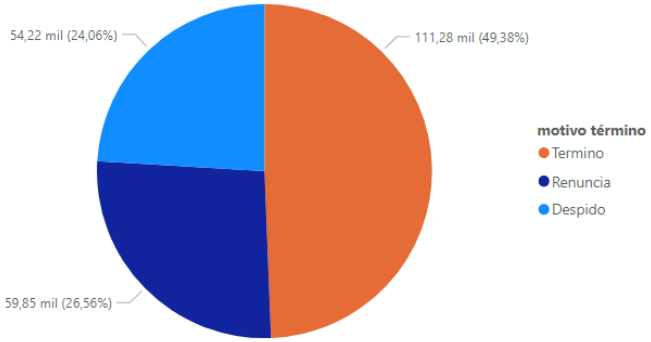
Ilustración 33: Nodo *“String Replace (Dictionary)"*



Fuente: Elaboración propia en base a KNIME

Esta transformación arrojo la siguiente composición de los datos en cuanto a la fuga, los cuales pertenecen al 50,62% como se muestran en la Ilustración 34.

Ilustración 34: Grafico de fuga



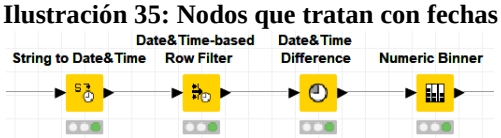
Fuente: Elaboración propia

Del total de los datos, se infiere que casi la mitad de los trabajadores terminan su contrato de forma normal, la otra mitad que corresponde a la fuga en sí, son más los que renuncian de los que despiden.

En los siguientes nodos tienen relación con las categorías de fechas, las cuales corresponde a “Fecha comienzo” y “Fecha término”. Esto servirá, ya que, se puede filtrar las fechas antes de la pandemia, debido a que las condiciones de los trabajadores y motivos de renuncia no son los mismos en ese hito de los acontecimientos.

Por consiguiente, se obtiene una nueva columna con la diferencia de tiempo entre el hito y la fecha de término, con el objetivo de posicionarnos en ese momento del tiempo y poder determinar los que aun mantenían contrato con su empresa y no se habían ido como “Permanencia”, el que será el elemento que se contrapondrá con la fuga. Esto se realiza mediante los nodos de la Ilustración 35.

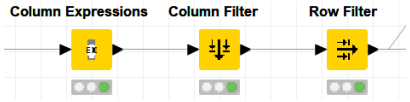
Ilustración 35: Nodos que tratan con fechas



Fuente: Elaboración propia en base a KNIME

Ya finalizando el proceso de tratamiento, en la Ilustración 36 se filtran las columnas de las fechas y se saca de los datos a los “Terminados”, debido a que no renunciaron ni fueron despedidos.

Ilustración 36: Nodos de clase y filtro

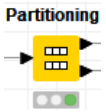


Fuente: Elaboración propia en base a KNIME

5.1.2. Construcción y evaluación del modelo

Empezando con la construcción del modelo de clasificación, se debe particionar los datos en entrenamiento y prueba, en este caso se tomó para entrenamiento un 80% de los datos y para la prueba un 20%. Esto se puede gracias al nodo *"Partitioning"* de la Ilustración 37.

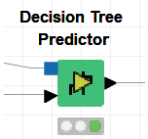
Ilustración 37: Nodo de partición



Fuente: Elaboración propia en base a KNIME

Por consiguiente, se procede a entrenar el algoritmo para crear el modelo correspondiente con los datos de entrenamiento definidos anteriormente mediante estratificación aleatoria en la Ilustración 38.

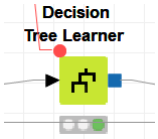
Ilustración 38: Nodo de entrenamiento



Fuente: Elaboración propia en base a KNIME

Con el modelo ya creado y entrenado, se ejecuta el predictor para que escriba los resultados en donde predecirá la clase de los datos de prueba, para comprobar su eficacia en la Ilustración 39.

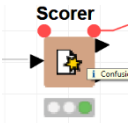
Ilustración 39: Nodo de predicción



Fuente: Elaboración propia en base a KNIME

El ultimo nodo que se muestra en la Ilustración 40 escribió los resultados en la tabla, próximo al atributo "Clase", con el fin de que sean comparadas por el nodo "Scorer" y así poder crear la matriz de confusión, para poder establecer la precisión y error del clasificador en cuestión.

Ilustración 40: Nodo constructor de la matriz de confusión



Fuente: Elaboración propia en base a KNIME

5.1.3. Evaluación

Se siguió este mismo proceso de construcción y evaluación, para tres modelos en esta primera iteración, los cuales son arboles de decisión, bayes ingenuo y bosques aleatorios.

Tabla 2: Resultados de precisión de la iteración 1

Modelo	Precisión
Arboles de decisión	61,5%
Bayes ingenuo	51,5%
Bosques aleatorios	64,2%

Fuente: Elaboración propia

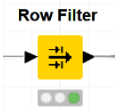
Los resultados que se muestran en la Tabla 2, nos dicen que el mejor clasificador hasta el momento es bosques aleatorios. Cabe recalcar que a pesar de que los resultados de la precisión de los algoritmos recién mostrados no son de los mejores se debe encontrar los parámetros idóneos y prepara de otra forma los datos para que estos algoritmos tengan un mejor desempeño.

5.2. Segunda iteración

En la segunda iteración, se procedió a cambiar el tipo de clase, pasando de una clase-múltiple a una clase binaria. Esto con el objetivo de obtener mayor exactitud a la hora de predecir con otro tipo de resultados similares a los que se pensó inicialmente.

Dentro de los casos de predicción de fuga de trabajadores, se tiene a la deserción de empleados o más bien llamado "*employee atrittion*", lo cual consta de la predicción binaria de renuncia o permanencia en la empresa. Para ello se filtran los datos con el nodo de la Ilustración 41, sacando del conjunto de datos, a los que resultaron despedidos. Esto tiene sentido, debido a que la empresa es la que toma esta decisión sobre el empleador, ya que, sabe cuándo puede ocurrir tal evento, considerando determinados factores, como necesidades de la misma empresa principalmente.

Ilustración 41: Nodo de filtrado de "Despidos"



Fuente: Elaboración propia

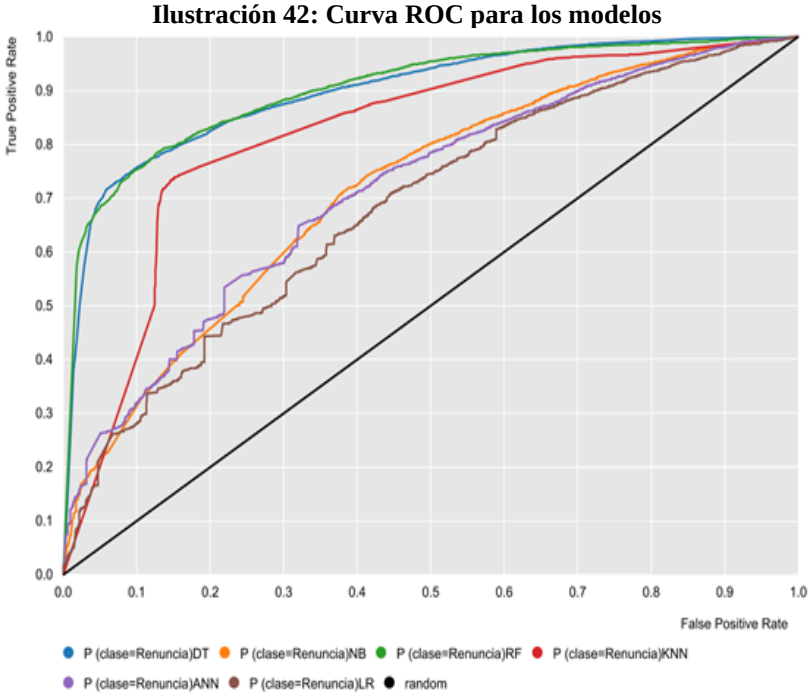
Tras ello, se ejecutaron los nodos de construcción de los modelos para obtener los resultados esperados. Los resultados se muestran en la Tabla 3.

Tabla 3: Resultados de precisión de la iteración 2

Modelo	Precisión
Arboles de decisión	82,8%
Bayes ingenuo	73,1%
Bosques aleatorios	83,4%
KNN	79,5%
Redes Neuronales Artificiales	74,5%
Regresión Logística	73,5%

Fuente: Elaboración propia

Ilustración 42: Curva ROC para los modelos



Fuente: Elaboración propia en base a KNIME

En la Ilustración 42, se muestra la curva ROC, que compara los verdaderos positivos con los falsos positivos, dependiendo de los umbrales de clasificacion que se pueden obtener. Dicho grafico representa a los modelos generados por arboles de decisión, bayes ingenuo, bosques aleatorios, KNN, redes neuronales artificiales y regresión logística. Se puede apreciar que tanto

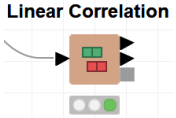
como el modelo de árboles de decisión, como bosques aleatorios, entregan mejores resultados, siendo levemente superior los bosques aleatorios por sobre los árboles de decisión, en cuanto a la precisión resultante. No obstante, en la gráfica mencionada no difieren en casi nada entre estas.

5.3. Tercera iteración

En esta tercera iteración, se realizan ciertos procedimientos y cambios a la construcción de los datos que alimentan los modelos de clasificacion. En conjunto con el CTO de Talana, se establece que los datos de entrada deben ser lo menos ruidoso posible, es decir, que no contengan datos ficticios de algún nodo que trate los valores perdidos. Por ello, se tratan estos valores por medio del método de *listwise deletion*, es decir, la eliminación de todos aquellos casos que poseen algún valor perdido. Se asume que los datos perdidos siguen un patrón MCAR, por ende, el procedimiento no agrega sesgos a los datos. Esta decisión podría revertirse posteriormente y decidir algún tipo de imputación más efectiva, que como se tenía anteriormente de solo imputar el valor más frecuente, la media y mediana, para así, tener mayor poder de análisis.

Se mantiene la columna de “Anticipo pactado”, ya que, los valores iguales a 0 en esta variable, son relevantes para el estudio en general. Además, se crea otra variable, a partir de las fechas de contratación y termino de contrato, la cual es llamada “Años en la empresa”. Esta variable es importante, ya que es una de las que recurrentemente está presente en este tipo de casos. Finalmente, en dicha iteración, se realizó el análisis de correlaciones, el cual se comprobó la correlación de todas las columnas involucradas entre sí, con el objetivo de poder identificar las que altamente estén correlacionadas para así evitar que ingresen al modelo. En la Ilustración 43 se muestra su nodo y en la Tabla 4, se identifica las variables y sus correlaciones con las demás.

Ilustración 43: Nodo de correlación lineal



Fuente: Elaboración propia en base a KNIME

Tabla 4: Resultados de correlación de las variables entre si

Primera columna	Segunda columna	Correlación
es pensionado	afp	0.36510196704332304
forma de pago	banco	0.3227320327636767
cargo	isapre	0.27591023706932166
cargo	banco	0.24028843361286908
banco	clase	0.23005698811919648
cargo	clase	0.1800480724980857
afp	clase	0.14801256798324539
forma de pago	clase	0.1428707115242947
sueldo base	anticipo pactado	0.14051073897009844
isapre	clase	0.1282125613006103
cargo	afp	0.11933030295663616
isapre	banco	0.11268310656133214
cargo	forma de pago	0.09911447047428294
es pensionado	clase	0.09583461719093385
afp	banco	0.09558874490267183
isapre	afp	0.08422638327478274
es pensionado	banco	0.07710054409950096
afp	forma de pago	0.07342887106346664
isapre	forma de pago	0.06606893187995379
horas jornada	anticipo pactado	0.05169828089441236
sueldo base	años en la empresa	0.05110054251352774
sueldo base	horas jornada	0.04516014744814837
es pensionado	forma de pago	0.043808552901783517
cargo	es pensionado	0.03980791641177071
es pensionado	isapre	0.017677329626548523
anticipo pactado	años en la empresa	0.010090398961034336
horas jornada	años en la empresa	0.0038687547504163066

Fuente: Elaboración propia

Generalmente, las variables con correlaciones que estén fuera del rango de [-0.7; 0.7], son eliminadas del modelo, para así evitar clasificaciones mal ejecutadas. Por ello, se puede notar que en las variables de estudio no se presentan correlaciones altas ni negativas entre las variables, por ende, no se elimina ninguna de estas y son aptas para ingresar a los modelos.

Con dichos cambios, en la Tabla 5 se procede a presentar los resultados precisión de los modelos en cuestión.

Tabla 5: Resultados de precisión de la iteración 3

Modelo	Precisión
Arboles de decisión	82,8%
Bayes ingenuo	67,2%
Bosques aleatorios	85,7%
KNN	76,1%
Redes Neuronales Artificiales	65,3%
Regresión Logística	64.9%

Fuente: Elaboración propia

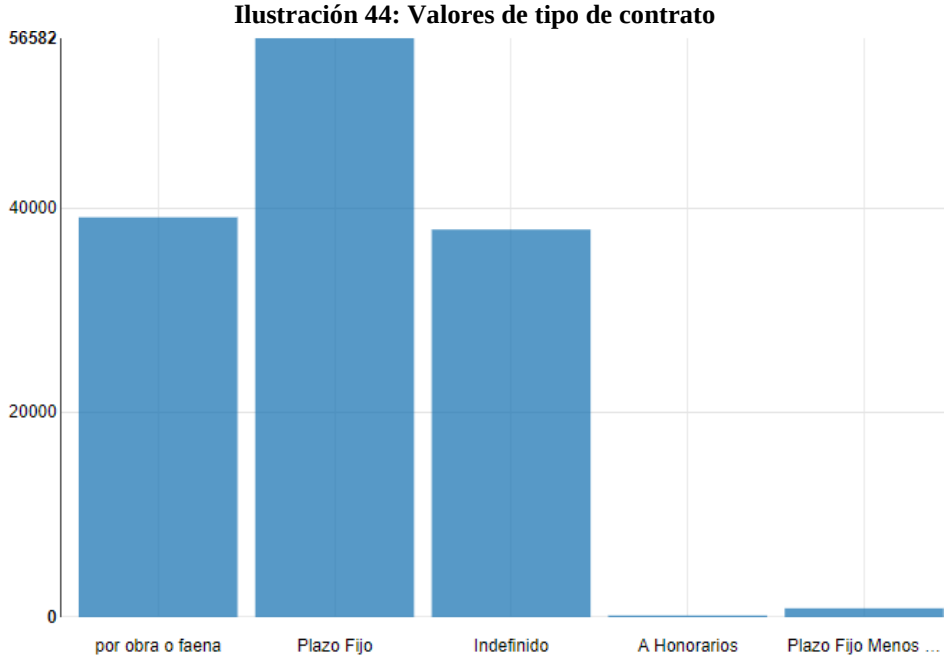
Como se puede observar en los resultados, todos los algoritmos empeoraron su rendimiento, excepto arboles de decisión y bosques aleatorios. Este último, incluso mejoro su rendimiento a comparación con los demás, pasando de un 83,4% a un 85,7% de precisión en sus predicciones.

