

# Consultoría de Ciencia de Datos para Procesos de Reclamos de Licencias Médicas en la SUSESO

José Munizaga-Rosas

Abril 2018

## Contenidos

1	Introducción .....	7
2	Predicción y Clasificación Proactiva de Reclamantes .....	10
2.1	Introducción .....	10
2.1.1	Los Datos .....	11
2.2	Uso de Arboles Clasificadores .....	13
2.2.1	Primer Ejemplo para Año 2016 .....	14
2.2.2	Segundo Ejemplo para Año 2016 .....	16
2.2.3	Nuevo Ejemplo para Año 2016 considerando más variables .....	18
2.2.4	Nuevo Ejemplo Usando Criterio de Reclamo Basado en Folio: Todas las Licencias .....	19
2.2.5	Nuevo Ejemplo Usando Criterio de Reclamo Basado en Folio: Todas las Licencias, Folio sin Hash .....	21
2.2.6	Ejemplo Adicional Usando Criterio de Reclamo Basado en Folio: Licencias Rechazadas .....	24
2.3	Uso de Redes Neuronales .....	26
2.3.1	Ejemplo Usando Criterio de Reclamo Basado en Folio .....	26
2.4	Uso de Modelos de Regresión .....	27
2.4.1	Ejemplo Usando Criterio de Reclamo Basado en Folio .....	28
2.5	Pruebas adicionales utilizando rechazos y reclamos definidos por SUSESO .....	30
2.5.1	Modelo de árbol de decisión utilizando identificaciones provistas por SUSESO .....	30
2.5.2	Uso de Modelo de <i>Ensemble</i> .....	32
2.6	Conclusiones .....	36
3	Asignación Inteligente de Casos .....	39
3.1	Introducción .....	39
3.2	Descripción de la Heurística Existente para el Problema .....	40
3.3	Críticas a la Heurística Existente .....	42
3.3.1	Medición de la Complejidad de un Reclamo .....	42
3.3.2	Determinación de la Especialidad Requerida por un Reclamo .....	42
3.3.3	Determinación de la Eficacia de un Profesional .....	43
3.3.4	La Heurística es Golosa .....	43
3.3.5	Algunos comentarios adicionales .....	48
3.4	Características Deseables de un Modelo de Asignación .....	49
3.4.1	Función Objetivo Razonable .....	49
3.4.2	Debiese considerar el costo de operar el sistema .....	49
3.4.3	Debiese ayudar a determinar el dimensionamiento requerido para entregar el servicio .....	49

3.4.4	Debiese incorporar demanda futura esperada en la decisión .....	49
3.4.5	Debiese incorporar el concepto de nivel de servicio .....	50
3.5	Conclusiones .....	50
4	Propuesta Automatizada de Resolución de Casos.....	52
4.1	Introducción .....	52
4.2	Breve Descripción de los Datos a Utilizar .....	52
4.3	La Metodología Propuesta.....	55
4.4	Aplicación de la Metodología a Licencias de Origen Común .....	56
4.4.1	Utilizando solo las Familias CIE10 correspondientes a F, M y J .....	56
4.4.2	Utilizando solo las Familias CIE10 correspondientes al complemento de la subsección	
4.1	58	
4.4.3	Utilizando Codificación de Variables Dummies .....	59
4.5	Aplicación de la Metodología a Licencias de Origen Maternal.....	62
4.6	Observaciones Adicionales .....	62
4.7	Conclusiones .....	62
4.8	Apéndice .....	63
4.8.1	Familias de Diagnóstico CIE10.....	63
5	Predicción y Clasificación de Comportamiento Post-Atención.....	65
5.1	Introducción .....	65
5.2	Predictor de Reconsideraciones .....	66
5.3	Ayuda a la Fiscalización .....	67
5.3.1	Primer Ejemplo: Comportamiento global Médicos de la SUSESO en el período 2014 –	
2017	68	
5.3.2	Segundo Ejemplo: Comportamiento Anual de Médicos de la SUSESO en el período	
2014 – 2017.....	69	
5.3.3	Comentarios Adicionales .....	70
5.4	Conclusiones .....	74
5.5	Apéndice .....	74
5.5.1	Agrupamiento de casos por médico .....	74
6	Conclusiones .....	77
6.1	Sugerencias para mejoras futuras .....	78
6.1.1	Mejora de las relaciones entre tablas de distintas bases de datos .....	78
6.1.2	Enlace a bases de datos proveniente de otros servicios .....	79
6.1.3	Creación de una Línea de Tiempo de cada caso .....	80
6.1.4	Mejoras en la clasificación de las licencias de acuerdo con el Código de Diagnóstico.	81
6.1.5	Respecto del concepto de Expediente.....	81

6.1.6	Creación de un modelo económico para los reclamos .....	81
6.1.7	Estrategias tendientes a eliminar posibles sesgos en los Dictámenes .....	82
6.1.8	Caracterización de la fuerza de trabajo .....	82
6.1.9	Codificación de variables categóricas como variables dummies .....	83
6.1.10	Codificación de valores faltantes como algún valor por defecto.....	83

## Figuras

Figura 1: Resumen del trabajo realizado .....	9
Figura 2: Árbol Clasificador, Segundo Ejemplo 2016 .....	17
Figura 3: Clasificador, Ejemplo 2016 - Mas Variables .....	19
Figura 4: Clasificador, Ejemplo 2016 - Otra Definición de Reclamación.....	21
Figura 5: Clasificador, Ejemplo 2016 - Otra Definición de Reclamación.....	23
Figura 6: Clasificador, Ejemplo 2016 - Otra Definición de Licencias Rechazadas .....	25
Figura 7: Diagrama de Flujo del Modelo Actualmente Propuesto .....	41
Figura 8: Frecuencias para cada familia CIE10 en el caso de licencias de origen común .....	54
Figura 9: Frecuencias para cada familia CIE10 en licencias de origen maternal .....	55
Figura 10: Evolución de Porcentajes de Aprobación/Rechazo .....	69
Figura 11: Evolución del Número de Reclamos .....	70

## Tablas

Tabla 1: Ejemplo ficticio, factores en uso .....	43
Tabla 2: Ejemplo de heurística para un caso hipotético .....	44
Tabla 3: Nuevo ejemplo, secuencia aleatoria distinta, mismos tiempos de llegada .....	46
Tabla 4: Optimización para ejemplo 1 .....	47
Tabla 5: Modelo de optimización para el segundo ejemplo.....	48
Tabla 6: Cálculo de probabilidades familias CIE 10 poco relevantes .....	59
Tabla 7: Propuesta conceptual de estimación de esfuerzo basado en naturaleza de casos cercanos	62
Tabla 8: Ejemplo de productividad diaria estimada por año para un médico.....	70
Tabla 9: Muestra de reporte de médicos que emiten licencias.....	73
Tabla 10: Detalle de los resultados de reclamo por médico.....	75
Tabla 11: Resumen de Conclusiones y Recomendaciones .....	79

## 1 Introducción

El presente documento entrega a través de sus páginas la documentación de los experimentos llevados a cabo, dentro del contexto de una consultoría de Ciencia de Datos (Business Intelligence) encargada por la SUSESO, para explorar y pilotear la aplicabilidad de algoritmos de aprendizaje automático en el contexto del flujo de reclamos de licencias hechos a la SUSESO.

En reunión de lanzamiento de la Consultoría (Kick-Off) realizada con fecha 07 de diciembre de 2017, se determinó la forma de cumplir con los objetivos del servicio, en los siguientes términos: existen cuatro dimensiones en la Consultoría (precisadas en el numeral 3 de los Términos de Referencia, que son: Predicción y clasificación proactiva de reclamantes; Asignación Inteligente de Casos, Propuesta automatizada de Resolución de Casos, y Predicción y clasificación de comportamiento post-atención), en todas las cuales corresponde realizar las labores consignadas en el numeral 4.1 sobre aspectos metodológicos (1. Levantamiento y análisis exploratorio de información existente; 2. Revisión bibliográfica de modelos similares en caso de que estos existan; 3. Consideración de modelos matemáticos y alternativas; 4. Pilotaje a pequeña escala de los modelos matemáticos para efectos de validar la hipótesis de efectividad de los modelos bajo consideración; 5. Determinación de camino a seguir basado en antecedentes recopilados en los pasos anteriores). Por su parte, toda esta labor a su vez se distribuye según materia, es decir, licencias médicas, seguro laboral y otros; en función de la información disponible en los sistemas de información de la Superintendencia.

El proceso que sigue una licencia una vez que esta es rechazada dependerá básicamente de si ella le concierne a una entidad aseguradora privada o al sistema público. Las diferencias que existen tienen que ver con el flujo intermedio de dichos rechazos, información que actualmente no se posee, y que eventualmente termina en un reclamo siendo ingresado a la SUSESO por los usuarios del sistema. Una vez que dicho reclamo ha sido ingresado a la SUSESO, esta debe asignar un profesional para que revise el caso, este profesional debe resolver si acoger o no el reclamo y finalmente de crea un dictamen que es remitido al reclamante. En algunas circunstancias, los reclamantes pueden no estar de acuerdo con el resultado (esencialmente su reclamo no ha sido acogido) y en dicho caso se reingresa el caso a la SUSESO. Como es posible ver, estas etapas del proceso son esencialmente las definidas dentro del contexto de esta consultoría y tienen ellas por objetivo el poder disminuir los tiempos actuales de resolución de reclamos (que de acuerdo con estadísticas oficiales es del orden de 120 días) para eventualmente producir dictámenes dentro de un plazo de 30 días máximo en un 95% de los casos.

De acuerdo con información entregada por la contraparte técnica de la SUSESO, el mayor tiempo de espera en el flujo de trabajo guarda relación con la conformación del expediente, por lo tanto, poseer capacidades predictivas que permitan anticipar los reclamos, puede ayudar a mejorar sustancialmente los tiempos de operación del sistema. Las siguientes etapas del proceso agregan también tiempo, y por lo tanto la utilización de información histórica que permita entender y caracterizar de mejor manera el fenómeno es una gran ayuda.