5.4. Cuarta iteración

En esta iteración, se agregan dos nuevas columnas para aumentar la información de los datos y aumentar el poder de decisión. Estos se llaman "Tipo de contrato" y "Distancia", los cuales se obtuvieron posteriormente a la tercera iteración.

- **Tipo de contrato:** es una variable categórica que se divide por los valores de "por obra o faena", "indefinido", "a honorarios", "plazo fijo", "plazo fijo menos 30 días" y "a honorarios" con un total de datos faltantes de 23.633. Estos datos son filtrados, obteniendo el grafico mostrado en la Ilustración 44.

Ilustración 44: Valores de tipo de contrato



Fuente: elaboración propia en base a KNIME

- **Distancia (Km):** como se mencionó anteriormente, se añadió la variable de tipo numérica, llamada "distancia", que es la distancia en kilómetros que tienen los trabajadores desde donde viven (residencia) hasta la empresa (trabajo), la cual se calculó mediante las columnas referentes a "comuna residencia", "calle residencia", "comuna

trabajo" y "calle trabajo", gracias a la API de Google Maps, la cual se automatizo en Visual Basic para agilizar dicho proceso de extracción. Se tiene que considerar sobre esta variable, que los números son aproximados y en algunos casos puede que la API haya confundido una calle o comuna del mismo nombre en otro lugar del país y se haya calculado dicha distancia, por ende, es un dato que se aproxima a la realidad, pero no lo representa en su totalidad. Esta última variable, cuenta con más de la mitad de los registros con valores perdidos, ya que, si no se disponía de los datos de trabajo y residencia en simultaneo, no se podía llevar a cabo el cálculo correspondiente. Por ende, se imputaron los valores por medio del predictor de bosques aleatorios para datos faltantes, más conocido como missForest, la cual se conecta al entorno de R Studio para llevar a cabo su cálculo. Este método nos sirve, debido a que no sesga los datos y no requiere de comprobar que los atributos sean linealmente dependientes entre sí. Finalmente, se realiza el análisis de correlación resultando que nuevamente ninguno de las variables presenta algún tipo de relación lineal con otra.

Con dichos cambios, en la Tabla 6 se procede a presentar los resultados de precisión de los modelos en cuestión.

Tabla 6: Resultados de precisión de la iteración 4

Modelo	Precisión
Arboles de decisión	82,7%
Bayes ingenuo	66,5%
Bosques aleatorios	86,1%
KNN	72,0%
Redes Neuronales Artificiales	76,5%
Regresión Logística	71.1%

Fuente: Elaboración propia

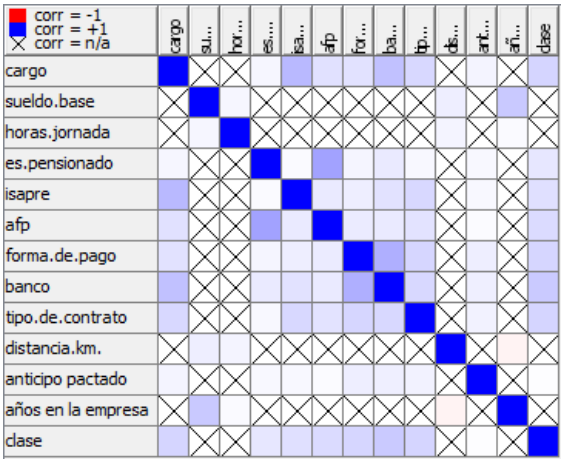
Como se puede apreciar, los únicos que bajaron rendimiento son los algoritmos de KNN y Bayes ingenuo, los demás aumentaron su rendimiento y a pesar de que los árboles de decisión y bosques aleatorios se mantienen de forma similar, estos superan a los demás significativamente, demostrando que se adaptan de buena manera a este tipo de casos.

5.5. Quinta iteración

En esta iteración se eliminan alrededor de 5.000 filas pertenecientes a valores atípicos o también llamados *Outliers*, de las columnas numéricas de "horas jornada", "sueldo" y "años en la empresa", ya que estos valores que están fuera de un rango aceptable pueden estar afectando el rendimiento de los clasificadores, debido a que no pertenecen a datos que se requieran analizar.

A pesar de que el conjunto de datos no cambió radicalmente como en iteraciones pasadas, también para esta dicha iteración se realiza el análisis de correlación en la Ilustración 45, encontrando de que, al igual que antes, todas son bajas y las columnas en su totalidad no afectarían en los resultados posteriores.

Ilustración 45: Grafico de correlaciones



Fuente: elaboración propia en base a KNIME

Con dichos cambios, en la Tabla 7 se procede a presentar los resultados de precisión de los modelos en cuestión.

Tabla 7: Resultados de precisión de la iteración 5

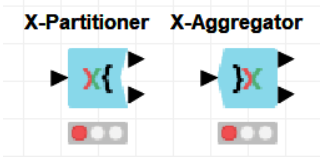
Modelo	Precisión
Arboles de decisión	82,8%
Bayes ingenuo	66,9%
Bosques aleatorios	85,5%
KNN	73,0%
Redes Neuronales Artificiales	72,4%
Regresión Logística	65.8%

Fuente: elaboración propia

Como se puede apreciar, los que bajaron rendimiento son los algoritmos de regresión logística, bosques aleatorios y las redes neuronales artificiales, donde los demás aumentaron su rendimiento. Se repite la tendencia de que los árboles de decisión y bosques aleatorios de superar a los demás clasificadores.

Para reafirmar la elección del modelo, ya que, es evidente la elección del modelo de bosques aleatorios, se realiza la validación cruzada de cada uno de estos modelos, gracias a los nodos que se ven en la Ilustración 46.

Ilustración 46: Nodos de validación cruzada



Fuente: elaboración propia en base a KNIME

Lo cual consiste en dividir el data set en k partes (k=10), y se corre el modelo k veces, con una parte como test y lo demás como entrenamiento, para así tener un conjunto de parámetros adicional que ayude a tener un mayor poder de decisión. A continuación, en la Tabla 8 se presenta las precisiones medias de los modelos y su respectiva desviación estándar.

Tabla 8: Resultados de la validación cruzada

Modelo	Precisión Media	Desviación Estándar
Arboles de decisión	82,6%	0,008
Bayes ingenuo	67,2%	0,009
Bosques aleatorios	85,8%	0,005
KNN	73,1%	0,008
Redes Neuronales Artificiales	72,9%	0,014
Regresión Logística	66.3%	0,009

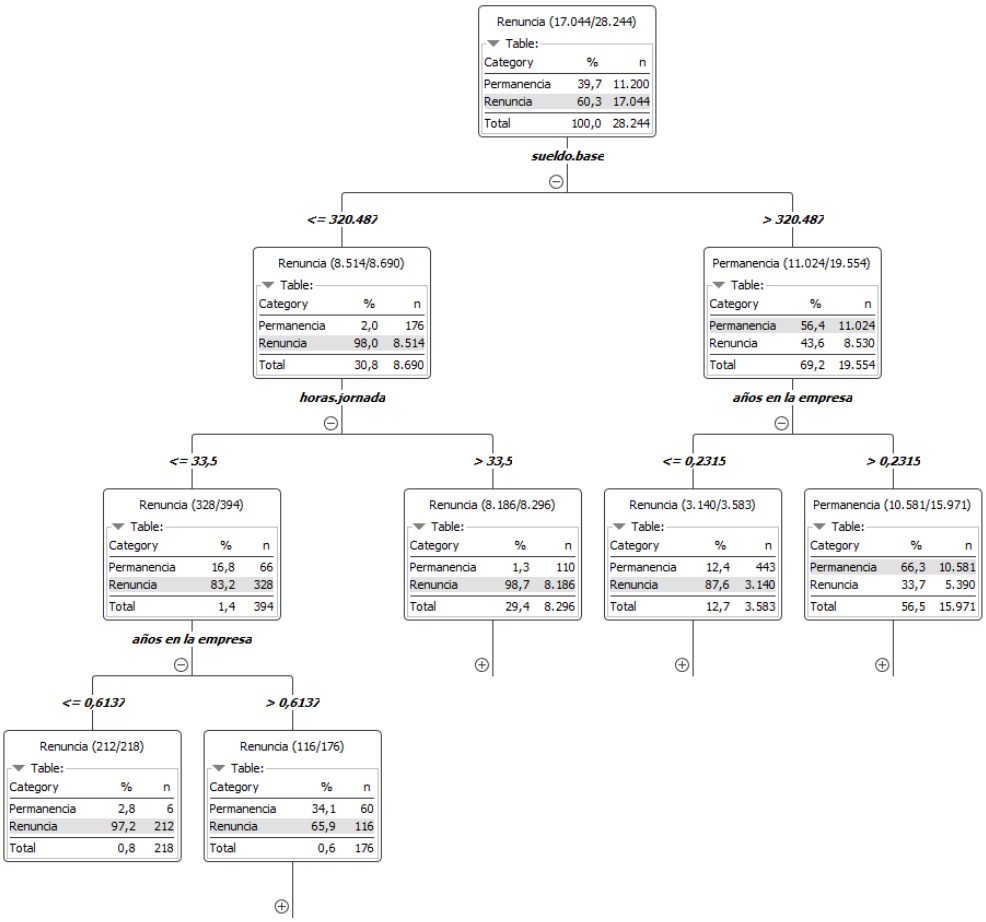
Fuente: elaboración propia

Con dichos parámetros, se puede destacar que se sigue la tendencia de las precisiones normales que se dieron anteriormente. Con respecto a la desviación estándar, se dio un valor bajo para todos los modelos, el cual, una vez más, el con mejor resultado es el modelo de bosques aleatorios. Con respecto a una baja desviación estándar, se puede concluir que, con datos nuevos y no entrenados, el modelo probablemente lo clasifique de forma correcta, ya que, su precisión es alta y varia muy poco con respecto a todos los datos.

A continuación, se muestran los análisis de los modelos resultantes con mejor precisión y un gráfico representativo, referente a un diagrama de cajas y bigotes por cada uno, con respecto a la variable más influyente del modelo:

- **Análisis de árbol de decisión:**

Ilustración 47: Árbol de decisión general resultante (muestra referencial)



Fuente: Elaboración propia en base a KNIME

Como se puede apreciar en la Ilustración 47, el modelo de árboles de decisión realiza cálculos para delimitar fronteras de decisión en las dimensiones de los datos, que permitan clasificar a nuevos registros en una de las categorías de la clase. La idea es iterativamente, generar particiones binarias en la región de interés, buscando que cada iteración genere un subgrupo lo más homogéneo posible. Por ello, el algoritmo en este caso, el cual cuenta con 528 hojas, empieza por la categoría del "Sueldo base", como nodo raíz y la que delimita gran parte de la clasificacion. Esta separa de los que ganan menor o igual a $320.487 por el lado izquierdo y a

los que ganan más de dicho monto al lado derecho. Por el lado izquierdo, el 98,0% de los datos contenidos en la rama se clasificaron como renuncia y es lógico, debido a que es menor al sueldo mínimo que hay en Chile actualmente y en donde se presenta la mayor cantidad de rotación por motivos de renuncia. Siguiendo por la rama, se toma la categoría de "Horas de jornada", para categorizar a los registros, los cuales se dividieron por los que trabajan más de 33,5 horas por el lado derecho y al lado izquierdo se encuentran los que trabajan menos de esa hora específicamente. Por consiguiente, se divide por "Años en la empresa" en donde se encuentra la primera hoja por al lado izquierdo, los cuales tienen menos de 0,613 años (7 meses). Esta hoja contiene 218 registros, en donde clasifico a un 97,2% como renuncia. Se puede decir entonces, que, si se gana menos del sueldo mínimo, se trabaja menos de 33,5 horas y se lleva menos de 7 meses, es muy probable que el empleado renuncie a su puesto de trabajo. Las mismas conclusiones se pueden realizar al lado derecho del modelo, ya que, se va dividiendo sucesivamente de forma similar, hasta encontrar una hoja en el cual termine de clasificar los datos en sus respectivas clases.

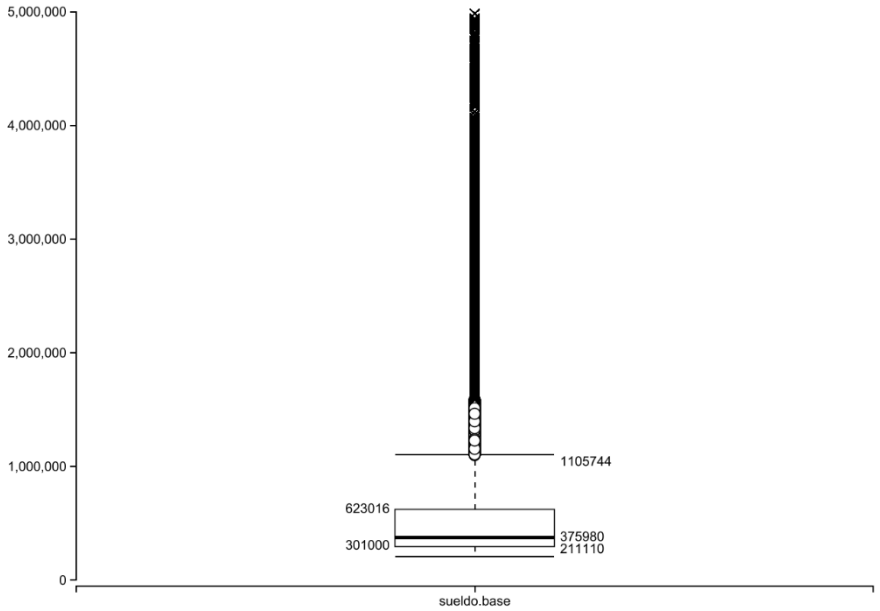
- **Análisis de bosques aleatorios:**

A diferencia del modelo de árboles de decisión, es que en lugar de entrenar un único árbol se entrenan varios, en el que cada árbol usa una parte distinta del *set* de datos original el cual se selecciona aleatoriamente. Con el bosque ya armado se hace la predicción, el cual seleccionará la clase del registro en base a la más votada por la mayoría de los árboles. Lo bueno del modelo y en ventaja al anterior, es que, si en los datos de entrenamiento hay ruido, este afectará probablemente sólo a unos cuantos árboles, pero no a la totalidad del bosque. De manera similar, las categorías que mayor se usaron para realizar las divisiones fueron el "sueldo base", "años en la empresa" y "cargo". El "sueldo base" fue candidato en el primer nivel para dividir unas 44 veces y esas mismas veces se usó, a su vez los "años en la empresa" fue candidato 40 veces y se usó 33 veces y "banco" fue candidato 38 veces y se usó 23 veces, teniendo en cuenta una totalidad de 165 modelos aleatorios distintos. Finalmente, la totalidad de árboles crearon un promedio de 5.351 nodos distribuidos en sus propias ramas.

- **Análisis del sueldo base:**

La variable más influyente fue el sueldo base, por lo cual, se procede a analizarlo gráficamente mediante el diagrama de cajas y bigotes. De igual manera que en la Ilustración 18, se considera los sueldos bajo los 5 millones, por las mismas razones descritas en tal análisis. A continuación, en la Ilustración 48, se puede apreciar el grafico de la variable en cuestión.