La consultoría se ha llevado a cabo entre los meses de Diciembre 2017 y Abril 2018, la metodología de trabajo ha requerido interacción regular con la contraparte técnica de la SUSESO, y su aplicación sostenida ha permitido conseguir los objetivos propuestos al inicio de ésta. Para efectos de poder entender las contribuciones principales de esta consultoría, es que se detalla en la Figura 1 los mayores hallazgos y se provee a modo de orientación un desglose del trabajo realizado. Los cuatro capítulos que siguen a esta introducción poseen el detalle de los experimentos hechos y las conclusiones obtenidas a partir de los datos. Finalmente, el informe se cierra con un capítulo de conclusiones globales en que se reiteran algunas de las conclusiones parciales, y además se perfilan posibles líneas de trabajo futuro que puedan mejorar aún más lo obtenible hoy en día. Para efectos de simplificar esta presentación se entrega una tabla también que resume las conclusiones principales por etapa y además se discuten las posibles líneas de desarrollo a futuro.



	Predicción y clasificación proactiva de reclamantes	Asignación Inteligente de Casos	Propuesta automatizada de Resolución de Casos	Predicción y clasificación de comportamiento post-atención
Objetivo	Poder identificar reclamantes una vez que su licencia ha sido rechazada pero con anterioridad a la presentación del reclamo a la SUSESO	Poder revisar la heurística existente y ver que tan apropiada es dentro del contexto de los datos disponibles	Crear una metodología que permita sugerir al profesional que estudia un caso el dictamen más razonable basado en información histórica	Ver si es que es posible predecir los casos que una vez dictaminados podrían presentar re-ingresos. Ver si es posible ayudar a fiscalizar
Descripción Fuentes Principales	<ul style="list-style-type: none"> <li>• Base de Datos SILMSIL</li> <li>• Base de Datos SISESAT</li> <li>• Base de Datos Expedientes Workflow</li> </ul>	<ul style="list-style-type: none"> <li>• Descripción de la Heurística propuesta entregado por la SUSESO</li> </ul>	<ul style="list-style-type: none"> <li>• Base de Datos creada por la SUSESO que "pega" registros SILMSIL con registros Expedientes Workflow</li> </ul>	<ul style="list-style-type: none"> <li>• Base de Datos filtrada por la SUSESO para identificar registros SILMSIL con re-ingreso</li> </ul>
Procesos Corridos	<ul style="list-style-type: none"> <li>• Implementación de árboles clasificadores en distintos casos: desbalanceado, balanceado, etc.</li> <li>• Utilización de modelos de regresión</li> <li>• Implementación de modelos de redes neuronales</li> <li>• Utilización de modelos de bagging y random forest</li> </ul>	<ul style="list-style-type: none"> <li>• Simulación de la Heurística propuesta de manera manual para un caso artificial que sirve de ilustración</li> <li>• Ejecución de modelo de optimización entero no-lineal simple para el mismo caso artificial y análisis de la brecha</li> </ul>	<ul style="list-style-type: none"> <li>• Análisis Exploratorio de Datos sobre los códigos de diagnóstico</li> <li>• Creación de algoritmo para identificar vecinos de casos (definición de métrica)</li> <li>• Pruebas con la metodología propuesta</li> </ul>	<ul style="list-style-type: none"> <li>• Implementación de árbol clasificador</li> <li>• Confección de indicadores ilustrativos usando EXCEL</li> <li>• Tests estadísticos</li> </ul>
Metodología	<ul style="list-style-type: none"> <li>• Implementación de algoritmos utilizando el lenguaje de programación R</li> <li>• Algunos análisis fueron hechos en GRETL</li> </ul>	<ul style="list-style-type: none"> <li>• Simulación manual de la heurística para un ejemplo simple</li> <li>• Construcción de un modelo de optimización simple y resuelto utilizando SOLVER de EXCEL</li> </ul>	<ul style="list-style-type: none"> <li>• Implementación de algoritmos utilizando el lenguaje de programación R</li> </ul>	<ul style="list-style-type: none"> <li>• Implementación de algoritmos utilizando el lenguaje de programación R</li> <li>• Algunos análisis fueron hechos en EXCEL y GRETL</li> </ul>

Figura 1: Resumen del trabajo realizado

## 2 Predicción y Clasificación Proactiva de Reclamantes

### 2.1 Introducción

El presente capítulo tiene por objetivo describir en algún nivel de detalle los experimentos hechos con los datos provistos por la SUSESO para poder explorar un prototipo de algoritmo de aprendizaje supervisado para poder predecir cuáles de las licencias rechazadas presentarán reclamo a la SUSESO.

La idea es bastante simple en esencia, el conjunto de datos posee dos tipos distintos de licencias, aquéllas que son rechazadas y que terminan presentando reclamos y aquéllas que son rechazadas y que no presentan reclamos a la SUSESO. Con esta información, año a año, es posible entonces utilizar algoritmos de aprendizaje automático supervisados y estudiar la precisión del clasificador. No obstante, la simpleza de la idea es menester notar que los datos no son balanceados, por este motivo hay que ser cuidadosos con la definición de lo que constituye un buen clasificador. En las circunstancias actuales, aproximadamente un 10% de las licencias que se rechazan presentan reclamo a la SUSESO. Si lo que se persigue es un clasificador con un buen porcentaje de aciertos, entonces la forma trivial de construir uno con un 90% de precisión es un clasificador que siempre dice que la licencia rechazada no presentará reclamo.

La implementación de la idea reveló ciertas características del problema que no habían sido consideradas en un principio, como por ejemplo el efecto de los datos desbalanceados mencionados anteriormente, inconsistencias en ciertos campos de la base de datos y algunas otras dudas que han sido convenientemente comunicadas, discutidas y acordadas mediante reuniones sostenidas con la contraparte técnica de la Superintendencia.

Los experimentos realizados mostraron que los árboles clasificadores son una buena opción al momento de clasificar licencias rechazadas que podrían presentar reclamación a la SUSESO, son modelos que además no están fijos en el tiempo y pueden ser entrenados automáticamente y de manera periódica en la medida de que nuevos datos sean recolectados. Los modelos de regresión producen resultados ligeramente mejores, sin embargo, la selección de variables podría ser arbitraria puesto que probar exhaustivamente todos los posibles modelos (i.e. utilizando todas las posibles combinaciones de variables) es un ejercicio costoso, los coeficientes en el modelo de regresión si bien tienen una interpretación, ésta es difícil de seguir para la gran mayoría de las personas y dejaría un modelo que no podría ser actualizado frecuentemente. Finalmente se intentaron modelos de redes neuronales pero el costo computacional de entrenamiento de este tipo de modelos resulta ser prohibitivo lo que sugiere el no adoptar este tipo de modelos puesto que ni siquiera pueden dar respuesta en tiempos razonables, hecho que obstaculiza su adopción en el contexto del proyecto PMI.

Finalmente, para cerrar esta sección, se debe hacer notar que el proceso de descubrimiento que se ha seguido para trabajar con los datos ha sido documentado en el reporte. Por este motivo se pueden observar variaciones de los ejemplos que fueron estudiados, experimentos que siguen un orden cronológico y muestran el aumento de entendimiento tanto del conjunto de datos como de los modelos aplicados a ellos.

### 2.1.1 Los Datos

Los datos que principalmente han sido provistos son de tres tipos: Bases de datos de licencias médicas, bases de datos de accidentes laborales y el registro de los reclamos presentados a SUSESO.

Una de las principales falencias de los datos en su forma actual es la imposibilidad de poder identificar qué licencias rechazadas en primera instancia han sido revertidas ya sea por apelación a la entidad aseguradora o por intervención de la COMPIN. Esto introduce un nivel de suciedad en los datos puesto que lo que se tiene por un lado son las licencias médicas rechazadas en primera instancia y después lo que se tiene es el archivo `expedientes_workflow.csv` que contiene la información relativa a los reclamos que han llegado a la SUSESO. Reconstruir esta historia intermedia es importante puesto que podría agregar información que permita identificar casos que no terminarán en reclamación ante la SUSESO.

Se llevó a cabo un pequeño análisis exploratorio de los datos, las cosas más relevantes que se observaron son las siguientes:

- La variable `articulo_77` reporta lo siguiente:

```
articulo_77
1 :      41
2 :3830953
ND:1639625
```

No se ha suministrado un diccionario de datos para entender lo que la categoría 1 y 2 significan. Llama la atención el alto número de faltantes en el campo (los que aquí aparecen marcados como ND)

- La variable `sexo_trabajador` reporta lo siguiente:

```
sexo_trabajador
1 :2029772
2 :3440837
NA's:      10
```

Llama la atención la prevalencia de un sexo respecto al otro, lo segundo es la existencia de 10 registros sin sexo.

- La variable `numero_dias_licencia` reporta las siguientes estadísticas:

```
numero_dias_licencia
Min. : 0.00000
```

```
1st Qu.: 5.00000
Median : 11.00000
Mean   : 14.53921
3rd Qu.: 20.00000
Max.   : 600.00000
```

Con la mediana teniendo un valor de 11 se tiene que el 50% de las licencias es de menos de 11 días durante el 2016 (que es el año donde se concentró el análisis).

- La variable `licencia_mat_suplementaria` reporta lo siguiente:

```
licencia_mat_suplementaria
1 : 370064
2 : 3460930
ND: 1639625
```

Llama la atención de que las licencias maternas suplementarias categoría 2 sumen más que la cantidad de trabajadores cuyo sexo es tipo 2 (esto asumiendo que este sexo es femenino), considerando que no es posible para un hombre solicitar una licencia maternal suplementaria.

- La variable `tipo_licencia_lm` reporta lo siguiente:

```
tipo_licencia_lm
1      : 4821280
3      : 237431
4      : 200455
7      : 173991
5      : 30629
Other) : 6737
NA's   : 96
```

Llama la atención los otros y los NA.

- La variable `recuperabilidad_laboral` reporta lo siguiente:

```
recuperabilidad_laboral
1 : 5374829
2 : 59686
NA's: 36104
```

Llama la atención los NA.

- La variable `inicio_tramite_invalidez` reporta lo siguiente:

```
inicio_tramite_invalidez
1 : 555087
2 : 4843125
NA's: 72407
```

Llama la atención los NA.