Ilustración 48: Diagrama de cajas y bigotes del sueldo base post preprocesamiento



Fuente: Elaboración propia

Se puede apreciar que, se considera como el primer cuartil al número 301.000, al segundo cuartil o mediana a 375.980 y al tercer cuartil a 623.016. Los valores de máximo y mínimo son 1.105.744 y 211.110 respectivamente y establecen que los números que estén fuera de tal rango son considerados atípicos. Se puede destacar que los datos se concentran en gran parte bajo los 623.016, ya que, estos representan el 75% de estos.

5.6. Segmentación

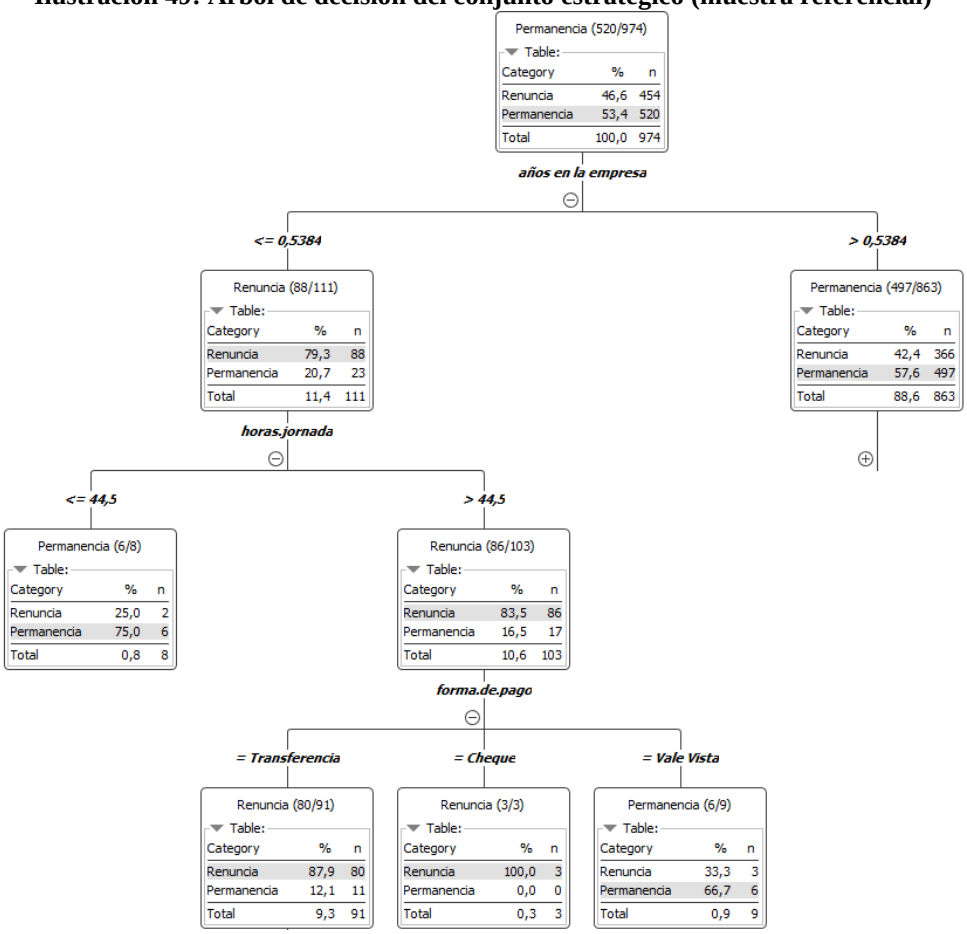
Con el nombre de "Iteración 6 (Niveles)", el objetivo de este apartado es analizar a los trabajadores por importancia, se segmentan los datos para crear nuevos modelos de predicción, en base a si el empleado pertenece a un nivel "Estratégico", "Táctico" y "Operativo". Los modelos de árboles de decisión y bosques aleatorios fueron los que dieron mejores resultados obtuvieron en la aplicación de la metodología. Adicionalmente, de la misma forma como con

el modelo general, se realiza el análisis por cada segmentación de la variable más influyente, la cual es el sueldo base para todos los casos.

5.6.1. Estratégico

Para poder analizar en que variables se basaron los modelos para predecir, se tomara como ejemplo a los árboles de decisión, la cual se presenta a continuación en la Ilustración 49 para el conjunto estratégico presentando 96 hojas.

Ilustración 49: Árbol de decisión del conjunto estratégico (muestra referencial)



Fuente: Elaboración propia en base a KNIME

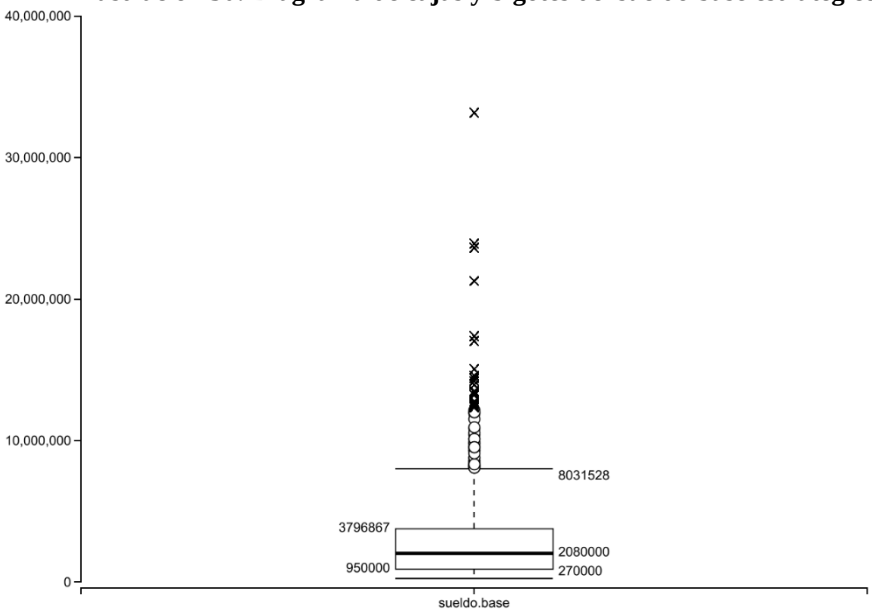
Como se puede ver, el modelo toma como nodo padre a la categoría de “Años en la empresa”. Separa de los que han estado en la empresa menor o igual a 0,5384 años, equivalente a 7 meses por el lado izquierdo y a los que han estado más de ese tiempo al lado derecho. Tomando como ejemplo el lado izquierdo, ya que presentan las primeras hojas del árbol, se tiene el 79,3% de los datos contenidos en la rama se clasificaron como renuncia para cargos altos. Acá, tomo la

variable de "Horas de jornada", con el criterio de mayor a 44,5 horas por el lado derecho y por el lado izquierdo menor o igual esta. Se puede concluir que, si se quiere predecir con el siguiente modelo a un trabajador de nivel estratégico, que haya estado menos de 7 meses y trabaje menos de 44,5 horas, es muy probable que el empleado permanezca en su puesto de trabajo.

Dicho modelo obtuvo una precisión del 66,4% y el de bosques aleatorios un 66,5%, esta baja precisión puede suceder debido a que el conjunto de datos contiene en su minoría a empleados de tipo estratégico, obteniendo un entrenamiento con muy pocos datos y con este tipo de variables, se hace complicado para el modelo discernir de buena forma, sufriendo del llamado *underfitting*.

A continuación, en la Ilustración 50, se puede apreciar el grafico de la variable de sueldo base para el nivel estratégico.

Ilustración 50: Diagrama de cajas y bigotes del sueldo base estratégico



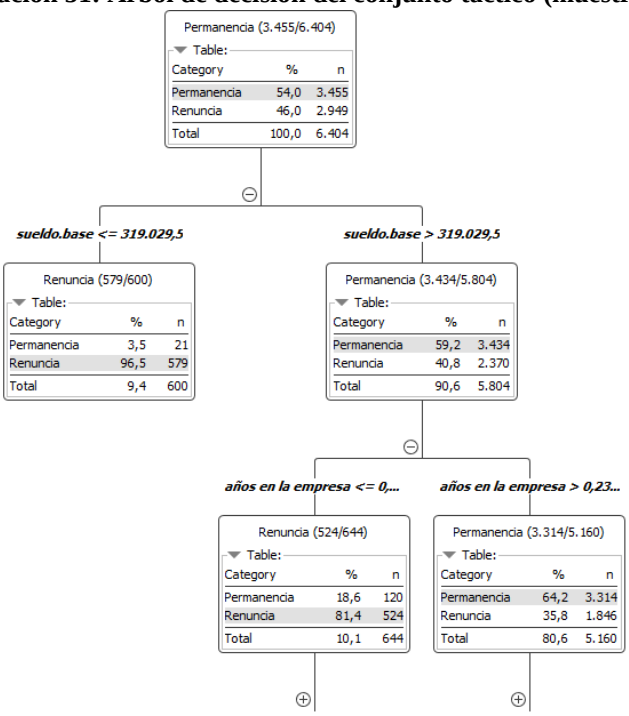
Fuente: Elaboración propia

Se puede apreciar que, se considera como el primer cuartil al número 950.000, al segundo cuartil o mediana a 2.080.000 y al tercer cuartil a 3.796.867. Los valores de máximo y mínimo son 8.031.528 y 270.000 respectivamente y establecen que los números que estén fuera de tal rango son considerados atípicos. Se destaca que los datos se concentran en gran parte bajo los 3.796.867, ya que, estos representan el 75% de estos.

5.6.2. Táctico

De la misma forma que antes, se tomara como ejemplo al árbol de decisión, la cual se presenta a continuación en la Ilustración 51 para el conjunto táctico presentando 153 hojas.

Ilustración 51: Árbol de decisión del conjunto táctico (muestra referencial)



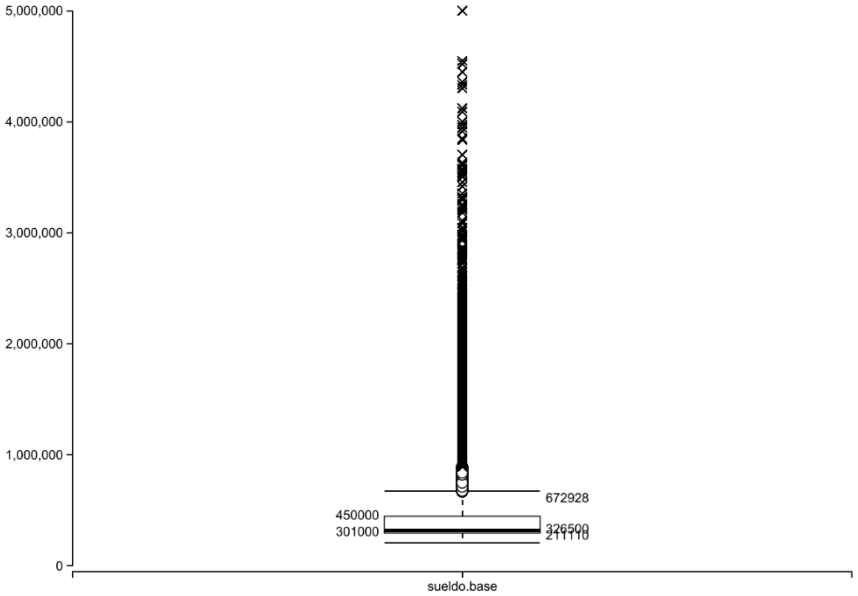
Fuente: Elaboración propia en base a KNIME

Como se puede ver, el modelo toma como referencia a la categoría de "Sueldo base". Separa a los que ganan menor o igual a $319.029 por la izquierda y a los que ganan más al lado derecho. Yendo hacia el lado izquierdo, se tiene el 96,5% de los datos contenidos en la rama se clasificaron como renuncia para cargos de nivel táctico. Como ahí se termina la rama, se considera esta una hoja la cual clasifica por el criterio del sueldo base solamente. Se puede concluir que, si se quiere predecir con el siguiente modelo a un trabajador de nivel táctico, que gane menos de $319.029, es muy probable que el empleado renuncie a su puesto de trabajo.

El modelo resultante tuvo una precisión del 72,8% y el de bosques aleatorios un 79,3%, en donde, se dio dicho resultado por como se dijo previamente, el conjunto de datos contiene en cierto grado, más información con respecto a los puestos estratégicos, obteniendo un entrenamiento con mayor cantidad de datos, por ende, aumenta el poder de predicción del modelo resultante con este tipo de variables.

A continuación, en la Ilustración 52, se puede apreciar el grafico de la variable de sueldo base para el nivel operativo, el cual también fue filtrado bajo los 5 millones como en la Ilustración 18 y Ilustración 48.

Ilustración 52: Diagrama de cajas y bigotes del sueldo base operativo



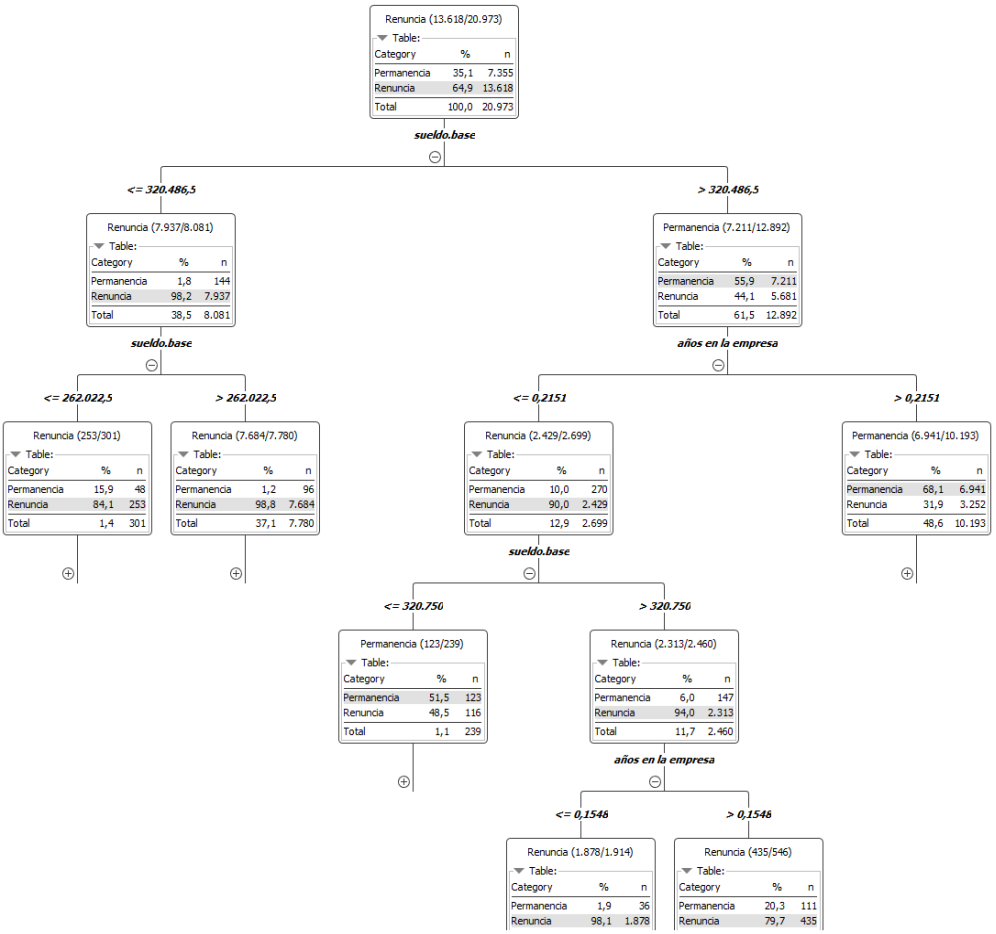
Fuente: Elaboración propia

Se puede apreciar que, se considera como el primer cuartil al número 301.000, al segundo cuartil o mediana a 326.500 y al tercer cuartil a 450.000. Los valores de máximo y mínimo son 672.928 y 211.110 respectivamente y establecen que los números que estén fuera de tal rango son considerados atípicos. Se destaca que los datos se concentran en gran parte bajo los 450.000, ya que, estos representan el 75% de estos.

5.6.3. Operativo

En el caso de los operarios, el árbol de decisión se presenta a continuación en la Ilustración 53, conteniendo 379 hojas.

Ilustración 53: Árbol de decisión del conjunto operativo (muestra referencial)



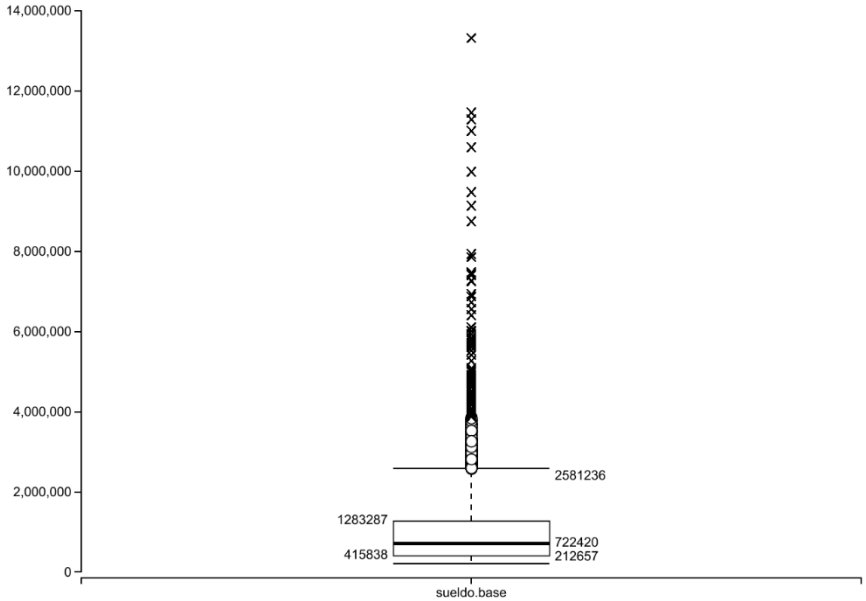
Fuente: Elaboración propia en base a KNIME

El presente árbol, su estructura es más robusta y complicado de llegar a una predicción de manera inmediata de forma manual. Pero sigue la misma lógica que las demás, separando los datos de entrenamiento y clasificando según esas separaciones o criterios. Se puede destacar, que las variables a las que mayor importancia le dio el modelo para la predicción fueron las variables de "Sueldo base" y "Años en la empresa", siendo la primera la de mayor importancia debido a que constituye al nodo padre.

El modelo resultante del conjunto tuvo una precisión del 86,2% y el de bosques aleatorios un 88,4%, por lo cual constituye modelos de buena predicción. Esto se cumple con lo descrito anteriormente, ya que, el conjunto operativo representa un porcentaje importante de los datos, el modelo tiene información suficiente para poder predecir de una forma más precisa, habiendo entrenado de buena forma con este tipo de variables en cuestión.

A continuación, en la Ilustración 54, se puede apreciar el grafico de la variable de sueldo base para el nivel estratégico.

Ilustración 54: Diagrama de cajas y bigotes del sueldo base táctico



Fuente: Elaboración propia

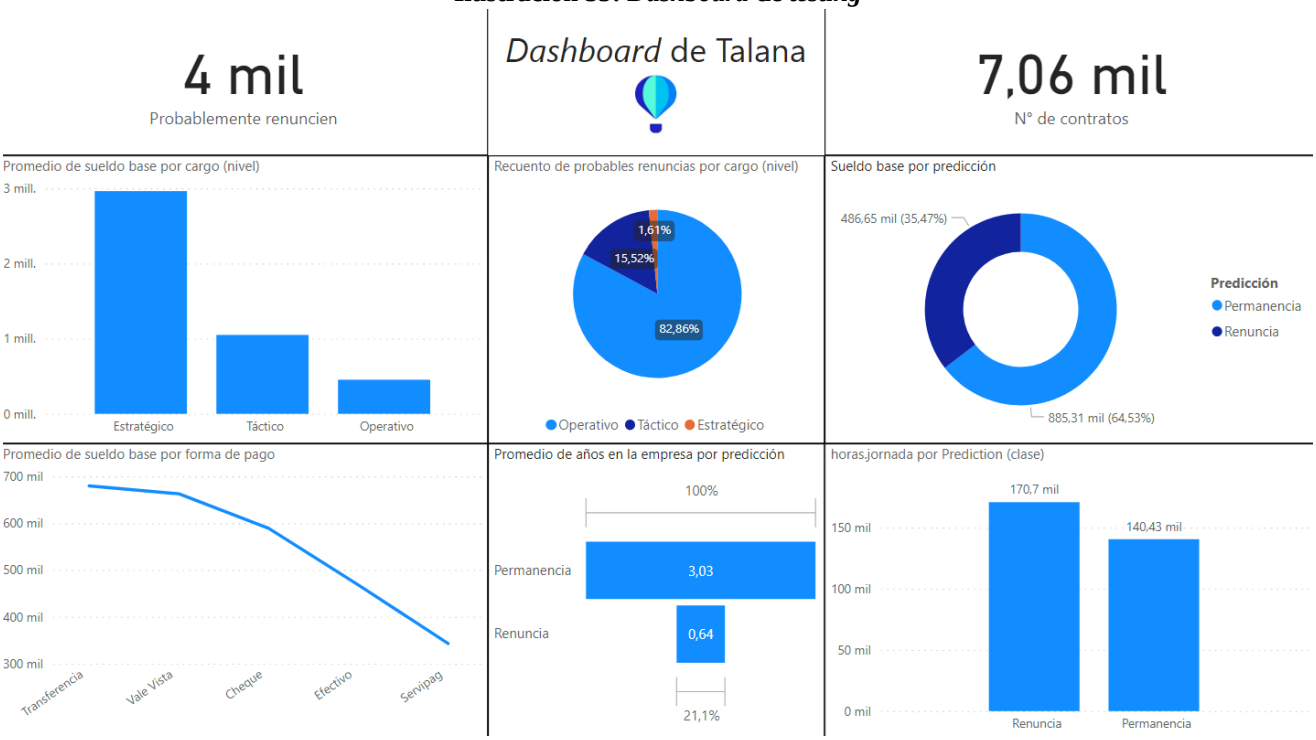
Se puede apreciar que, se considera como el primer cuartil al número 415.838, al segundo cuartil o mediana a 722.420 y al tercer cuartil a 1.283.287. Los valores de máximo y mínimo son 2.581.236 y 212.657 respectivamente y establecen que los números que estén fuera de tal rango son considerados atípicos. Se destaca que los datos se concentran en gran parte bajo los 1.283.287, ya que, estos representan el 75% de estos en general.

5.7. *Dashboard* de validación

Con los modelos ya creados, analizados y validados, se procede a crear un *dashboard* en Power BI que muestre, de manera visual, los datos resultantes que se obtuvieron en la predicción del modelo con mejor precisión (bosques aleatorios), a través de tablas y gráficos interactivos. La idea es que se muestre la información general y resumida con respecto a las predicciones y como estos se relacionan con las demás variables predictoras. Algunos de los gráficos mostrados en la Ilustración 55 son: promedio de sueldo base v/s cargo (nivel), promedio de años en la empresa v/s predicción, recuento de probables renuncias v/s cargo (nivel), entre otros.

Estos gráficos se complementan dentro de la hoja para así ofrecer un mayor entendimiento de los resultados.

Ilustración 55: *Dashboard* de *testing*



Fuente: Elaboración propia

5.8. Propuesta de implementación

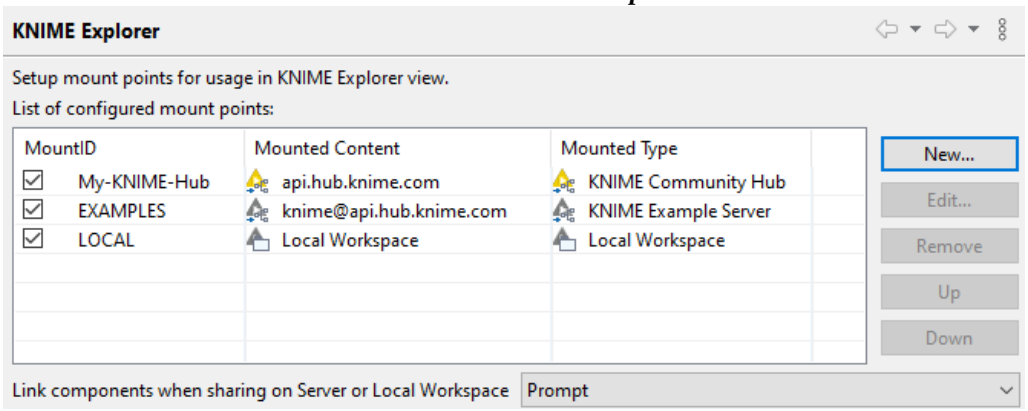
Lo ideal es que la información o predicción que se requiera por parte de una empresa, cliente de Talana, esta pueda introducir la información de dicho empleado y el modelo realice la predicción, siendo esta capturada para que se muestre en la aplicación de Talana o "Talana Next".

Por ello, en *KNIME* existe el protocolo para poder conectar los flujos de trabajo del software con un servidor *REST*, este protocolo se llama *"KNIME Server REST API"*. Para ello se configura el explorador con un nuevo espacio o *WorkSpace*, como se muestra en la Ilustración 56 y Ilustración 57.

Ilustración 56: Icono para configurar el espacio de trabajo

Fuente: Elaboración propia

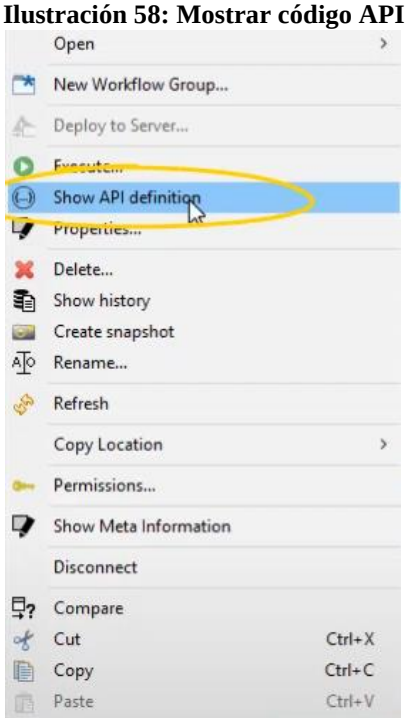
Ilustración 57: *KNIME Explorer*



Fuente: Elaboración propia

Como se muestra en la ilustración anterior, se crea un espacio nuevo con la *URL* del servidor *REST*. Acá los flujos de trabajo creados deben estar en este nuevo espacio para que estos estén conectados al servicio. Para que dicha operación funcione, deben estar los nodos llamados "*Container Input (Table)*" y "*Container Output (Table)",* los cuales se encuentran en el flujo del modelo con mayor precisión (bosques aleatorios) en la quinta iteración, para un eventual requerimiento para el *web service.*

Ilustración 58: Mostrar código API



Fuente: Elaboración propia

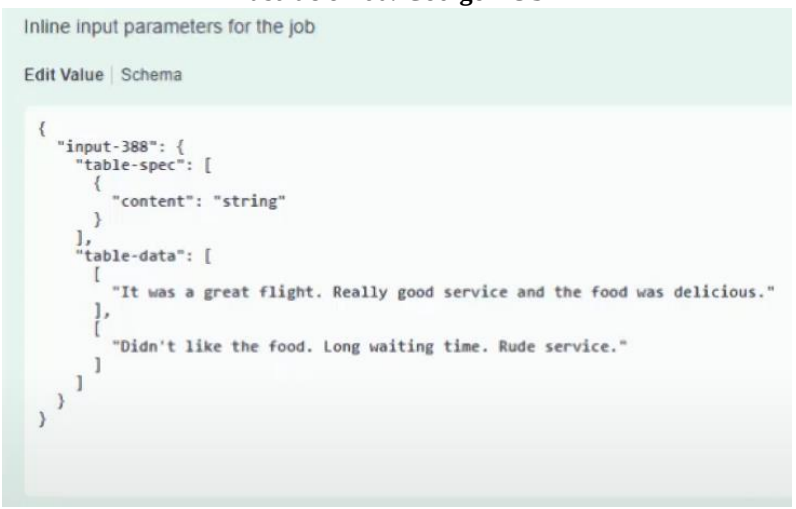
En la Ilustración 58, se ve como obtener los códigos, cliqueando el flujo donde se ubica el modelo, los cuales interactúan para enviar y recibir información, como lo son el *GET* y *POST* de la Ilustración 59, los cuales se obtienen del *host* de *Swagger*.

Ilustración 59: Mostrar código

Fuente: Elaboración propia

En la Ilustración 60, se ve el ejemplo de un código para ingresar los datos que se requieren. Dicha tabla tiene una columna de tipo *string.*

Ilustración 60: Código POST



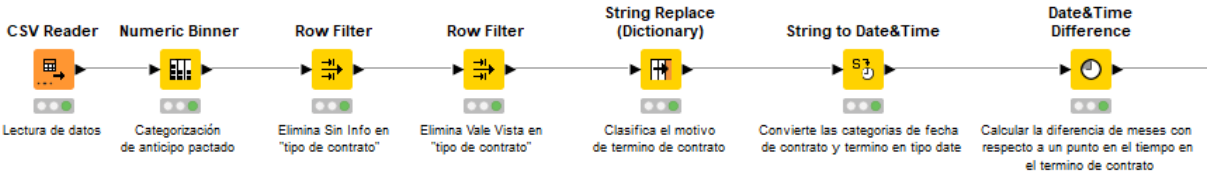
Fuente: Elaboración propia

5.9. Predicción en fecha

Aparte de lo ejecutado en las secciones anteriores, también se quiso realizar otro enfoque para la empresa Talana. Este enfoque busca la predicción en función del tiempo, es decir, cuando un trabajador es más probable que renuncie.

Para ello, se realiza un tratamiento de datos similar al proceso anterior, finalizando con la categorización de la clase, el cual se hace referente a los años en que el trabajador duro en su puesto. En la Ilustración 61 se muestra parte del proceso.