- Similarmente para las siguientes variables llama la atención la cantidad de NA:
  - tipo\_reposo: 629
  - lugar\_reposo: 242516
  - tipo\_profesional: 51127
  - licencia\_modificada: 1640980
  - estado\_resolucion: 146927
  - tipo\_resolucion: 2663458
  - redictamen: 1639660
  - causa\_rechazo: 4572604
  - derecho\_a\_subsidio: 168173
  - Etc.
  
- Algunas variables no reportan valores, o bien sólo tienen los valores por defecto:
  - num\_dias\_cotizacion\_pagados: 0
  - num\_dias\_pagados\_periodo\_info: 0
  - monto\_subsidio\_pagado: -1
  - monto\_subsidio\_diario: -1
  - monto\_cotizaciones\_seg\_cesant: -1
  - monto\_otras\_cotizaciones: -1
  - monto\_liquido\_pagado: 0
  - remuneracion\_imp\_mes\_anterior: 0

Como se puede apreciar hay bastantes potenciales problemas en la base de datos, por ejemplo, uno de los datos importantes que podría haber servido como proxy del sueldo aparece sólo con ceros, por lo que resulta relevante que la SUSESO pueda obtener, al menos la información de las cotizaciones previsionales vinculadas a un RUT en una determinada fecha. También es importante notar que los primeros análisis hechos fueron ejecutados con la totalidad de la información de licencia, esto para efectos de poder tener una idea aproximada del tipo de comportamiento de los modelos, antes de construir un modelo muy detallado. Posteriormente se deberá implementar un filtro para identificar las licencias rechazadas sobre la base del campo `estado_resolucion` y el campo `derecho_a_subsidio`, los valores para dichos campos deben ser 2, 3 o 4 y B o C respectivamente.

## 2.2 Uso de Árboles Clasificadores

Un árbol clasificador es un tipo de algoritmo de aprendizaje supervisado que permite construir un árbol (estructura de datos) que posibilita clasificar una observación de acuerdo con valores de ciertos atributos. Por ejemplo, un árbol clasificador para determinar si un individuo tiene menos de 10 años basado en el atributo altura podría establecer que un individuo será clasificado como menor a 10 años si su altura es menor que 1.40 mts. de lo contrario este será considerado mayor a 10 años. ¿Es este método uno bueno? La respuesta es probablemente no, esto puesto que pueden existir diferencias entre niños y niñas, ubicación geográfica, etc. Por lo mismo, el método tendrá un error (i.e., no podrá clasificar correctamente a un 100% de los casos).

El ejemplo previo, a pesar de toda su evidente simpleza, muestra la naturaleza del problema en cuestión: Hay observaciones para los cuales se conoce la clasificación y el objetivo del algoritmo es poder “entrenar” un modelo de tal forma que se minimice alguna medida de error (la más obvia siendo el número de casos mal clasificados). Los métodos basados en árboles clasificadores destacan por su simpleza y por ser bastante decentes en un rango amplio de aplicaciones. Dentro de los modelos más frecuentemente usados de árboles clasificadores se encuentran C4.5 y CART. Habitualmente, estos modelos requieren que los atributos de las observaciones sean números reales, no obstante lo anterior, es posible utilizar variables categóricas en versiones adaptadas de los algoritmos originales sin ningún tipo de problemas, por este motivo se prefiere usualmente este tipo de algoritmos por sobre otros que tienen requerimientos más restrictivos en cuanto al formato y naturaleza de los datos. No es el objetivo de Informe precisar los detalles relativos a la formulación matemática de los modelos, sino más bien ver qué tipo de respuesta se puede obtener utilizando estas herramientas en un caso concreto.

Es importante notar que los árboles clasificadores operan por la vía de separar los datos utilizando planos (rectas en  $\mathbb{R}^2$ ) y que por lo tanto en función de la estructura del conjunto de datos, esta técnica podría no ser apropiada para los propósitos esperados. Los árboles clasificadores que se han modelado han utilizado la librería `tree` de R, el paquete estadístico de código abierto para análisis estadístico. Lo otro importante de notar es que los árboles clasificadores proveen una buena orientación respecto de las variables que podrían ser posteriormente incorporadas en un modelo de regresión multivariada. Finalmente, es importante precisar que los árboles clasificadores reciben como entrada una base de datos y proveen como salida un modelo de árbol, si los datos cambian el modelo también podría cambiar por lo que son particularmente útiles en el contexto de la creación de modelos de clasificación con una alta presencia de dinamismo en los datos. Otras técnicas tales como modelos de regresión requieren la especificación a priori de las variables a utilizar en el modelo lo cual podría eventualmente requerir constantes actualizaciones del modelo, los árboles por otro lado se entrenan de manera automática y como ya se ha mencionado el resultado viene en la forma de un árbol clasificador que además identifica las variables que permiten discriminar observaciones particulares presentadas al modelo.

### 2.2.1 Primer Ejemplo para Año 2016

El primer ejemplo que se corrió fue con toda la base de datos SILMSIL del año 2016. Se consideró que éste era un buen año a utilizar puesto que los datos son más recientes, ha habido un esfuerzo más claro de normalizar y depurar mejor los datos y la información al parecer es más completa. No se decidió utilizar el 2017 puesto que hay muchos casos limítrofes en que la licencia ha sido rechazada y sin embargo la reclamación aún no ha sido presentada a la SUSESO. Se estima que los casos del 2016 ya han tenido tiempo suficiente para ser apelados y presentados a todas las instancias intermedias de tal forma que, si hubo algún reclamo generado de un rechazo de alguna licencia del 2016, este debería estar presente en el archivo `expedientes_workflow.csv`.

Antes de poder correr el árbol clasificador, fue necesario indicarle a R cuáles variables son categóricas, esto puesto que si no hay una identificación explícita de las variables categóricas entonces el algoritmo presenta errores. El comando de R usado para indicar variables categóricas se puede ilustrar con el siguiente ejemplo:

```
SILMSIL$inicio_tramite_invalidez = as.factor(SILMSIL$inicio_tramite_invalidez)
```

Lo otro que se hizo fue crear un campo `Reclamo` que básicamente aproxima la licencia reclamada viendo si es que el RUT del reclamante aparece en el archivo `expedientes_workflow.csv`, razón por la cual, si el RUT aparece, entonces se codifica como `Reclama`, y, en caso contrario, se codifica como `No Reclama`. Esta es una primera aproximación a las licencias que presentan reclamos, sin embargo, a priori se identifican deficiencias con este método puesto que licencias rechazadas en primera instancia que fueron revertidas posteriormente, ya sea por la misma aseguradora o la COMPIN, aparecerán como reclamadas.

Cuando se corrió el árbol clasificador en el conjunto mencionado de datos, se obtuvo el siguiente output:

```
node), split, n, deviance, yval, (yprob)
  * denotes terminal node

1) root 3495061 2739266.0 No Reclama ( 0.86709245 0.13290755 )
 2) derecho_a_subsidio: A,B,ND 3082234 2104450.0 No Reclama ( 0.89243970
0.10756030 )
  4) numero_dias_licencia < 12.5 1852585 1019435.0 No Reclama ( 0.92148053
0.07851947 ) *
  5) numero_dias_licencia > 12.5 1229649 1045153.0 No Reclama ( 0.84868690
0.15131310 ) *
 3) derecho_a_subsidio: C 412827 518909.2 No Reclama ( 0.67784568
0.32215432 )
  6) numero_dias_licencia < 12.5 148484 130001.2 No Reclama ( 0.84114787
0.15885213 ) *
  7) numero_dias_licencia > 12.5 264343 358576.3 No Reclama ( 0.58611728
0.41388272 ) *
```

Y el resumen del árbol es:

```
Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "derecho_a_subsidio" "numero_dias_licencia"
Number of terminal nodes: 4
Residual mean deviance: 0.7305073 = 2553165 / 3495057
Misclassification error rate: 0.1329076 = 464520 / 3495061
```

lo que indica que la tasa de mala clasificación es de un 13% (que en este caso corresponden a los casos con licencia rechazada que presentan reclamación a la SUSESO). Claramente, este árbol si bien es matemáticamente correcto, para los propósitos de la SUSESO es inútil puesto que no permite identificar licencias que son rechazadas y que serán posteriormente reclamadas, el algoritmo opta por lo simple y

clasifica a todos como No Reclama y con ello produce un clasificador con una tasa del orden de 87% de asertividad lo que es bastante bueno.

Para solucionar los problemas vistos en el ejemplo, se deberá muestrear la base de datos para construir ejemplos balanceados en los que los casos de Reclama y No Reclama estén en las mismas proporciones. Esta técnica se sugiere en la literatura como una de las posibilidades para lidiar con bases de datos desbalanceadas.