Ilustración 61: Preparación de datos para predecir fecha de renuncia



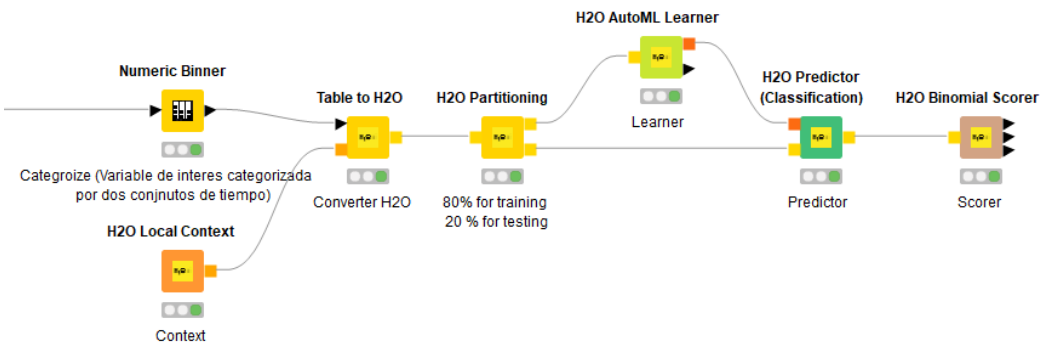
Fuente: Elaboración propia

El objetivo es que, con datos nuevos, se haga una predicción de cuando renunciaría si es que se diera el caso, en un periodo o rango de tiempo. Mencionado lo anterior, se procede a utilizar los nodos requeridos para minar los datos y así poder obtener resultados.

Antes de precisar los procesos realizados, se debe mencionar que se intentó con varios tipos de clase, de los cuales, el que dio mejor resultado fue categorizar binariamente la variable en que renunciara en "Menos de 6 meses" y "Mayor a 6 meses".

Primero, a cambio de los nodos que se habían utilizado anteriormente, en este enfoque se optó por utilizar el "AutoML" de H2O como se muestra en la Ilustración 62, el cual procesa y optimiza varios algoritmos de *machine learning* a la vez (*Random Forest, Gradient Boosting Machine, Deep Learning*, etc.), mostrando el mejor resultado de estos. Un contra de este método es que no precisa el algoritmo, solo muestra los mejores resultados obtenidos.

Ilustración 62: Nodos de AutoML

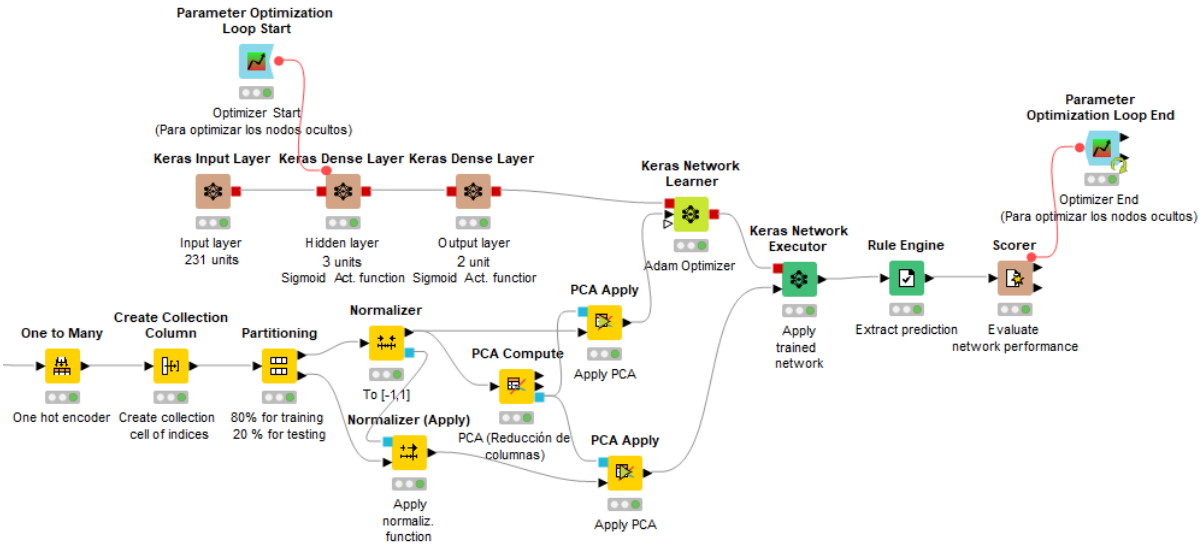


Fuente: Elaboración propia

Acá, la precisión mayor que obtuvieron los algoritmos fue de un 80,3%, prediciendo de mejor manera los que renunciaran en menos de 6 meses desde su contratación, ya que, dicha categoría obtuvo el menor error en los resultados obtenidos.

También se realizó la prueba de los resultados que nos podría dar minar datos con el algoritmo de Keras como se muestra en la Ilustración 63, lo cual ejecuta una red neuronal para dicha tarea. Para su ejecución, se convirtió a las columnas de tipo *string*, en numéricas, dando así mayor cantidad de columnas. Con el objetivo de reducir la cantidad de columnas, se utilizó el método PCA para la reducción de variables y así el algoritmo tuviera menos carga de procesamiento. Se utilizaron 231 nodos de entrada, 3 nodos ocultos y 2 nodos de salida, con la función de activación del tipo Sigmoide.

Ilustración 63: Nodos de *Deep Learning Keras*



Fuente: Elaboración propia

Acá, la precisión mayor que se obtuvo fue de un 77,9%, teniendo menor precisión que en el anterior procedimiento. Para aumentar esta precisión, se optimizo la cantidad de nodos ocultos como se muestra en la, obteniendo un total de 33 capas.

Ilustración 64: Optimización de capas ocultas

Row ID	I Capas	D Objective value
Best parameters	33	0.784

Fuente: KNIME

Con dicho número de capas ocultas se obtiene una precisión de 78,4% aumentando ligeramente, pero sigue siendo menor que antes, por lo que se recomienda seguir el anterior procedimiento para mejores resultados.

Con ello, se puede saber aproximadamente si un trabajador prolongara su estadía en la empresa por más de seis meses, o en cambio renunciara antes de ese periodo, teniendo que gastar tiempo y dinero en la contratación del puesto demasiado luego a lo que se tiene previsto, por ello con dicha información se fortalece la toma de decisión de recursos humanos y de cómo afrontar la situación del empleado en cuestión.

CAPÍTULO 6:

EVALUACION DE IMPACTO

6.1. Impacto económico

En general, el presente proyecto busca impactar de manera sostenida e importante en el aspecto económico de las empresas y de Talana. Por un lado, significando un ahorro en costos por parte de las empresas y por otro, en la obtención de ganancias por la información entregada a estas, en términos de mayor demanda o cobrando por esta. Como se menciona en el capítulo 1.2, se busca la predicción de la fuga de trabajadores mediante técnicas de minería de datos y *machine learning*, para las empresas con tal de transformar los datos en dinero. Dicho beneficio económico, es lo que se calculará en el presente capitulo, con tal de estimar el impacto que se tendrá a la hora de que el proyecto este implementado.

Con respecto al cálculo, es importante destacar que el beneficio por retener a un trabajador importante dentro de las empresas no es constante, es decir, varia por el tipo de cargo que ostenta cada uno. Un empleado que pertenece al área financiera de la empresa y resuelve gran parte de los dineros de esta, costara más que un empleado que se dedique a la limpieza y aseo, en términos de reclutamiento y aprendizaje. Por ende, se llevará a cabo el cálculo de estos costos, que a su vez será el beneficio de las empresas por ahorrarse dichos costos, dividiéndolos en categorías como se realizó anteriormente, por niveles organizacionales, destacando a los cargos más influyentes o de mayor ocurrencia dentro de las empresas.

En el primer nivel, que corresponde al estratégico, se destacan los cargos de director, gerente y decano. En dicho nivel, se determina que el aprendizaje total se alcanza a los 6 meses, la cual corresponde a lo que demora en ser productivo completamente el trabajador en su cargo (Ramírez, 2020). Mediante reunión con el CTO de Talana, se estableció la Ecuación 2, Ecuación 3, Ecuación 4 y Ecuación 5, las cuales muestran los cálculos del ahorro de entrenamiento por cargo al mes, lo que significa el beneficio de haber tomado medidas para permanencia del trabajador en el puesto en cuestión.

Ecuación 2: Renuncia promedio del cargo al mes

$$RP_{Cargo}\ (Mes) = RP_{Cargo}\ (Año) * \frac{12\ Meses}{1\ Año}$$

Fuente: Elaboración propia en base a Talana

Ecuación 3: Ahorro en entrenamiento de cargo estratégico al mes

$$AE_{Cargo\ Estratégico}/Mes = \frac{SP_{Cargo}\ (\$CLP) * T_{Estratégico}}{RP_{Cargo}\ (Mes)}$$

Fuente: Elaboración propia en base a Talana

Ecuación 4: Ahorro en entrenamiento de cargo táctico al mes

$$AE_{Cargo\ Táctico}/Mes = \frac{SP_{Cargo}\ (\$CLP) * T_{Táctico}}{RP_{Cargo}\ (Mes)}$$

Fuente: Elaboración propia en base a Talana

Ecuación 5: Ahorro en entrenamiento de cargo operativo al mes

$$AE_{Cargo\ Operativo}/Mes = \frac{SP_{Cargo}\ (\$CLP) * T_{Operativo}}{RP_{Cargo}\ (Mes)}$$

Fuente: Elaboración propia en base a Talana

Donde, $AE =$ ahorro de entrenamiento, $SP =$ sueldo promedio, $RP = renuncia\ promedio$ y $T =$ tiempo de entrenamiento.

A continuación, como ejemplo, se muestra el cálculo del ahorro de un gerente en la Ecuación 6 y los demás resumidos en la Tabla 9.

Ecuación 6: Ahorro en entrenamiento de gerente al mes

$$RP\ (Mes) = 4{,}38\ Años * \frac{12\ Meses}{1\ Año} = 52{,}56\ Meses$$

$$AE/Mes = \frac{4.081.810\ \$CLP\ * 6}{52{,}56\ Meses} = 465.960 \frac{\$CLP}{Mes}$$

Fuente: Elaboración propia

Tabla 9: Ahorro en entrenamiento de cargos operativos

Cargo	Salario promedio ($CLP)	Renuncia promedio (Meses)	Ahorro en entrenamiento/Mes
Director	2.277.752	37,32	366.198
Gerente	4.081.810	52,56	465.960
Decano	1.569.916	34,98	269.282

Fuente: Elaboración propia

En el segundo nivel, que corresponde al táctico, se destacan los cargos de jefe, *product owner*, ingeniero, líder, capitán e instructor. En dicho nivel, se estima que el aprendizaje total se alcanza a los 6 meses para ser productivo (Ramírez, 2020). En modo de ejemplo, se muestra el cálculo del ahorro de un jefe en la Ecuación 7 y los demás resumidos en la Tabla 10.

Ecuación 7: Ahorro en entrenamiento de jefe al mes

$$RP\ (Mes) = 2{,}825\ Años * \frac{12\ Meses}{1\ Año} = 33{,}90\ Meses$$

$$AE/Mes = \frac{1.442.146\ \$CLP\ * 6}{33{,}90\ Meses} = 255.247\ \frac{\$CLP}{Mes}$$

Fuente: Elaboración propia

Tabla 10: Ahorro en entrenamiento de cargos tácticos

Cargo	Salario promedio ($CLP)	Renuncia promedio (Meses)	Ahorro en entrenamiento/Mes
Jefe	1.442.146	33,90	255.247
Pr. Owner	2.454.865	27,00	545.526
Ingeniero	1.780.236	24,00	445.060
Líder	1.636.519	31,07	316.052
Capitán	815.471	20,51	238.582
Instructor	1.026.086	6,50	946.574

Fuente: Elaboración propia

En el tercer y último nivel, que corresponde al operativo, se destacan los cargos de vendedor, operador, tens, soldador, utilero, conductor, bodeguero, vigilante y reponedor. En dicho nivel, se estima que el aprendizaje total o entrenamiento se alcanza al tercer mes para ser productivo (Marriaga, 2017). En modo de ejemplo, se muestra el cálculo del ahorro de un vendedor en la Ecuación 8 y los demás resumidos en la Tabla 11.

Ecuación 8: Ahorro en entrenamiento de vendedor al mes

$$RP\ (Mes) = 1{,}159\ Años * \frac{12\ Meses}{1\ Año} = 13{,}91\ Meses$$

$$AE/Mes = \frac{369.644\ \$CLP\ * 3}{13{,}91\ Meses} = 79.722\ \frac{\$CLP}{Mes}$$

Fuente: Elaboración propia

Tabla 11: Ahorro en entrenamiento de cargos operativos

Cargo	Salario promedio ($CLP)	Renuncia promedio (Meses)	Ahorro en entrenamiento/Mes
Vendedor	369.644	13,91	79.722
Operador	344.519	5,59	184.827
Tens	433.072	20,21	64.293
Soldador	503.662	4,69	322.035
Utilero	476.796	6,06	236.037
Conductor	417.651	13,76	91.032
Bodeguero	338.226	8,84	114.732
Vigilante	314.946	23,00	41.073
Reponedor	271.278	13,04	62.391

Fuente: Elaboración propia

Posterior a los cálculos realizados, se busca establecer una métrica que pueda abarcar de manera general el beneficio del proyecto en la perspectiva de los clientes de la empresa, por ello, se define el ahorro total promedio que podría haber tenido una empresa por la obtención de dicha información y posterior toma de medidas para la retención de los trabajadores importantes.

Este número sirve para medir el beneficio del ahorro a partir de los datos históricos, el cual puede distar en amplio margen al beneficio real que pueda haber tenido cada empresa, debido a las estimaciones realizadas y de que depende ampliamente de la cantidad de trabajadores que se consideran importantes y del tipo de cargo que ostentan. Teniendo claro lo anterior, se comienza calculando en una nueva columna el ahorro de entrenamiento mensual (AE/Mes) de cada trabajador existente en los datos que haya renunciado. Posteriormente, se estima que el total de estos trabajadores, solo un 5% fueron considerados importantes para sus respectivas empresas, por diversas razones como buen rendimiento de sus funciones, puntualidad, entre otros. Luego, se agrupa por cada empresa, la suma del ahorro de entrenamiento que hubiesen tenido al retener la totalidad de sus trabajadores más importantes, considerando la mitad de las empresas en cuestión, debido a que son las que estarían interesadas en contar con la predicción a partir de los datos. Finalmente, a partir de lo anterior, se obtiene la métrica anteriormente mencionada del ahorro total promedio por empresa, la cual da un estimado de 1.928.137 $CLP/Mes. Cabe destacar, que la toma de medidas para la retención de los empleados importantes es probable que no siempre resulte como se requiere, es decir, que de igual forma el trabajador renunciara, ya que, puede tener razones muy fuertes para tomar dicha decisión.

6.2. Impacto social

El impacto económico es importante para comprender los efectos relacionados con el dinero que resultan de la implementación del proyecto. El dinero no es la única variable que incide en esta fase, por ello, se hace importante analizar dicho efecto para el presente capitulo. El efecto o impacto al que se hace referencia es al impacto social, que, en general de los proyectos, explica las consecuencias que traerá la implementación para una cierta comunidad. En este caso, el efecto que causara es netamente hacia los trabajadores de las empresas que son clientes

de Talana, por lo tanto, se realiza un análisis enfocado a como impactara en sus trabajos, socialmente hablando, la retención de los empleados importantes.