## 2.2.2 Segundo Ejemplo para Año 2016

Variables utilizadas en el modelo:

- articulo\_77
- edad\_trabajador
- sexo\_trabajador
- numero\_dias\_licencia
- tipo\_licencia\_lm
- recuperabilidad\_laboral
- inicio\_tramite\_invalidez
- tipo\_reposo
- lugar\_reposo
- tipo\_profesional
- num\_dias\_previos\_autorizados
- causa\_rechazo
- derecho\_a\_subsidio
- actividad\_laboral\_trabajador
- ocupacion\_trabajador
- calidad\_trabajador
- num\_dias\_incapacidad
- monto\_subsidio\_liquido
- monto\_cotizaciones\_salud
- monto\_cotizaciones\_pension
- monto\_base\_calculo\_subsidio
- monto\_liquido\_pagado
- remuneracion\_imp\_mes\_anterior

La prueba tomó un total de 4.78 minutos en ejecutarse. El árbol clasificador es el siguiente:

```
node), split, n, deviance, yval, (yprob)
  * denotes terminal node
```

```
1) root 920857 1276000 Reclama ( 0.4952 0.5048 )
  2) derecho_a_subsidio: A,B,ND 745722 1025000 No Reclama ( 0.5549 0.4451 )
    4) número_dias_licencia < 12.5 402310 526600 No Reclama ( 0.6382 0.3618 ) *
      5) número_dias_licencia > 12.5 343412 473600 Reclama ( 0.4573 0.5427 ) *
    3) derecho_a_subsidio: C 175135 193400 Reclama ( 0.2409 0.7591 ) *
```

Y la precisión del árbol clasificador es la siguiente:



```

Classification tree:
tree(formula = Reclamo ~ articulo_77 + edad_trabajador + sexo_trabajador +
      numero_dias_licencia + tipo_licencia_lm + recuperabilidad_laboral +
      inicio_tramite_invalidez + tipo_reposo + lugar_reposo + tipo_profesional +
      num_dias_previos_autorizados + causa_rechazo + derecho_a_subsidio +
      actividad_laboral_trabajador + ocupacion_trabajador + calidad_trabajador +
      num_dias_incapacidad + monto_subsidio_liquido + monto_cotizaciones_salud +
      monto_cotizaciones_pension + monto_base_calculo_subsidio +
      monto_liquido_pagado + remuneracion_imp_mes_anterior, data = train.set)
Variables actually used in tree construction:
[1] "derecho_a_subsidio" "numero_dias_licencia"
Number of terminal nodes: 3
Residual mean deviance: 1.296 = 1194000 / 920900
Misclassification error rate: 0.3744 = 344799 / 920857

```

**La matriz de confusión es<sup>1</sup>:**

prediction	No Reclama	Reclama
No Reclama	122194	68200
Reclama	92887	146138

29% bien clasificado (no reclamantes) y 34% bien clasificado (reclamantes) lo que da un total aproximado de 63% de casos bien clasificados, que por lo demás se traduce en un 69% aproximadamente de casos de licencias reclamadas bien clasificadas ( $\approx 68200/214338$ ).

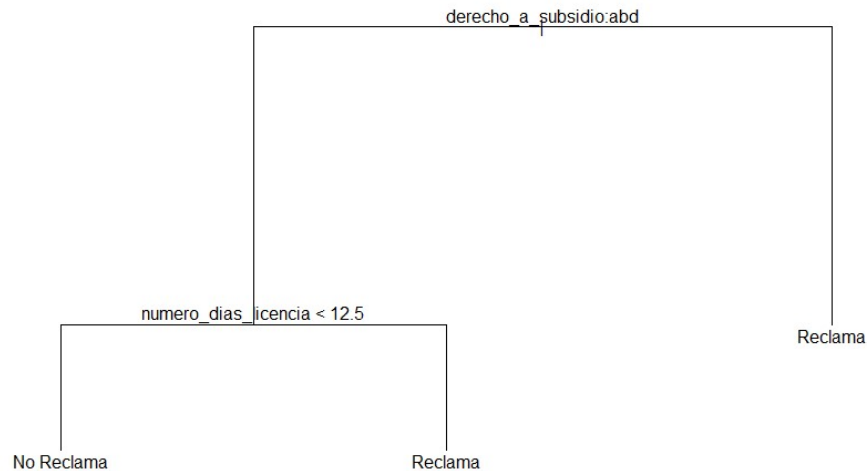


Figura 2: Árbol Clasificador, Segundo Ejemplo 2016

<sup>1</sup> Se denomina matriz de confusión a la tabla que resume la calidad del árbol clasificador, esto es, indica cuantos casos han sido correctamente clasificados (diagonal) y cuales han sido erróneamente clasificados (fuera de la diagonal). La matriz de confusión permite entender la precisión del modelo cuando éste es aplicado al conjunto de datos de prueba (datos que han sido separados y no han sido utilizados para entrenar el modelo)

### 2.2.3 Nuevo Ejemplo para Año 2016 considerando más variables

Un ejemplo similar al anterior se corrió, esta vez sin embargo se agregaron variables adicionales a las usadas anteriormente. Esta idea surge a raíz de interacciones con la contraparte técnica de la SUSESO en el contexto del proceso formal de discusión y clarificaciones relativas a la base de datos existente. Las variables adicionales que se agregaron a las ya usadas en el análisis son:

- licencia\_mat\_suplementaria
- licencia\_modificada
- tipo\_licencia\_medica\_resuelta
- estado\_resolucion
- tipo\_resolucion
- redictamen
- tipo\_entidad\_pagadora\_subsidio

La prueba tomó un total de 5.42 minutos en ejecutarse<sup>2</sup>. El árbol clasificador es el siguiente:

```
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

1) root 893899 1239008.0 Reclama ( 0.4925344 0.5074656 )
  2) num_dias_incapacidad < 0.5 160097 168547.2 Reclama ( 0.2195919
0.7804081 ) *
  3) num_dias_incapacidad > 0.5 733802 1009289.0 No Reclama ( 0.5520835
0.4479165 )
    6) numero_dias_licencia < 12.5 397791 521388.4 No Reclama ( 0.6365906
0.3634094 ) *
    7) numero_dias_licencia > 12.5 336011 462713.7 Reclama ( 0.4520388
0.5479612 ) *
```

Y la precisión del árbol clasificador es la siguiente:

```
Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "num_dias_incapacidad" "numero_dias_licencia"
Number of terminal nodes: 3
Residual mean deviance: 1.289467 = 1152649 / 893896
Misclassification error rate: 0.370967 = 331607 / 893899
```

lo que es una "ligera" mejora respecto del clasificador anterior (notar que la muestra es distinta eso sí).

La matriz de confusión es:

---

<sup>2</sup> Todos los tiempos reportados en el presente documento son relativos al uso de una máquina con un procesador Intel Core i7 7th Gen, con 16 GB de memoria RAM y disco duro SSD

prediction	No Reclama	Reclama
No Reclama	121716	68256
Reclama	93365	146082

lo cual da estadísticas similares a las ya obtenidas.

Lo notable de este experimento es que la adición de variables parece no cambiar mucho los resultados puesto que ellas no permiten discriminar adecuadamente. La otra cosa notable es que al haber cambiado la instancia de los datos (recordar que estamos muestreando para balancear los casos), el árbol ligeramente cambia.

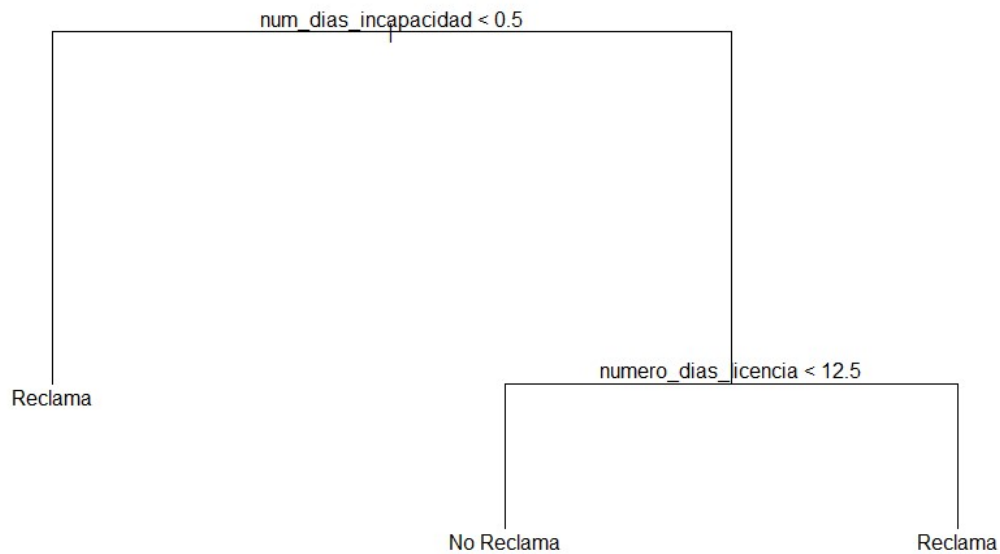


Figura 3: Clasificador, Ejemplo 2016 - Mas Variables

#### 2.2.4 Nuevo Ejemplo Usando Criterio de Reclamo Basado en Folio: Todas las Licencias

Como todos los modelos anteriores muestran, el porcentaje de datos bien clasificados no es tan significativo como se esperaría, pudiendo ser una de las posibles razones detrás de esto es la suciedad existente en los datos respecto de la identificación de cuáles licencias fueron reclamadas y cuáles no. Habiéndolo cotejado con la contraparte técnica de la Superintendencia, se acordó que el criterio de licencia rechazada con presentación de reclamación a SUSESO vendrá determinada por la presencia del folio de la licencia en alguno de los campos folio del archivo `expedientes_workflow.csv`. El criterio es entonces que la licencia es reclamada si su folio está relacionado a algún "expediente". Antes de proceder al ejemplo final donde además se reduce el universo total de licencias sólo a las rechazadas, primero probaremos la calidad de un clasificador sobre todas las licencias emitidas.

Se programó una rutina en R que busca la presencia de un folio en alguna de las 30 columnas y asigna entonces al folio de dicha licencia `Reclama`, en caso contrario asignará `No Reclama`. Notar que con este criterio se tiene que los reclamos ascienden 97229 del universo de 5470619 licencias médicas emitidas, es decir aproximadamente 1.78% de las licencias emitidas reclaman, lo que no es contradictorio con las estadísticas oficiales, ya que la cantidad de reclamos es del orden de un 10 a 15 % del universo de licencias rechazadas que bien se sabe es un 10 a 15% del universo de licencias emitidas.

La prueba tomó un total de 5.25 minutos en ejecutarse. El árbol clasificador es el siguiente:

```
node), split, n, deviance, yval, (yprob)
  * denotes terminal node

1) root 125148 173292.000 Reclama ( 0.48001566 0.51998434 )
 2) estado_resolucion: 1,3,4 65649 65574.650 No Reclama ( 0.80069765
0.19930235 )
 4) numero_dias_licencia < 9.5 28138 9656.704 No Reclama ( 0.95884569
0.04115431 ) *
 5) numero_dias_licencia > 9.5 37511 46911.400 No Reclama ( 0.68206659
0.31793341 )
 10) num_dias_previos_autorizados < 52.5 26168 27762.460 No Reclama (
0.77717059 0.22282941 ) *
 11) num_dias_previos_autorizados > 52.5 11343 15661.430 Reclama (
0.46266420
0.53733580 ) *
 3) estado_resolucion: 2,5 59499 45108.990 Reclama ( 0.12618699
0.87381301 )
 6) causa_rechazo: 3,ND 3403 4514.555 No Reclama ( 0.62151043
0.37848957 ) *
 7) causa_rechazo: 1,2,4,5 56096 35510.470 Reclama ( 0.09613876
0.90386124 )
 14) articulo_77: 2 43071 19845.990 Reclama ( 0.06124771
0.93875229 ) *
 15) articulo_77: ND 13025 13440.730 Reclama ( 0.21151631
0.78848369 ) *
```

Y la precisión del árbol clasificador es la siguiente:

```
Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "estado_resolucion" "numero_dias_licencia"
"num_dias_previos_autorizados"
[4] "causa_rechazo" "articulo_77"
Number of terminal nodes: 6
Residual mean deviance: 0.7262299 = 90881.86 / 125142
Misclassification error rate: 0.151165 = 18918 / 125148
```

lo que es una mejora sustancial respecto del clasificador anterior (notar que la muestra es distinta eso sí).

La matriz de confusión en el conjunto de datos de test es:

prediction	No Reclama	Reclama
No Reclama	23677	4161
Reclama	5509	24991

lo cual da estadísticas mejores a las ya obtenidas. La clasificación correcta de licencias que presentan reclamos es de un 85% aproximadamente<sup>3</sup>.

Lo notable de este experimento es que en función de cómo definamos las licencias que terminan reclamando, los resultados cambian notoriamente. Es preciso acordar o medir de mejor manera quiénes son los que reclaman, puesto que esto genera un impacto notable en el resultado.

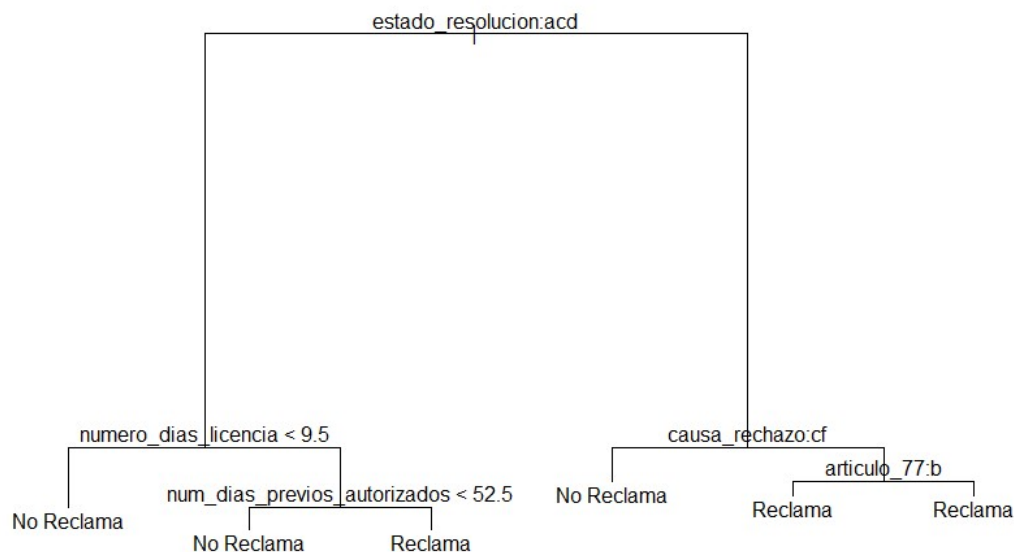


Figura 4: Clasificador, Ejemplo 2016 - Otra Definición de Reclamación

### 2.2.5 Nuevo Ejemplo Usando Criterio de Reclamo Basado en Folio: Todas las Licencias, Folio sin Hash

Los datos inicialmente fueron provistos con los números de folio escritos utilizando una tabla Hash. Para efectos de poder analizar si es que el uso de códigos Hash introduce errores, se volvió a correr el mismo algoritmo de la subsección anterior sólo

<sup>3</sup> Este porcentaje se calcula utilizando los valores tomados de la matriz de confusión, que indica que de un total de 29152 = 24991 + 4161 licencias que debiesen haber reclamado, hay 24991 que han sido correctamente clasificadas, esto entrega  $24991/29152 \approx 0.85$ , es decir 85%

que esta vez los archivos utilizados presentaban el número de folio sin ningún tratamiento (uso de Hash).

Se programó una rutina en R que busca la presencia de un folio en alguna de las 30 columnas y asigna entonces al folio de dicha licencia `Reclama`, en caso contrario asignará `No Reclama`. Notar que con este criterio se tiene que los reclamos ascienden 121645 del universo de 5470619 licencias médicas emitidas, es decir aproximadamente 2.22% de las licencias emitidas reclaman, lo que no es contradictorio con las estadísticas oficiales, ya que la cantidad de reclamos es del orden de un 10 a 15 % del universo de licencias rechazadas que bien se sabe es un 10 a 15% del universo de licencias emitidas. Surgió en este punto del trabajo entonces la duda de sobre cómo se pueden identificar de manera adecuada las licencias rechazadas, y se averiguó de la contraparte técnica de la Superintendencia el criterio que es utilizado por ellos, criterio que será aplicado en la próxima subsección.

Se ejecutó la prueba. El árbol clasificador resultante es el siguiente:

```
node), split, n, deviance, yval, (yprob)
  * denotes terminal node

1) root 154417 213900.0 Reclama ( 0.48456 0.51544 )
 2) derecho_a_subsidio: A 75733 69370.0 No Reclama ( 0.82866
0.17134 )
 4) numero_dias_licencia < 9.5 32655 9386.0 No Reclama ( 0.96739
0.03261 ) *
 5) numero_dias_licencia > 9.5 43078 50800.0 No Reclama ( 0.72350
0.27650 )
 10) num_dias_previos_autorizados < 58.5 30840 30460.0 No Reclama (
0.80467
0.19533 )
 20) num_dias_incapacidad < 0.5 997 402.6 Reclama ( 0.05115
0.94885 ) *
 21) num_dias_incapacidad > 0.5 29843 27220.0 No Reclama ( 0.82984
0.17016 ) *
 11) num_dias_previos_autorizados > 58.5 12238 16950.0 No Reclama (
0.51896
0.48104 ) *
3) derecho_a_subsidio: B,C,ND 78684 67430.0 Reclama ( 0.15337
0.84663 )
6) licencia_mat_suplementaria: 1,ND 35219 38010.0 Reclama ( 0.23033
0.76967 )
12) redictamen: 2,ND 20854 26720.0 Reclama ( 0.33941 0.66059 )
 24) numero_dias_licencia < 11.5 4766 6224.0 No Reclama ( 0.64079
0.35921 ) *
 25) numero_dias_licencia > 11.5 16088 18100.0 Reclama ( 0.25012
0.74988 ) *
13) redictamen: 1 14365 7433.0 Reclama ( 0.07198 0.92802 ) *
7) licencia_mat_suplementaria: 2 43465 26500.0 Reclama ( 0.09102
0.90898 ) *
```

Y la precisión del árbol clasificador es la siguiente:

```
Classification tree:
tree(formula = f, data = train.set)
```

```

Variables actually used in tree construction:
[1] "derecho_a_subsidio"          "numero_dias_licencia"
"num_dias_previos_autorizados"
[4] "num_dias_incapacidad"      "licencia_mat_suplementaria"  "redictamen"
Number of terminal nodes:  8
Residual mean deviance:  0.7268 = 112200 / 154400
Misclassification error rate: 0.1477 = 22807 / 154417

```

lo que es bastante cercano al clasificador anterior. Se debe notar que la muestra es distinta y eso podría explicar las diferencias entre los resultados y los clasificadores.

La matriz de confusión en el conjunto de datos de test es:

prediction	No Reclama	Reclama
No Reclama	31598	6803
Reclama	5101	29485

lo cual da estadísticas mejores a las ya obtenidas. La clasificación correcta de licencias que presentan reclamos es de un 82% aproximadamente<sup>4</sup>.

Este ejemplo muestra que el uso de códigos Hash para los folios ha ocasionado algo de ruido en los resultados que se obtienen luego de aplicar el modelo.

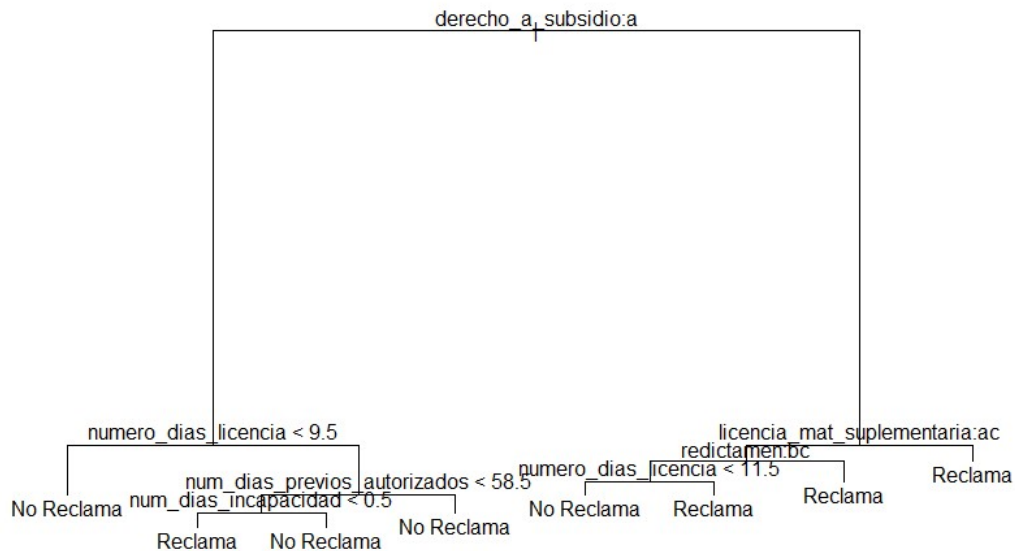


Figura 5: Clasificador, Ejemplo 2016 - Otra Definición de Reclamación

<sup>4</sup> Este porcentaje se calcula utilizando los valores tomados de la matriz de confusión, que indica que de un total de 36228 = 29485 + 6803 licencias que debiesen haber reclamado, hay 29485 que han sido correctamente clasificadas, esto entrega 29485/36228 ≈ 0.82, es decir 82%

## 2.2.6 Ejemplo Adicional Usando Criterio de Reclamo Basado en Folio: Licencias Rechazadas

Los ejemplos anteriores fueron corridos sobre todos los registros disponibles para el año 2016. Discusiones con contraparte técnica de la Superintendencia permitieron entender el criterio a utilizar para filtrar el universo de licencias para determinar sólo aquellas que han sido rechazadas. Para ello se programó una rutina en `\textsf{R}` para filtrar los registros sobre la base de los campos `\verb+estado_resolucion+` y `\verb+derecho_a_subsidio+`, los valores para dichos campos deben ser `\verb+2, 3 o 4+` y `\verb+B o C+` respectivamente. Estas licencias son denominadas *licencias rechazadas*. Posteriormente, la misma técnica de identificación de folio descrita en la subsección anterior permite identificar aquellas licencias que terminaron presentando reclamos (*licencias reclamadas*). La cantidad de licencias consideradas rechazadas al aplicar este criterio alcanza las 534590 licencias, de un total de 5470619 licencias para el 2016 lo que representa 9.8% aproximadamente.