En el contexto anteriormente mencionado, es complejo poder determinar algún beneficio que fuese cuantitativo de este tipo de impacto, por ende, se determina los elementos que afectan la rutina de los trabajadores de manera cualitativa, esto en base a la "Gestión del talento y técnicas de retención de personal clave" (Rojo, 2014) . Las variables sociales identificadas en la implementación de medidas para la retención de trabajadores importantes y disminuir la rotación no deseada a partir de la predicción de los datos son:

- **Ambiente agradable:** la retención de trabajadores importantes y ante las medidas tomadas, permite no sólo conservar a un empleado de buen rendimiento, sino afectar de manera directa al ambiente de trabajo. Esto mejoraría en la eficiencia del desarrollo de las actividades, en la interacción entre los miembros del equipo de trabajo y en el comportamiento afuera de este, ya que, en un mal ambiente, se nota en el comportamiento con los demás a la hora de una jornada finalizada.
- **Mayor compromiso:** a medida en que aumente el interés con los trabajadores, en sus trabajos y haya un mayor compromiso, automáticamente se crea un vínculo más fuerte con la organización, bajando las probabilidades de renuncia. Este compromiso también se traduce en intensidad y mayor esfuerzo del trabajador en no defraudar en lo que realiza y también en una mayor identificación con la organización, teniendo un sentimiento positivo hacia esta, dependiendo de su percepción en el trabajo. Psicológicamente, el trabajador estará a gusto, ya que, sabe con mayor seguridad de que su trabajo es importante para la empresa, afectando también a su entorno, no solo en el trabajo, sino en su vida cotidiana y en la relación con sus cercanos.
- **Satisfacción laboral:** en relación con el punto anterior, el cual conecta con el presente concepto, el trabajador estará más cómodo con su trabajo mediante las medidas tomadas, es decir, tendrá mayor satisfacción laboral. Se puede decir que la motivación y satisfacción son aspectos complementarios dentro de una organización. Se motiva mediante la satisfacción de necesidades de distinta índole, las cuales al ser superadas generan nuevas necesidades las que nuevamente buscan ser satisfechas generando el sentido de motivación.

- **Tiempo disponible:** al tener una flexibilidad de horarios como medida, repercute de forma directa sobre la calidad de vida de los profesionales de la organización y de sus familias. Las dificultades habituales en cualquier vida familiar como llevar o recoger los niños del colegio, ir al médico, realizar actividades deportivas, atender a las necesidades de los mayores, hacer compras o gestionar cualquier otro asunto de carácter personal pueden reducirse de forma considerable.

CONCLUSIONES

Como conclusión, es más que sabido que las organizaciones, cada vez buscan en como implementar nuevas metodologías y construir proyectos para la utilización de sus datos a su beneficio, tanto interno como externo de la organización. Por esa línea, la utilización de herramientas de minería de datos como el *machine learning* para mejorar su servicio ayudarán de buena forma a Talana a dar un paso más en su relación con la tecnología actual.

Como acercamiento, se realizó el diagnóstico de la situación actual, el cual se pudo determinar que los datos que la empresa obtiene de sus clientes lo hacía con el sistema de *Web Service REST,* la cual estandariza la obtención de los datos para todos los clientes. Determinando la realización del trabajo de minería de datos, ya que, se tiene gran parte de las variables a pesar de que la calidad de estos no fuera la óptima.

Se puede destacar que se llevó a cabo las etapas de la metodología propuesta, tales como el descubrimiento de trabajo, enfoque analítico, recursos de datos, preparación de datos, construcción del modelo, evaluar modelo, aprender actualizando y evaluar el impacto. La flexibilidad de esta permitió construir un trabajo mediante iteraciones, volviendo hacia atrás, evaluando los cambios que presentaban los modelos por diferentes tipos de tratamiento en los datos de entrada, tales como, agregación de datos, imputación de datos, eliminación de *outliers*, etc.

Se comprueba y se hace énfasis en que el tiempo de trabajo se gasta mayoritariamente en el preprocesamiento de los datos y para que estos tengan la mayor calidad posible al entrar en los modelos de predicción. En el presente caso, se recalca que la calidad de los datos que se dispuso para dicho trabajo no fue de la calidad esperada, teniendo que preprocesarlos de forma intensiva y sistemática. Además, no se pudo tener en posesión datos de tipo psicométrico, los cuales habrían sido información más que útil en la construcción de los modelos y su optimización en las predicciones finales.

El programa KNIME, ayudo al cumplimiento de los procedimientos de la metodología, ya que, abarca cada una de ellas en su entorno de ejecución. Dicho *software,* fue útil a la hora de su funcionamiento, permitiendo realizar rápidos análisis, debido a la utilización de nodos

que lo hacía más claro en la construcción de diferentes procesos. Cada nodo está relacionado con las etapas metodológicas, permitiendo efectuar en los datos su preparación mediante procesos ETL (*Extract, transform and load*), visualización, construcción de modelos, evaluación e implementación de estos. A su vez, la adaptación fue rápida por dicho motivo y probablemente, fuera requerir de mayor tiempo con programas que usaran la algoritmia, a pesar de conocerlos de cerca. Para alguien que empieza en el mundo del *machine learning* y no esté ligado a la programación, se recomienda dicho programa por lo comentado anteriormente.

Por consiguiente, se eligió el modelo de bosques aleatorios para este tipo de datos, con una precisión de 85,5%, una precisión media de validación cruzada de 85,8% y desviación estándar de 0,005, demostrando una alta precisión y buen ajuste a nuevos datos que el modelo se enfrente. En segmentación, se recomienda utilizar el modelo referente a cargos operativos por la alta precisión y mejor entrenamiento con respecto a los demás. Se puede decir que el algoritmo de bosques aleatorios se ajustó bien, debido a que el conjunto de datos es grande, especial para estos casos y de que funciona para problemas lineales y no lineales.

También, se realizó un enfoque distinto para la empresa Talana. Este enfoque predice en función del tiempo, es decir, cuando un trabajador es más probable que renuncie. Para ello se utilizaron dos métodos, AutoML y Deep Learning Keras. Estos resultaron con una predicción del 80,3% y 78,4% respectivamente, siendo mejor el primero.

Se realiza una propuesta de implementación para del proyecto mediante el servicio REST, que es el mismo que se tiene en la empresa, con tal de que sea de una manera rápida y se pueda mantener a lo largo del tiempo, pudiendo interactuar con los clientes y la aplicación de la empresa.

Finalmente, se realiza la evaluación de impacto del proyecto, en términos económicos y sociales. Se pudo determinar los beneficios que tendrían las empresas al obtener la información de la predicción de la renuncia de sus empleados, los cuales dependerán de cuantos empleados son importantes para estas y el tipo de cargo que ostentan. Se considero como beneficio global a la suma del ahorro promedio de las empresas, resultando un valor de 1.928.137 $CLP/Mes. En cuanto al análisis del impacto social, se logra establecer que la implementación del presente proyecto busca mejorar ciertas variables que son importantes para

los trabajadores en términos sociales dentro y fuera de la empresa. Se establece parámetros cualitativos dentro del trabajo, esto a raíz de un mayor bienestar social con los pares empleados y sus familias.

Se recomienda del presente trabajo, poder agregar variables de tipo psicométrico como se menciona anteriormente, para poder verificar los rendimientos de los modelos y de cómo se comportan con las nuevas variables agregadas y la construcción de sus relaciones. También seguir investigando sobre si poder identificar en qué momento del tiempo o en cuanto renunciara un empleado, si es posible, clasificando la clase de una manera distinta o agregando otro tipo de variables que ayuden a poder determinar los posibles tiempos en que un trabajador renuncie de su cargo. Por último, se recomienda tomar la propuesta de implementación para ponerla en ejecución, la cual se adecua al sistema actual.

BIBLIOGRAFÍA

Ahmed, B. (2018). *A Lean Design Thinking Methodology (LDTM) for Machine Learning and Modern Data Projects.*

Aprendesas. (2015). *Arboles de decisión en SAS*. Obtenido de http://sasybi.blogspot.com/2015/05/arboles-de-decision-en-sas.html

Dai, W., & Zhu, Z. (2020). *Employee Resignation Prediction Model Based on Machine Learning.*

Gandhi, R. (2018). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Obtenido de https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Gartner. (2018). *Herramientas del Data Mining.* Obtenido de https://www.lisdatasolutions.com/blog/herramientas-del-data-mining/

Gartner. (2020). *gartner.com*. Obtenido de https://www.gartner.com/en/information-technology/glossary/big-data

Gartner. (2021). *2021 Gartner Magic Quadrant for Analytics and Business Intelligence Platforms.* Obtenido de https://info.microsoft.com/ww-Landing-2021-Gartner-MQ-for-Analytics-and-Business-Intelligence-Power-BI.html?LCID=EN-US

Gascó, T. (2019). *Definición de Teorema de Bayes*. Obtenido de https://www.economiasimple.net/glosario/teorema-de-bayes

IBM. (2019). *ibm.com*. Obtenido de https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview

Información de Mercados. (2017). *Información de Mercados*. Obtenido de https://www.informaciondemercados.cl/talana-plataforma-chilena-que-simplifica-la-gestion-de-recursos-humanos/

Jara, R. (2015). *Detección de fuga de operarios en una empresa del sector minero utilizando minería de datos.* Santiago.

KD Nuggets. (2014). *KD Nuggets*. Obtenido de https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Martínez, C. (2012). *Aplicación de técnicas de minería de datos para mejorar el proceso de control de gestión en entel.* Santiago.

Microsoft. (s.f.). *Power BI reports,filters and highlighting.* Obtenido de https://docs.microsoft.com/es-es/power-bi/create-reports/power-bi-reports-filters-and-highlighting

Molina, I. (2020). *¿Cómo aprende una red neuronal de IA?* Obtenido de https://es.quora.com/C%C3%B3mo-aprende-una-red-neuronal-de-IA

Onuralp, I. (2017). *An Approach for Predicting Employee Churn by Using Data Mining.*

Parra, F. (2019). *Estadística y Machine Learning*. Obtenido de https://bookdown.org/content/2274/portada.html

Rojo, F. (2014). *Gestión del talento y técnicas de retención de personal clave.* Concepción.

Romero, E. (2013). Obtenido de https://estebanromero.com/2013/05/design-thinking-una-vision-global/

Salcedo, L. (2018). *Introducción al Machine Learning #9 - K Vecinos más cercanos (Clasificación y Regresión)*. Obtenido de https://pythondiario.com/2018/01/introduccion-al-machine-learning-9-k.html

Salunkhe, T. P. (2018). *Improving Employee Retention by Predicting Employee Attrition using Machine Learning.*

Staryfurman, L. (s.f.). *The Power MBA*. Obtenido de https://www.thepowermba.com/es/business/design-thinking/

Talana. (2021). *Talana*. Obtenido de https://talana.com/es/clientes/

U. de Alcalá. (2021). *Master data scientist*. Obtenido de https://www.master-data-scientist.com/que-es-masters-in-data-science/

Vallés, J. (2013). *https://josepvalles.wordpress.com*. Obtenido de https://josepvalles.wordpress.com/2013/12/29/chuleta-guia-tutorial-open-refine-google-refine/

XMS. (s.f.). *¿Qué es Power BI y cuáles son sus características?* Obtenido de https://www.xms.cl/que-es-power-bi-y-cuales-son-caracteristicas/

ANEXOS

Anexo 1: *Endpoint* de "Persona"

Propiedad	Tipo GET	Tipo POST, PUT	Uso	Opcional	Descripción
id	integer				ID único
fechaCreacion	datetime		GET	Si	Fecha y hora de creación
rut	string	string	GET, POST, PUT	No	Rut, sin puntos, con guion y DV
nombre	string	string	GET, POST, PUT	No	Nombre
apellidoPaterno	string	string	GET, POST, PUT	No	Apellido Paterno
apellidoMaterno	string	string	GET, POST, PUT	No	Apellido Materno
sexo	char(1)	char(1)	GET, POST, PUT	Si	M/F
fechaNacimiento	date	date	GET, POST, PUT	Si	Fecha de nacimiento
nacionalidad	string	string	GET, POST, PUT	Si	Pais, como código ISO
username	string	string	GET, POST, PUT		Nombre de usuario
permisos	list[permisos]		GET	Si	Lista de permisos adicionales
email	string	string	GET, POST, PUT	Si	Email principal
detalles	list[detalles]		GET, POST	Si	Detalles personales
permisos	list[permiso]		GET, POST	Si	Lista de permisos de usuario
externalReference	list[externalReference]		GET	Si	Lista de identificadores

Fuente: elaboración propia en base a Talana

Anexo 2: Objeto "Detalle" proveniente de "Persona"

Propiedad	Tipo POST	Uso	Opcional	Descripción
id				ID único
fechaCreacion	date	GET, POST	Si	Fecha y hora de creación
validoDesde	date	GET, POST	Si	Fecha desde la cual es válido el detalle
email	string	GET, POST	Si	Email principal del detalle
telefono	string	GET, POST	Si	El teléfono del detalle de la persona
celular	string	GET, POST, PUT	Si	Celular del detalle de la persona
direccionCalle	string	GET, POST	Si	Calle del domicilio
direccionNumero	string	GET, POST	Si	Numero de domicilio
direccionDepartamento	string	GET, POST	Si	Número del departamento
direccionComuna	integer	GET, POST	Si	Puntero a [unidadOrganizacional]
direccionCiudad	integer	GET, POST	Si	Puntero a [unidadOrganizacional]
estadoCivil	string	GET, POST	Si	Estado civil de la persona
nivelEducacional	string	GET, POST	Si	Nivel educacional de la persona
colegio	string	GET, POST	Si	Colegio de la persona
institucionEstudiosSuperiores	string	GET, POST	Si	Institución superior de la persona
profesión	string	GET, POST	Si	Profesión de la persona
observaciones	string	GET, POST	Si	Agregar cualquier observación sobre la persona
contratosDeEmergencia	string	GET, POST	Si	Agregar cualquier contrato de emergencia de la persona

Fuente: elaboración propia en base a Talana

Anexo 3: Objeto "Permiso" proveniente de "Persona"

ID	Nombre	Descripción
1	ver_asistencia_mobile	Mostrar la sección de marcación de asistencia en el App
2	marcacion_libre	Permitir marcar sin estar presente en la sucursal
3	marcacion_root	Permitir marcar, aunque el teléfono esté hackeado, y se posible falsear la ubicación del GPS
4	permitir_ingreso_talana	Permitir ingresar a la plataforma y al app. Sin este permiso, el usuario no se puede loguear
5	permitir_marcacion_mobile	Permitir marcar desde el móvil
6	permitir_contratos_paralelos	Permitir más de un contrato paralelo

Fuente: elaboración propia en base a Talana

Anexo 4: *Endpoint* de "Contratos"

Propiedad	Tipo GET	POST, PUT	Opcional	Descripción
id	integer		No	Id único
empleado	integer	integer	No	Puntero a objeto [persona]
codigo	string	string	Si	Código del contrato
fechaCreacion	datetime		Si	Fecha de Creación original de estas condiciones contractuales
tipoContrato	integer	integer	Si	Puntero a objeto [tipoContrato].