Hay que notar que, para efectos de construir un árbol clasificador, al igual que antes, es necesario considerar un conjunto balanceado puesto que las licencias reclamadas siguen representando aproximadamente 10% del total de licencias rechazadas por lo que el clasificador tenderá a ignorar los casos minoritarios (reclamos). En efecto, si se mira el universo de datos de licencias rechazadas para el 2016 y se identifican los reclamos por el folio se obtienen las siguientes estadísticas:

No Reclama	Reclama
461103	73487

es decir, los casos de reclamo representan aproximadamente un 13.75% de los casos. Este número es cercano al número identificado por la contraparte técnica de la Superintendencia, pero ligeramente menor que el identificado para 2016. Una posible explicación para la diferencia es que podría haber ciertos casos de folios que no están siendo considerados, en particular hay ciertos expedientes que utilizan 29 campos de folios y el último `folio30` tiene listas separadas por comas, la función utilizada en R no puede hacerse cargo de separar dichas listas presentes en el campo `folio30`.

Se corrió el algoritmo. El árbol clasificador resultante es el siguiente:

```
node), split, n, deviance, yval, (yprob)
  * denotes terminal node

1) root 101935 141300 No Reclama ( 0.50160 0.49840 )
  2) monto_subsidio_liquidado < -0.5 63223 82890 Reclama ( 0.36373
0.63627 )
    4) numero_dias_licencia < 9.5 7167 8795 No Reclama ( 0.69680
0.30320 ) *
    5) numero_dias_licencia > 9.5 56056 70380 Reclama ( 0.32114
0.67886 )
      10) causa_rechazo: 2,3,4,5,ND 34339 45880 Reclama ( 0.38851
0.61149 ) *
      11) causa_rechazo: 1 21717 22590 Reclama ( 0.21462
0.78538 ) *
  3) monto_subsidio_liquidado > -0.5 38712 45400 No Reclama ( 0.72678
```



```

0.27322 )
6) numero_dias_licencia < 11.5 10560 5676 No Reclama ( 0.92405
0.07595 ) *
7) numero_dias_licencia > 11.5 28152 36360 No Reclama ( 0.65278
0.34722 ) *

```

Y la precisión del árbol clasificador es la siguiente:

```

Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "monto_subsidio_liquido" "numero_dias_licencia" "causa_rechazo"
Number of terminal nodes: 5
Residual mean deviance: 1.17 = 119300 / 101900
Misclassification error rate: 0.3017 = 30752 / 101935

```

lo que es un empeoramiento respecto del clasificador anterior (notar que la muestra es distinta eso sí).

La matriz de confusión en el conjunto de datos de test es:

prediction	No Reclama	Reclama
No Reclama	14017	5678
Reclama	7891	16507

lo cual da estadísticas peores a las ya obtenidas. La clasificación correcta de licencias que presentan reclamos es de un 75% aproximadamente.

Este experimento confirma que el algoritmo es sensible a la definición del universo, puesto que los resultados cambian de manera no menor.

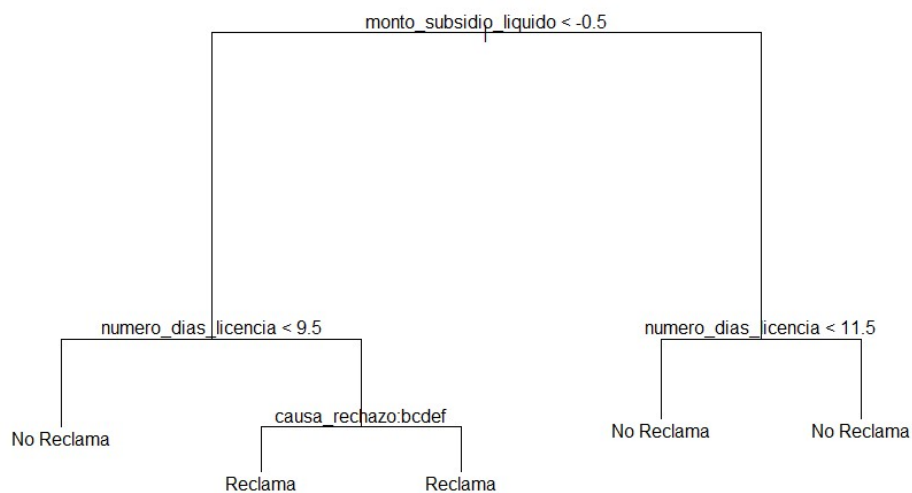


Figura 6: Clasificador, Ejemplo 2016 - Otra Definición de Licencias Rechazadas

## 2.3 Uso de Redes Neuronales

Una red neuronal, es un tipo de modelo que trata de imitar la forma en que las neuronas operan en el cerebro humano. El modelo básico es tal que si la suma ponderada (por pesos) de todas las conexiones entrantes a una neurona superan un cierto valor preestablecido, entonces la neurona emitirá una señal (o producirá sinapsis) lo que a su vez comunicará con todas las neuronas con las cuales ella está conectada. Hay varias variantes de este modelo básico, lo que es importante en todo caso es notar que el algoritmo de entrenamiento de una red neuronal procede primero aplicando ejemplos para los cuales la red es entrenada, es decir, los pesos de las conexiones entre neuronas son modificados de tal forma de producir el resultado esperado para ese ejemplo. Después de presentar varios ejemplos a la red neuronal, esta aprende de ellos y fija los valores de los pesos de las conexiones basados en los ejemplos presentados. Después de terminado el proceso de entrenamiento, la red neuronal necesita ser probada en el conjunto de validación separado para tal efecto antes de proceder al entrenamiento (al igual que en el caso de los árboles clasificadores, este porcentaje asciende a 30% de los datos para el conjunto balanceado).

### 2.3.1 Ejemplo Usando Criterio de Reclamo Basado en Folio

Inicialmente se consideraron las mismas variables que para el árbol de decisión, claro que en este caso estas fueron codificadas utilizando variables binarias (variables dummy). El listado de variables utilizadas en el modelo viene dado por:

```
articulo_772, articulo_77ND, edad_trabajador,
sexo_trabajador2, numero_dias_licencia, tipo_licencia_lm2,
tipo_licencia_lm3, tipo_licencia_lm4, tipo_licencia_lm5,
tipo_licencia_lm6, tipo_licencia_lm7, recuperabilidad_laboral2,
inicio_tramite_invalidez2, tipo_reposo2, lugar_reposo2, lugar_reposo3,
lugar_reposoND, tipo_profesional2, tipo_profesional3,
num_dias_previos_autorizados, causa_rechazo2, causa_rechazo3,
causa_rechazo4, causa_rechazo5, causa_rechazoND, derecho_a_subsidioB,
derecho_a_subsidioC, derecho_a_subsidioND, actividad_laboral_trabajador1,
actividad_laboral_trabajador2, actividad_laboral_trabajador3,
actividad_laboral_trabajador4, actividad_laboral_trabajador5,
actividad_laboral_trabajador6, actividad_laboral_trabajador7,
actividad_laboral_trabajador8, actividad_laboral_trabajador9,
actividad_laboral_trabajadorND, ocupacion_trabajador12,
ocupacion_trabajador13, ocupacion_trabajador14, ocupacion_trabajador15,
ocupacion_trabajador16, ocupacion_trabajador17, ocupacion_trabajador18,
ocupacion_trabajador19, calidad_trabajador2, calidad_trabajador3,
calidad_trabajador4, num_dias_incapacidad, monto_subsidio_liquido,
monto_cotizaciones_salud, monto_cotizaciones_pension,
monto_base_calculo_subsidio, monto_liquido_pagado,
remuneracion_imp_mes_anterior, licencia_mat_suplementaria2,
licencia_mat_suplementariaND, licencia_modificada2, licencia_modificadaND,
tipo_licencia_medica_resuelta2, tipo_licencia_medica_resuelta3,
tipo_licencia_medica_resuelta4, tipo_licencia_medica_resuelta5,
tipo_licencia_medica_resuelta6, tipo_licencia_medica_resuelta7,
estado_resolucion2, estado_resolucion3, estado_resolucion4,
estado_resolucion5, tipo_resolucion2, tipo_resolucionND,
```

redictamen2, redictamenND, tipo\_entidad\_pagadora\_subsidioB,  
tipo\_entidad\_pagadora\_subsidioC, tipo\_entidad\_pagadora\_subsidioD,  
tipo\_entidad\_pagadora\_subsidioE, tipo\_entidad\_pagadora\_subsidioF,  
tipo\_entidad\_pagadora\_subsidioG, tipo\_entidad\_pagadora\_subsidioH,  
tipo\_entidad\_pagadora\_subsidioND.

Se intentó entrenar la red neuronal sobre este conjunto de variables, sin embargo, después de aproximadamente dos horas de espera, el algoritmo no exhibe convergencia a nada. Se decidió abortar la corrida y correr el algoritmo con menos variables.

Se corrió con menos variables y el resultado fue similar. Las variables usadas en este caso son:

articulo\_772, articulo\_77ND, edad\_trabajador, sexo\_trabajador2,  
numero\_dias\_licencia, tipo\_licencia\_lm2, tipo\_licencia\_lm3,  
tipo\_licencia\_lm4, tipo\_licencia\_lm5, tipo\_licencia\_lm6,  
tipo\_licencia\_lm7, recuperabilidad\_laboral2, inicio\_tramite\_invalidez2,  
num\_dias\_previos\_autorizados, causa\_rechazo2, causa\_rechazo3,  
causa\_rechazo4, causa\_rechazo5, causa\_rechazoND, derecho\_a\_subsidioB,  
derecho\_a\_subsidioC, derecho\_a\_subsidioND, num\_dias\_incapacidad,  
licencia\_mat\_suplementaria2, licencia\_mat\_suplementariaND,  
licencia\_modificada2, licencia\_modificadaND,  
tipo\_licencia\_medica\_resuelta2,  
tipo\_licencia\_medica\_resuelta3, tipo\_licencia\_medica\_resuelta4,  
tipo\_licencia\_medica\_resuelta5, tipo\_licencia\_medica\_resuelta6,  
tipo\_licencia\_medica\_resuelta7, estado\_resolucion2,  
estado\_resolucion3, estado\_resolucion4, estado\_resolucion5,  
tipo\_resolucion2, tipo\_resolucionND

Después de averiguar en foros y algunos textos, se observa que las redes neuronales no escalan bien en términos de número de variables o número de casos. Se recomienda no utilizarlas en este caso puesto que demoran demasiado y no permitirían su uso a la escala necesaria.

## 2.4 Uso de Modelos de Regresión

Un modelo de regresión es un modelo que trata de especificar una variable (llamada dependiente) en función de un conjunto de variables o regresores (llamados variables independientes). La típica forma funcional de una regresión es de la forma:

$$y = \beta_0 + \sum_{j=1}^n \beta_j x_j$$

donde  $y$  es la variable dependiente y los  $x_j$  son las variables independientes.

En el caso particular de que la variable a ser regresionada sea binaria (como es el caso) entonces el modelo de regresión se conoce con el nombre de modelo logit (o *logistic regression*). Un modelo relacionado es el modelo probit en que se pretende estimar la probabilidad de que una observación determinada pertenezca a alguna clase particular (esto para el caso de variables binarias).

Hay que notar que el principal problema al que se enfrenta el investigador cuando se usan modelos de regresión tiene que ver con la especificación del modelo. La especificación del modelo guarda relación con la selección de variables que intervienen en el. Una alternativa podía ser el uso de búsquedas secuenciales de especificación, es decir, probar todas las posibles combinaciones de variables. Lamentablemente, en presencia de un alto número de variables esto no es factible. Por ejemplo, si tenemos 30 variables independientes para un modelo, el número de posibles combinaciones de ellas es  $2^n$ , lo que es lo mismo que 1.073.741.824 posibles combinaciones ( $2^{30}$ ), si el computador pudiese evaluar 1000 regresiones por segundo (lo que es optimista) entonces demoraría aproximadamente 12 días y medio en revisar todas las posibles especificaciones, lo que a todas luces no es muy práctico.

Junto al problema anterior, tenemos el hecho de que los modelos son ajustados con datos y que cuando esos datos cambian (como lo será en el caso de las licencias médicas) entonces habrá que cambiar el modelo, lo que significa un alto costo de mantención puesto que el fenómeno es dinámico. Esto es un punto negativo cuando se compara con los árboles de clasificación que producen como salida un modelo que ha sido automáticamente ajustado.

#### 2.4.1 Ejemplo Usando Criterio de Reclamo Basado en Folio

Se seleccionaron las siguientes variables para correr un modelo:

```
tipo_formulario, articulo_77, edad_trabajador, sexo_trabajador,
numero_dias_licencia, licencia_mat_suplementaria, tipo_licencia_lm,
recuperabilidad_laboral, inicio_tramite_invalidez, licencia_modificada,
tipo_licencia_medica_resuelta, num_dias_previos_autorizados,
estado_resolucion, tipo_resolucion, redictamen, causa_rechazo,
derecho_a_subsidio, num_dias_incapacidad, motivo_rechazo
```

Se procedió a correr el modelo utilizando el software de código abierto GRETl en una base de datos balanceada, se tuvieron que botar los registros con observaciones faltantes (de lo contrario la regresión no corre) y para las 133580 observaciones que quedaron se obtuvo lo siguiente:

```
Function evaluations: 172
Evaluations of gradient: 34
```

```
Model 6: Ordered Logit, using observations 1-133580
Dependent variable: Reclamo
Standard errors based on Hessian
```

	coefficient	std. error	z	p-value	
tipo_formulario	-0.0633504	0.0147130	-4.306	1.66e-05	***
articulo_77	-5.34052	0.119430	-44.72	0.0000	***
edad_trabajador	0.0138091	0.000754230	18.31	7.04e-075	***
sexo_trabajador	0.379099	0.0194475	19.49	1.25e-084	***
numero_dias_li~	0.0577191	0.00115105	50.14	0.0000	***
licencia_mat_s~	1.92627	0.0600622	32.07	1.11e-225	***
tipo_licencia~	0.637924	0.0373577	17.08	2.24e-065	***
recuperabilida~	-0.192410	0.0392960	-4.896	9.76e-07	***
inicio_tramite~	0.0701207	0.0307778	2.278	0.0227	**

licencia_modif~	-0.982915	0.0262889	-37.39	5.90e-306	***
tipo_licencia_~	-0.639419	0.0280713	-22.78	7.50e-115	***
num_dias_previ~	0.00530754	0.000123794	42.87	0.0000	***
estado_resoluc~	-0.340496	0.0157866	-21.57	3.53e-103	***
tipo_resolucion	-0.247867	0.0314726	-7.876	3.39e-015	***
redictamen	0.873696	0.0669328	13.05	6.08e-039	***
causa_rechazo	0.886436	0.0189898	46.68	0.0000	***
derecho_a_subs~	0.712245	0.0153649	46.36	0.0000	***
num_dias_incap~	-0.0472719	0.00125566	-37.65	0.0000	***
motivo_rechazo	0.221361	0.00614310	36.03	2.45e-284	***
cut1	-0.0521200	0.107470	-0.4850	0.6277	
Mean dependent var	1.606056	S.D. dependent var	0.488624		
Log-likelihood	-43188.24	Akaike criterion	86416.47		
Schwarz criterion	86612.52	Hannan-Quinn	86475.20		

Number of cases 'correctly predicted' = 118408 (88.6%)  
Likelihood ratio test: Chi-square(19) = 93591.6 [0.0000]

Este reporte para el modelo Logit indica que para el conjunto de entrenamiento se obtiene una precisión del 88.6%, es decir, el modelo ajustado aplicados a los datos es capaz de predecir correctamente 118408 casos de un total de 133580.

Se procedió a correr un modelo probit codificando para ello la variable Reclamo como binaria (1 si Reclama, o si no) y los resultados de la regresión fueron:

Model 7: Probit, using observations 1-133580  
Dependent variable: DReclamo\_2  
Standard errors based on Hessian

	coefficient	std. error	z	p-value	
const	-0.00872669	0.0556447	-0.1568	0.8754	
tipo_formulario	-0.0249121	0.00779468	-3.196	0.0014	***
articulo_77	-3.05071	0.0634982	-48.04	0.0000	***
edad_trabajador	0.00784612	0.000401114	19.56	3.34e-085	***
sexo_trabajador	0.208326	0.0103929	20.05	2.23e-089	***
numero_dias_lice~	0.0281152	0.000533051	52.74	0.0000	***
licencia_mat_sup~	1.12763	0.0333281	33.83	6.20e-251	***
tipo_licencia_lm	0.333315	0.0165761	20.11	6.25e-090	***
recuperabilidad_~	-0.0816375	0.0217267	-3.757	0.0002	***
inicio_tramite_i~	0.0143328	0.0170459	0.8408	0.4004	
licencia_modific~	-0.540280	0.0139565	-38.71	0.0000	***
tipo_licencia_me~	-0.344632	0.0122865	-28.05	4.02e-173	***
num_dias_previos~	0.00244905	5.34724e-05	45.80	0.0000	***
estado_resolucion	-0.174396	0.00856048	-20.37	2.95e-092	***
tipo_resolucion	-0.131722	0.0151190	-8.712	2.98e-018	***
redictamen	0.472029	0.0333954	14.13	2.33e-045	***
causa_rechazo	0.456566	0.00867250	52.65	0.0000	***
derecho_a_subsid~	0.459110	0.00848051	54.14	0.0000	***
num_dias_incapac~	-0.0221624	0.000611178	-36.26	6.48e-288	***
motivo_rechazo	0.117029	0.00312633	37.43	1.12e-306	***
Mean dependent var	0.606056	S.D. dependent var	0.488624		
McFadden R-squared	0.508904	Adjusted R-squared	0.508681		
Log-likelihood	-43983.88	Akaike criterion	88007.75		
Schwarz criterion	88203.80	Hannan-Quinn	88066.49		

```
Number of cases 'correctly predicted' = 118182 (88.5%)
f(beta'x) at mean of independent vars = 0.373
Likelihood ratio test: Chi-square(19) = 91157.6 [0.0000]
```

```
          Predicted
          0      1
Actual 0  43264  9359
        1   6039 74918
```

Excluding the constant, p-value was highest for variable 9 (inicio\_tramite\_invalidez)

```
Test for normality of residual -
Null hypothesis: error is normally distributed
Test statistic: Chi-square(2) = 499.6
with p-value = 3.26058e-109
```

Se ve que este modelo es tan bueno como el anterior. Sin embargo, la variable `inicio_tramite_invalidez` ya no es estadísticamente significativa. También se observa que los residuos no son normalmente distribuidos lo que eventualmente trae como consecuencia que el estimador no sea el mejor lineal insesgado.