empleadorRazonSocial	integer	integer	No	Puntero a objeto [razonSocial]
cargo	string	string	Si	
fechaContratacion	date	date	Si	Fecha de contratación original del trabajador
desde	date	date	No	Desde cuándo rigen estas condiciones contractuales
hasta	date	date	Si	Hasta cuándo rigen estas condiciones. Puede estar vacío
unidadOrganizacional		integer	Si	Puntero a objeto [unidadOrganizacional]
sucursal	[sucursal]	integer	Si	Objeto sucursal
grupos	list[grupo]	list[integer]	Si	Lista de los grupos al cual pertenece la persona
anexo	string	string	Si	
centroCosto	[centroCosto]	integer	Si	Objeto centroCosto
jornada	integer	integer	Si	Puntero a objeto [jornada]
horasDeLaJornada	integer	integer	Si	Cantidad de horas semanales de trabajo
codigoFranquiciaSence	integer	integer	Si	
nivelSence	string	string	Si	
sindicato	integer	integer	Si	Puntero a objeto [sindicato]
jefe	integer	integer	Si	Puntero a objeto [persona]
esPensionado	char(1)	char(1)	Si	N=No | S=Si | C=Si, pero cotiza | A = Activo > 65 años | X = Expatriado
tramoAsignacionPrevisional	integer	integer	Si	Puntero a objeto [tramoAsignacionFamiliar]
zonaAsignacionPrevisional	integer	integer	Si	Puntero a objeto [ubicacionGeografica]
correspondeAsignacionMaternal	boolean	boolean	Si	Si es que tiene asignación maternal
isapre	integer	integer	Si	Puntero a objeto [prevision]
montoPactadoIsapre	float	float	Si	Monto Pactado con Isapre. Si está vacío, se asume 7%
montoPactadoIsapreMoneda	string	string	Si	Moneda monto pactado. UF=UF | $=Pesos \| 7+GES=7%+Ges en UF \| 7+GES$=7% + GES en pesos
afp	integer	integer	Si	Puntero a objeto [afp]
adscribeASeguroCesantiaParaContratosPreviosA2002	boolean		No	Si el trabajador está contratado con fecha anterior al 2002, si adscribe o no al seguro de cesantía
apvMonto	float		Si	Monto primer APV
apvMoneda	string		Si	Moneda APV: UF | $
apvInstitucion	integer		Si	Puntero a [institucionAPV]
apvTipo	char(1)		Si	Tipo de APV: “A” | “B”
apvCuentaDos	float		Si	Monto cuenta dos
apvCuentaDosMoneda	string		Si	Moneda cuenta dos
depositoConvenidoMonto	float		Si	Monto Depósito Convenido
depositoConvenidoMoneda	string		Si	Moneda Depósito Convenido: UF | $
retencionJuducialDestinatario	string		Si	Nombre de persona destinatario de retención judicial
sueldoPatronal	boolean		Si	¿Es sueldo patronal? Sólo para socios de la empresa
sueldoBase	integer	integer	Si	Sueldo Base mensual en $
sueldoFormaPago	string	string	Si	Forma de pago de sueldo.
sueldoBanco	id	id	Si	Puntero a [banco]
sueldoCuentaCorriente	string	string	Si	Número de cuenta corriente para depósito
sueldoCuentaCorrienteTipo	string	string	Si	Tipo de cuenta corriente: Cuenta Vista, Ahorro y Corriente
sueldoTipoPago	string	string	Si	Forma de cálculo sueldo: mensual | diario |hora
valorHoraExtraPactada	float	float	Si	
mesesImponiblesreconocidos	integer	integer	Si	"Meses que se reconocen como trabajados de antes de contratar a la persona. Se usan para los días progresivos
mesesImponiblesReconocidosDesde	date	date	Si	
vacacionesReconocidoDesde	date	date	Si	Beneficio. Fecha de contratación utilizada
asignacionMovilizacion	integer	integer	Si	Movilización mensual en $
asignacionColacion	integer	integer	Si	Colación mensual en $
anticipoPactado	integer	integer	Si	
fechaDeContratacionReconocidaParaAnosDeServicio	date	date	Si	Beneficio. Fecha de Contratación para utilizar para cálculo de años de servicio
pagaTresPrimerosDiasLicencia	boolean	boolean	Si	Beneficio. ¿Se subsidian los 3 primeros días de licencia?
mantieneRentaLiquidaLicencia	boolean	boolean	Si	Beneficio. ¿Mantiene la renta líquida durante las licencias?
diasAdministrativos	integer	integer	Si	Beneficio. Días administrativos por trabajador durante el periodo
beneficiosInfoAdicional	string	string	Si	String con beneficios adicionales

indemnizacionSinTopeAnos	boolean	boolean	Si	Beneficio. Al calcular la indemnización por años de servicio al finiquitar, ¿se quita el tope de años?
indemnizacionSinTopeRenta	boolean	boolean	Si	Beneficio. Al calcular la indemnización por años de servicio al finiquitar, ¿se quita el tope de renta?
diasAdicionalesVacaciones	integer	integer	Si	Beneficio. Días adicionales de vacaciones por año.
creadoPor	integer		Si	Creador del contrato. Puntero a [persona]
descripcionDelCargo	string	string	Si	Descripción en texto del cargo
clausulasAdicionales	string	string	Si	
detalleAnexoContrato	string	string	Si	
documentoEsContratoOAnexo	string	string	No	
claseSalarial	[claseSalarial]	integer	Si	
rolPrivado	boolean	boolean	Si	
asignacionZonaExtrema	float		Si	Factor de asignación adicional por zona extrema
unidadOrganizacionalDetails	[unidadOrganizacional]		Si	Objeto detallado [unidadOrganizacional]
tipoContratoDetails	[tipoContrato]		Si	Objeto detallado [tipoContrato]
externalReference	list[externalReference]		Si	Objeto detallado con los Ids de este contrato en otros sistemas
finiquitado	boolean	boolean	Si	¿Fue finiquitado?
motivoEgreso	[motivoEgreso]	integer	Si	Motivo de término de contrato (si fue finiquitado)
INE	[codigoIne]	integer	Si	Objeto detallado con el código INE del trabajador
userDefinedFields	list[userDefinedField]		Si	Listado con campos personalizados

Fuente: elaboración propia en base a Talana

Anexo 5: *Endpoint* de "Contratos v2"

Propiedad	Tipo GET	Opcional	Descripción
id	integer	No	Id único
empleado	integer	No	Puntero a objeto [persona]
codigo	string	Si	Código del contrato
fechaCreacion	datetime	Si	Fecha de Creación original de estas condiciones contractuales
tipoContrato	integer	Si	Puntero a objeto [tipoContrato]. Se especifican detalles en tipoContratoDetails
empleadorRazonSocial	integer	No	Puntero a objeto [razonSocial]
cargo	string	Si	
fechaContratacion	date	Si	Fecha de contratación original del trabajador
desde	date	No	Desde cuándo rigen estas condiciones contractuales
hasta	date	Si	Hasta cuándo rigen estas condiciones. Puede estar vacío
unidadOrganizacional		Si	Puntero a objeto [unidadOrganizacional]
sucursal	[sucursal]	Si	Objeto sucursal
grupos	list[grupo]	Si	Lista de los grupos al cual pertenece la persona
anexo	string	Si	
centroCosto	[centroCosto]	Si	Objeto centroCosto
jornada	integer	Si	Puntero a objeto [jornada]
horasDeLaJornada	integer	Si	Cantidad de horas semanales de trabajo
codigoFranquiciaSence	integer	Si	
nivelSence	string	Si	
sindicato	integer	Si	Puntero a objeto [sindicato]
jefe	integer	Si	Puntero a objeto [persona]
esPensionado	char(1)	Si	N=No | S=Si | C=Si, pero cotiza | A = Activo > 65 años | X = Expatriado
tramoAsignacionPrevisional	integer	Si	Puntero a objeto [tramoAsignacionFamiliar]
zonaAsignacionPrevisional	integer	Si	Puntero a objeto [ubicacionGeografica]
correspondeAsignacionMaternal	boolean	Si	Si es que tiene asignación maternal
isapre	integer	Si	Puntero a objeto [prevision]
montoPactadoIsapre	float	Si	Monto Pactado con Isapre. Si está vacío, se asume 7%
montoPactadoIsapreMoneda	string	Si	Moneda monto pactado. UF=UF | $=Pesos \| 7+GES=7%+Ges en UF \| 7+GES$=7% + GES en pesos
afp	integer	Si	Puntero a objeto [afp]
adscribeASeguroCesantiaParaContratosPreviosA2002	boolean	No	Si el trabajador está contratado con fecha anterior al 2002, si adscribe o no al seguro de cesantía

apvMonto	float	Si	Monto primer APV
apvMoneda	string	Si	Moneda APV: UF | $
apvInstitucion	integer	Si	Puntero a [institucionAPV]
apvTipo	char(1)	Si	Tipo de APV: “A” | “B”
apvCuentaDos	float	Si	Monto cuenta dos
apvCuentaDosMoneda	string	Si	Moneda cuenta dos
depositoConvenidoMonto	float	Si	Monto Depósito Convenido
depositoConvenidoMoneda	string	Si	Moneda Depósito Convenido: UF | $
retencionJuducialDestinatario	string	Si	Nombre de persona destinatario de retención judicial
sueldoPatronal	boolean	Si	¿Es sueldo patronal? Sólo para socios de la empresa
sueldoBase	integer	Si	Sueldo Base mensual en $
sueldoFormaPago	string	Si	Forma de pago de sueldo.
sueldoBanco	id	Si	Puntero a [banco]
sueldoCuentaCorriente	string	Si	Número de cuenta corriente para depósito
sueldoCuentaCorrienteTipo	string	Si	Tipo de cuenta corriente: Cuenta Vista | Cuenta de Ahorro | Cuenta Corriente
sueldoTipoPago	string	Si	Forma de cálculo sueldo: mensual | diario |hora
valorHoraExtraPactada	float	Si	
mesesImponiblesreconocidos	integer	Si	"Meses que se reconocen como trabajados de antes de contratar a la persona. Se usan para los días progresivos
mesesImponiblesReconocidosDesde	date	Si	
vacacionesReconocidoDesde	date	Si	Beneficio. Fecha de contratación utilizada para cálculo de vacaciones progresivas
asignacionMovilizacion	integer	Si	Movilización mensual en $
asignacionColacion	integer	Si	Colación mensual en $
anticipoPactado	integer	Si	
fechaDeContratacionReconocidaParaAnosDeServicio	date	Si	Beneficio. Fecha de Contratación para utilizar para cálculo de años de servicio
pagaTresPrimerosDiasLicencia	boolean	Si	Beneficio. ¿Se subsidian los 3 primeros días de licencia?
mantieneRentaLiquidaLicencia	boolean	Si	Beneficio. ¿Mantiene la renta líquida durante las licencias?
diasAdministrativos	integer	Si	Beneficio. Días administrativos por trabajador durante el periodo
beneficiosInfoAdicional	string	Si	String con beneficios adicionales
indemnizacionSinTopeAnos	boolean	Si	Beneficio. Al calcular la indemnización por años de servicio al finiquitar, ¿se quita el tope de años?
indemnizacionSinTopeRenta	boolean	Si	Beneficio. Al calcular la indemnización por años de servicio al finiquitar, ¿se quita el tope de renta?
diasAdicionalesVacaciones	integer	Si	Beneficio. Días adicionales de vacaciones por año.
creadoPor	integer	Si	Creador del contrato. Puntero a [persona]
descripcionDelCargo	string	Si	Descripción en texto del cargo
clausulasAdicionales	string	Si	
detalleAnexoContrato	string	Si	
documentoEsContratoOAnexo	string	No	
claseSalarial	[claseSalarial]	Si	
rolPrivado	boolean	Si	
asignacionZonaExtrema	float	Si	Factor de asignación adicional por zona extrema
unidadOrganizacionalDetails	[unidadOrganizacional]	Si	Objeto detallado [unidadOrganizacional]
tipoContratoDetails	[tipoContrato]	Si	Objeto detallado [tipoContrato]
externalReference	list[externalReference]	Si	Objeto detallado con los Ids de este contrato en otros sistemas
finiquitado	boolean	Si	¿Fue finiquitado?
motivoEgreso	[motivoEgreso]	Si	Motivo de término de contrato (si fue finiquitado)
INE	[codigoIne]	Si	Objeto detallado con el código INE del trabajador
userDefinedFields	list[userDefinedField]	Si	Listado con campos personalizados

Fuente: elaboración propia en base a Talana

Anexo 6: *Endpoint* de "Contratos resumido"

Propiedad	Tipo GET	Uso	Opcional	Descripción
id	integer	GET	No	Id único
empleado	integer	GET	No	Puntero a objeto [persona]
codigo	string	GET	Si	Código del contrato
fechaCreacion	datetime	GET	Si	Fecha de Creación original de estas condiciones contractuales

tipoContrato	integer	GET	Si	Puntero a objeto [tipoContrato]. Se especifican detalles en tipoContratoDetails
empleadorRazonSocial	integer	GET	No	Puntero a objeto [razonSocial]
cargo	string	GET	Si	
fechaContratacion	date	GET	Si	Fecha de contratación original del trabajador
hasta	date	GET	Si	Hasta cuándo rigen estas condiciones. Puede estar vacío
finiquitado	boolean	GET	Si	¿Fue finiquitado?
rolPrivado	boolean	GET	Si	¿Manejar como Rol Privado?
unidadOrganizacional		GET	Si	Puntero a objeto [unidadOrganizacional]
sucursal	[sucursal]	GET	Si	Objeto sucursal
jornada	integer	GET	Si	Puntero a objeto [jornada]
horasDeLaJornada	integer	GET	Si	Cantidad de horas semanales de trabajo
centroCosto	[centroCosto]	GET	Si	Objeto centroCosto
personaDetails	list[detalles]	GET	Si	Detalles personales
userDefinedFields	list[userDefinedField]	GET	Si	Listado con campos personalizados
activo	boolean	GET		(calculado) Si el contrato está activo al momento de consumir el servicio

Fuente: elaboración propia en base a Talana

Anexo 7: *Endpoint* de "Cargos"

Propiedad	Tipo GET	Uso	Opcional	Descripción
id	integer	GET	No	Id único
parent	integer	GET	Si	ID del nodo padre (cuando se construye como árbol)
code	string	GET	Si	Código del cargo
name	string	GET	No	Nombre del Cargo
level	integer	GET	Si	Nivel de indentación en el árbol
node_type	string	GET	No	Tipo de nodo en el árbol. Puede ser “familia” o “cargo”
visible	Boolean	GET	No	Si es o no visible para el usuario
created_on	datetime	GET	Si	Fecha y hora de creación
modified_on	datetime	GET	Si	Fecha y hora de última modificación

Fuente: elaboración propia en base a Talana

Anexo 8: *Endpoint* de "Vacaciones"

Propiedad	Tipo	Descripción
id	integer	ID único
empleado	integer	Puntero a [persona]
vacacionesDesde	date	Fecha inicio de vacaciones
numeroDias	float	Cantidad de días solicitados
jornada	char(1)	M=Mañana | T=Tarde (para medios días)
mediosDias	boolean	¿Sólo medio día?
vacacionesHasta	date	Último día de vacaciones solicitado
vacacionesRetorno	date	Fecha de reincorporación
aprobada	char(1)	¿Aprobada?: A=Aprobada | P=Pendiente | R=Rechazada
aprobadaPor	integer	Puntero a [persona]
creadaPor	integer	Puntero a [persona]
fechaAprobacion	datetime	Fecha de Aprobación
detallesTrabajador	[persona]	Datos detallados del trabajador
externalReference	list[externalReference]	Objeto detallado con los Ids de este contrato en otros sistemas
fechaCreacion	datetime	Fecha de Creación

Fuente: elaboración propia en base a Talana

Anexo 9: *Endpoint* de "Vacaciones resumido"

Propiedad	Tipo	Descripción
id	integer	ID único
empleado	integer	Puntero a [persona]
vacacionesDesde	date	Fecha inicio de vacaciones
numeroDias	float	Cantidad de días solicitados
mediosDias	boolean	¿Sólo medio día?
vacacionesHasta	date	Último día de vacaciones solicitado
vacacionesRetorno	date	Fecha de reincorporación
fechaAprobacion	datetime	Fecha de Aprobación
fechaCreacion	datetime	Fecha de Creación

Fuente: elaboración propia en base a Talana

Anexo 10: *Endpoint* de "Ausentismo"

Propiedad	Tipo GET	Tipo POST, PUT	Uso	Opcional	Descripción
id	integer		GET	Si	ID único
empleado	integer	integer	GET, POST, PUT	No	Creación, id o RUT del trabajador.

fechaDesde	date	date	GET, POST, PUT	No	Fecha inicio de ausencia
numeroDias	float	float	GET, POST, PUT	No	Cantidad de días solicitados
jornada	char(1)		GET	Si	M=Mañana | T=Tarde (para medios días)
mediosDias	boolean	boolean	GET, POST, PUT	Si	¿Sólo medio día?
fechaHasta	date		GET	Si	Último día de ausencia
fechaRetorno	date		GET	Si	Fecha de reincorporación
aprobada	char(1)		GET	Si	¿Aprobada?: A=Aprobada | P=Pendiente | R=Rechazada
aprobadaPor	integer	integer	GET	Si	Puntero a [persona]
creadoPor	integer		GET	Si	Puntero a [persona]
fechaAprobacion	datetime		GET	Si	Fecha de Aprobación
detallesTrabajador	[persona]		GET	Si	Datos detallados del trabajador
externalReference	list[externalReference]		GET	Si	Objeto detallado de este contrato en otros sistemas
fechaCreacion	datetime		GET	Si	Fecha de Creación
numeroLicencia	string	string	GET, POST, PUT	Si	Número de Licencia médica (opcional)
medicoLicencia	string	string	GET, POST, PUT	Si	Nombre del médico que emitió la licencia (opcional)
tipoAusencia	string	string	GET, POST, PUT	No	Tipo de ausencia. Puntero a [tipoAusencia]
documentacion	string	string	GET	Si	Documentación adicional presentada por el trabajador (opcional)
esContinuacion	boolean	boolean	GET, POST, PUT	Si	¿es continuación de una licencia anterior?

Fuente: elaboración propia en base a Talana

Anexo 11: *Endpoint* de "Ausentismo resumido"

Propiedad	Tipo GET	Uso	Opcional	Descripción
id	integer	GET	Si	ID único
empleado	integer	GET	No	Puntero a [persona]
fechaDesde	date	GET	No	Fecha inicio de ausencia
numeroDias	float	GET	No	Cantidad de días solicitados
fechaHasta	date	GET	Si	Último día de ausencia
fechaRetorno	date	GET	Si	Fecha de reincorporación
fechaAprobacion	datetime	GET	Si	Fecha de Aprobación
fechaHoraAprobacion	datetime	GET	Si	Fecha y hora de Aprobación
fechaCreacion	datetime	GET	Si	Fecha de Creación
tipoAusencia	string	GET	No	Tipo de ausencia. Puntero a [tipoAusencia]
motivo	string	GET	Si	Motivo de ausencia

Fuente: elaboración propia en base a Talana

Anexo 12: *Endpoint* de "Prorrateo por centro de costo"

Propiedad	Tipo GET	Uso	Descripción
id	integer	GET	Id único
empleado	integer	GET	Puntero a objeto [persona]
empleado__details	Object	GET	Detalles desreferenciados del objeto de empleado
idContrato	string	GET	Identificador de la contratación
fechaCreacion	datetime	GET	Fecha de Creación original de estas condiciones contractuales
empleadorRazonSocial	integer	GET	Puntero a objeto [razonSocial]
cargo	string	GET	
fechaContratacion	date	GET	Fecha de contratación original del trabajador
hasta	date	GET	Hasta cuándo rigen estas condiciones. Puede estar vacío
desde	date	GET	Desde cuándo rigen estas condiciones.
distribution	list	GET	Lista de centros de costo, con el porcentaje asignado durante el periodo.

Fuente: elaboración propia en base a Talana

Anexo 13: *Endpoint* de "Centralización contable"

Propiedad	Tipo	Descripción
razonSocial__nombre	String	Razón social del empleador
razonSocial	String	Rut de la razón social del empleador
nombreTrabajador	String	Nombre del Trabajador (para ítems desglosados por trabajador)
rutTrabajador	String	Rut del Trabajador (para ítems desglosados por trabajador)
sucursal__codigo	String	Sucursal asignada al trabajador (código)
sucursal__nombre	String	Sucursal asignada al trabajador (nombre)
item	String	El nómbre del ítem (ej: sueldo base)
item__codigo	String	Código asignado al ítem
item__nombreParametro	String	Cödigo de ítem asignado por Talana
year	Integer	Año del periodo centralizado
month	Integer	Mes del periodo centralizado
centroCosto__nombre	String	Nombre del centro de costo asignado a esa línea
centroCosto__codigo	String	Código del centro de costo asignado a esa línea

debe	Integer	Valor del “debe”
haber	Integer	Valor del “haber”
cuentaContable__codigo	String	Código cuenta Contable asignada
cuentaContable__nombre	String	Nombre cuenta Contable asignada

Fuente: elaboración propia en base a Talana

Anexo 14: *Endpoint* de "Asignación de ítems de pago"

Propiedad	POST	Uso	Opcional	Descripción
tipo_item	string	POST	No	Nombre del ítem de pago
valorFijo	float	POST	No	Valor para insertar o actualizar
glosa	string	POST	No	Texto descriptivo que se muestra la liquidación de sueldo del trabajador
id_contrato	integer	POST	Si	Id del contrato
rut_empleado	string	POST	Si	Rut del empleado
prorrateo	json	POST	Si	Estructura para contabilización (dependerá de la implementación)

Fuente: elaboración propia en base a Talana

Anexo 15: *Endpoint* de "Descarga de documentos de personas"

Propiedad	Tipo	Uso	Descripción
empleado	integer	GET	Id persona
empleado_detalles	[persona]	GET	Datos detallados del trabajador
adjunto	string	GET	URL del documento
nombre	string	GET	Id del contrato
categoria	string	GET	Nombre de la categoría
puedeVerloElTrabajador	boolean	GET	Visibilidad del documento para el dueño de éste
fechaCreacion	datetime	GET	Fecha y hora de creación

Fuente: elaboración propia en base a Talana

Anexo 16: *Endpoint* de "Creación de documentos de personas"

Propiedad	POST	Uso	Opcional	Descripción
empleado	integer	POST	Si (si existe rut)	Id persona
rut	string	POST	Si (si existe empleado)	Rut de la persona
adjunto	multipart	POST	No	El documento
nombre	string	POST	No	Id del contrato
categoria	string	POST	Si	Nombre de la categoría
puedeVerloElTrabajador	boolean	POST	Si	Visibilidad del documento para el dueño de éste

Fuente: elaboración propia en base a Talana

Anexo 17: *Endpoint* de "Días administrativos"

Propiedad	Tipo	Descripción
id	integer	ID único
empleado	integer	Puntero a [persona]
desde	date	Fecha inicio de solicitud
numeroDias	float	Cantidad de días solicitados
jornada	char(1)	M=Mañana | T=Tarde (para medios días)
mediosDias	boolean	¿Sólo medio día?
hasta	date	Último día solicitado
retorno	date	Fecha de reincorporación
aprobada	char(1)	¿Aprobada?: A=Aprobada | P=Pendiente | R=Rechazada
aprobadaPor	integer	Puntero a [persona]
creadaPor	integer	Puntero a [persona]
detallesTrabajador	[persona]	Datos detallados del trabajador
fechaCreacion	datetime	Fecha de Creación

Fuente: elaboración propia en base a Talana

Anexo 18: *Endpoint* de "Días administrativos resumido"

Propiedad	Tipo	Descripción
id	integer	ID único
empleado	integer	Puntero a [persona]
desde	date	Fecha inicio de solicitud
numeroDias	float	Cantidad de días solicitados
mediosDias	boolean	¿Sólo medio día?
hasta	date	Último día solicitado
retorno	date	Fecha de reincorporación
fechaCreacion	datetime	Fecha de Creación
fechaAprobacion	datetime	Fecha de Aprobación

Fuente: elaboración propia en base a Talana

Anexo 19: *Endpoint* de "Enrolamiento de firma digital"

Propiedad	Tipo	Descripción
id	integer	ID único
empleado	integer	Puntero a [persona]
rut	String	Rut del Trabajador
vigente	Boolean	Flag que muestra si el enrolamiento está vigente
tipo	String	Tipo de enrolamiento
fechaCreacion	Date	Fecha de creación del enrolamiento
mobileNumber	String	Número de teléfono del enrolamiento
creadoPor	integer	Puntero a [persona]

Fuente: elaboración propia en base a Talana

Anexo 20: *Endpoint* de "Solicitud de firma digital"

Propiedad	Tipo	Descripción
id	integer	ID único
empleado	integer	Puntero a [persona]
documentUrl	String	URL de verificación del documento
documentType	String	Tipo de documento
documentReference	integer	ID único del documento
signed	boolean	Flag que muestra si el documento está firmado
requestTS	datetime	Fecha de creación de la solicitud
uuid	String	Código único de la solicitud

Fuente: elaboración propia en base a Talana

Anexo 21: *Endpoint* de "Solicitud de firma digital" asociado al registro

Propiedad	Tipo	Descripción
id	integer	ID único
requestedUser	integer	Puntero a [persona]
signed	boolean	Flag que muestra si esta solicitud de firma por empleado fue firmada
status	String	Estado de la solicitud de firma
token	String	ID único del documento
signatureTS	datetime	Timestamp de la firma
TSASignature	datetime	Timestamp de la TAS de la firma
ip	String	IP desde donde se firmo
userAgent	String	Agente desde el cual se firmo
passVerification	String	Hash de verificación de la firma

Fuente: elaboración propia en base a Talana

Anexo 22: *Endpoint* de "Asignación de personas a turnos"

Propiedad	Tipo	Descripción
id	integer	ID único
fromDate	string	Fecha de inicio de asignación, en formato "YYYY-mm-dd"
toDate	string	Fecha de fin de asignación, en formato "YYYY-mm-dd"
person	integer	Persona asignada. Puntero a [persona]
workShift	integer	El turno asignado. Puntero a [workShift]

Fuente: elaboración propia en base a Talana

Anexo 23: *Endpoint* de "Inyección y visualización de marcas"

Propiedad	Tipo GET	Tipo POST	Uso	Opcional	Descripción
id	integer				ID único
person	[persona]		GET	-	Detalles de la persona que marcó
rut		string	POST	Si	Rut, sin puntos, con guión y DV
card_id		string	POST	Si	El código de tarjeta asignado
office	integer	integer	GET, POST	Si	ID en Talana de la sucursal
direction	char(1)	char(1)	GET, POST	Si	Dirección de la marca. (Tipo Marca) "E" = Enter "X" = Exit
TS	timestamp	timestamp	GET, POST	No	Fceha y Hora real de marca. Formato "2018-07-13T17:22:21"
sourceMark	string	string	GET, POST	Si	Fuente de la marca (usar como Tipo de Actualización de Registro)
received_datetime	timestamp	timestamp	GET, POST	No	Fecha y hora de recepción de marca Formato "2018-07-13T17:22:21"

Fuente: elaboración propia en base a Talana

Anexo 24: *Endpoint* de "Días trabajados por contrato"

Propiedad	Tipo	Descripción
id	integer	ID del contrato
idContrato	string	ID único de la contratación

empleado	[persona]	Empleado
codigo	string	Código del contrato
empleadorRazonSocial	[razonSocial]	Razón social bajo el cual está contratado el trabajador
diasTrabajados	float	Número de días trabajados, en base a los días del mes
diasTrabajadosReales	float	Número de días calculados como días del mes – días con licencia – días de vacaciones – días de ausencia, en base a 30

Fuente: elaboración propia en base a Talana

Anexo 25: *Endpoint* de "Turnos y horarios asignados por trabajador"

Propiedad	Tipo	Descripción
person	integer	Id de persona
personName	string	Nombre y apellidos de la persona
rut	string	Rut de la persona
days	[asignacion]	Lista de objetos de tipo "asignacion"

Fuente: elaboración propia en base a Talana

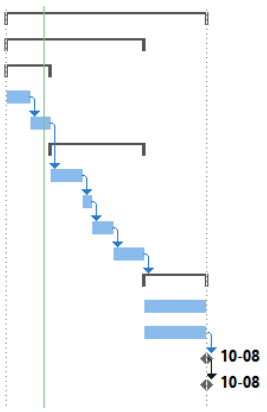
Anexo 26: Objeto "Asignación" de "Turnos y horarios asignados por trabajador"

Propiedad	Tipo	Descripción
workshift	string	Nombre del turno asignado
card	string	Código de la tarjeta asignado para ese día
entranceTime	timestamp	Fecha y hora de entrada del trabajador
exitTime	timestamp	Fecha y hora de salida del trabajador
hasToMark	boolean	¿debe marcar ese día?
reason	string	La razón por la cual no debe marcar ese día

Fuente: elaboración propia en base a Talana

Anexo 27: Actividades

Proyecto de titulo	**97 días**	**lun 29-03-21**	**mar 10-08-21**	
Implementación de la propuesta metodológica	**67 días**	**lun 29-03-21**	**mar 29-06-21**	
Diagnósticar	**22 días**	**lun 29-03-21**	**mar 27-04-21**	
Establecer objetivos y oportunidad	12 días	lun 29-03-21	mar 13-04-21	
Recopilar y entender datos	10 días	mié 14-04-21	mar 27-04-21	4
Desarrollar	**45 días**	**mié 28-04-21**	**mar 29-06-21**	
Preparar datos	15 días	mié 28-04-21	mar 18-05-21	5
Aplicar técnicas	5 días	mié 19-05-21	mar 25-05-21	7
Medir / Evaluar	10 días	mié 26-05-21	mar 08-06-21	8
Implementar	15 días	mié 09-06-21	mar 29-06-21	9
Ajustes finales al proyecto	**30 días**	**mié 30-06-21**	**mar 10-08-21**	**10**
Corregir etapas anteriores	30 días	mié 30-06-21	mar 10-08-21	
Redactar informe final	30 días	mié 30-06-21	mar 10-08-21	
Ajustes terminados	0 días	mar 10-08-21	mar 10-08-21	13
Proyecto terminado	0 días	mar 10-08-21	mar 10-08-21	14



Fuente: elaboración propia