## 2.5 Pruebas adicionales utilizando rechazos y reclamos definidos por SUSESO

Para poder enfrentar el problema ya identificado de tener un menor número de licencias identificadas como rechazadas menor a lo que indican las estadísticas históricas, problema que igualmente afecta las licencias reclamadas, se solicitó a SUSESO que creara un conjunto de datos para análisis consistente del conjunto de licencias identificadas como rechazadas y de aquellas, el conjunto de licencias identificadas como reclamadas. Se corrió un modelo de árbol de decisión basado en estos datos y a continuación se intentó un modelo de *ensemble* que permite corregir la tendencia al sobreajuste que presentan los modelos de árboles de decisión.

### 2.5.1 Modelo de árbol de decisión utilizando identificaciones provistas por SUSESO

Basado en el set de datos provisto por SUSESO, se corrió el mismo tipo de modelo de árbol clasificador utilizado anteriormente. Los resultados se muestran a continuación:

```
Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "estado_resolucion"      "num_dias_licencia"      "licencia_materna_sup"
Number of terminal nodes: 5
Residual mean deviance: 1.019 = 56960 / 55920
Misclassification error rate: 0.2348 = 13133 / 55924
node), split, n, deviance, yval, (yprob)
* denotes terminal node

1) root 55924 77240 1 ( 0.4645 0.5355 )
2) estado_resolucion: 1,3,4 12907 8437 0 ( 0.8992 0.1008 ) *
3) estado_resolucion: 2 43017 54800 1 ( 0.3340 0.6660 )
6) num_dias_licencia < 11.5 7231 9463 0 ( 0.6384 0.3616 )
12) licencia_materna_sup: *,1 5088 5731 0 ( 0.7492 0.2508 ) *
13) licencia_materna_sup: 2 2143 2836 1 ( 0.3752 0.6248 ) *
7) num_dias_licencia > 11.5 35786 41920 1 ( 0.2725 0.7275 )
```

```

14) licencia_materna_sup: *,1 20595 26920 1 ( 0.3601 0.6399 ) *
15) licencia_materna_sup: 2 15191 13040 1 ( 0.1537 0.8463 ) *

```

```

prediction    0    1
             0 11962 1604
             1 13403 24080

```

De la matriz de confusión del ejercicio podemos concluir que las licencias médicas rechazadas que presentan reclamos y que terminan mal clasificados son del orden del 6%, por otro lado, las licencias médicas rechazadas que no presentan reclamos son mal clasificadas en un 52%, lo que coloca dudas sobre la eficacia del método utilizado. Cabe notar que los resultados resumidos en la matriz de confusión corresponden a los que se obtienen de aplicar el árbol clasificador a los datos que se han separado como test (un 30% de la muestra balanceada), mirado globalmente, el error el método es de un 30% aproximadamente.

Dado que se debe construir un conjunto balanceado para entrenar el árbol, surge la duda de si estos resultados obtenidos dependen de la muestra en cuestión. Para responder a esta pregunta es que se decidió correr el modelo varias veces (para distintas muestras) y ver que tan sensible es la calidad del modelo. Los resultados de las distintas corridas se muestran a continuación. Como se puede desprender de los resultados tanto de entrenamiento como de testeo, las corridas con bastante consistentes en términos de su calidad, así que podemos concluir que la forma en que se construye el conjunto balanceado no introduce ruido adicional al proceso (esto probablemente debido al gran número de casos disponibles).

#### 2.5.1.1 Corrida adicional

```

Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "estado_resolucion"      "num_dias_licencia"      "licencia_materna_sup"
Number of terminal nodes: 5
Residual mean deviance: 1.02 = 56880 / 55750
Misclassification error rate: 0.2354 = 13127 / 55756
node), split, n, deviance, yval, (yprob)
* denotes terminal node

```

```

1) root 55756 76960 1 ( 0.4613 0.5387 )
2) estado_resolucion: 1,3,4 12958 8661 0 ( 0.8958 0.1042 ) *
3) estado_resolucion: 2 42798 54260 1 ( 0.3297 0.6703 )
6) num_dias_licencia < 10.5 5781 7333 0 ( 0.6700 0.3300 )
12) licencia_materna_sup: *,1 4099 4272 0 ( 0.7846 0.2154 ) *
13) licencia_materna_sup: 2 1682 2251 1 ( 0.3906 0.6094 ) *
7) num_dias_licencia > 10.5 37017 43660 1 ( 0.2765 0.7235 )
14) licencia_materna_sup: *,1 21384 28010 1 ( 0.3625 0.6375 ) *
15) licencia_materna_sup: 2 15633 13690 1 ( 0.1590 0.8410 ) *

```

```

prediction    0    1
             0 11889 1397
             1 13630 24133

```

#### 2.5.1.2 Corrida adicional

```

Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "estado_resolucion"      "num_dias_licencia"      "licencia_materna_sup"
Number of terminal nodes: 4
Residual mean deviance: 1.036 = 57860 / 55830
Misclassification error rate: 0.244 = 13624 / 55835
node), split, n, deviance, yval, (yprob)
  * denotes terminal node

```

- 1) root 55835 77120 1 ( 0.4642 0.5358 )
- 2) estado\_resolucion: 1,3,4 12873 8438 0 ( 0.8988 0.1012 ) \*
- 3) estado\_resolucion: 2,5 42962 54730 1 ( 0.3340 0.6660 )
- 6) num\_dias\_licencia < 10.5 5866 7416 0 ( 0.6729 0.3271 ) \*
- 7) num\_dias\_licencia > 10.5 37096 44020 1 ( 0.2804 0.7196 )
- 14) licencia\_materna\_sup: \*,1 21369 28120 1 ( 0.3683 0.6317 ) \*
- 15) licencia\_materna\_sup: 2 15727 13880 1 ( 0.1610 0.8390 ) \*

```

prediction      0      1
              0 15911 2481
              1  9540 23117

```

### 2.5.1.3 Corrida adicional

```

Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "estado_resolucion"      "num_dias_licencia"      "tipo_archivo"
"licencia_materna_sup"
Number of terminal nodes: 5
Residual mean deviance: 1.016 = 57130 / 56200
Misclassification error rate: 0.2322 = 13052 / 56209
node), split, n, deviance, yval, (yprob)
  * denotes terminal node

```

- 1) root 56209 77580 1 ( 0.4611 0.5389 )
- 2) estado\_resolucion: 1,4 13009 8590 0 ( 0.8977 0.1023 ) \*
- 3) estado\_resolucion: 2,3 43200 54770 1 ( 0.3297 0.6703 )
- 6) num\_dias\_licencia < 11.5 7247 9501 0 ( 0.6363 0.3637 )
- 12) tipo\_archivo: LM 2233 2961 1 ( 0.3780 0.6220 ) \*
- 13) tipo\_archivo: UNICO 5014 5625 0 ( 0.7513 0.2487 ) \*
- 7) num\_dias\_licencia > 11.5 35953 41790 1 ( 0.2678 0.7322 )
- 14) licencia\_materna\_sup: \*,1 20648 26790 1 ( 0.3520 0.6480 ) \*
- 15) licencia\_materna\_sup: 2 15305 13170 1 ( 0.1543 0.8457 ) \*

```

prediction      0      1
              0 11957 1463
              1 13720 23909

```

### 2.5.2 Uso de Modelo de Ensemble

Dado que aparentemente el modelo de árboles de decisión tiende a sobre ajustar, es que se ha decidido utilizar métodos de *ensemble*. Los métodos de *ensemble* se basan en la idea de que las fluctuaciones entre distintas muestras pueden ser compensadas por la vía de construir un modelo que considere varias repeticiones para varias instancias y lo que posteriormente se hace es promediar las respuestas obtenidas si es que los arboles son de regresión y si no se utiliza votación por mayoría si es que los arboles son de clasificación. Por ejemplo, si construyo 10 modelos de árbol



clasificador basado en 10 muestras distintas, y en 8 de aquellos modelos una licencia rechazada se clasifica como reclamo entonces el modelo de *ensemble* predecirá reclamo para esta licencia como respuesta de salida (la mayoría así lo opina). Se utilizaron dos algoritmos disponibles, estos difieren en algunos detalles de su implementación, pero ambos están basados en los mismos principios. Uno de ellos (Bosque Aleatorio) tiene además la ventaja de producir un listado ordenado de variables relevantes que se podría considerar como una ventaja al momento de socializar el modelo al público. En todos estos modelos el indicador clave de la calidad del entrenamiento es el *out-of-bag* (OOB) error que es un estimador del error del método. Se experimento de todas maneras separando un conjunto como test y se aplicó el modelo ajustado para predecir sobre este conjunto de test, estos datos vienen entregados como matrices de confusión en cada uno de los casos. Se percibe que el estimador OOB es en general generoso en comparación con los tests realizados con los datos que fueron separados para tal efecto.

### 2.5.2.1 Bagging

Los resultados de aplicar la técnica de *bagging* vienen dados a continuación.

#### 2.5.2.1.1 Corrida 1

Bagging classification trees with 25 bootstrap replications

```
Call: bagging.data.frame(formula = f, data = train.set, coob = TRUE)
```

Out-of-bag estimate of misclassification error: 0.1979 (Mejor que el 23% de un solo arbol clasificador)

```
my.prediction.rf      0      1
                    0 17042  3793
                    1  8559 21655
```

Esto se traduce en un 14.9% de malas clasificaciones para los reclamos, sin embargo, baja la mala clasificación de rechazos que no reclaman es del orden de 33.4%.

#### 2.5.2.1.2 Corrida 2

Bagging classification trees with 25 bootstrap replications

```
Call: bagging.data.frame(formula = f, data = train.set, coob = TRUE)
```

Out-of-bag estimate of misclassification error: 0.1982

```
my.prediction.rf      0      1
                    0 16072  3851
                    1  9273 21853
```

Esto se traduce en un 14.99% de malas clasificaciones para los reclamos y 36.59% para los rechazos que no reclaman

### 2.5.2.1.3 Corrida 3

Bagging classification trees with 25 bootstrap replications

```
Call: bagging.data.frame(formula = f, data = train.set, coob = TRUE)
```

Out-of-bag estimate of misclassification error: 0.1989

```
my.prediction.rf      0      1
                    0 16129  3710
                    1  9460 21750
```

Esto se traduce en un 14.58% de malas clasificaciones para los reclamos y 36.97% para los rechazos que no reclaman

### 2.5.2.2 Bosque Aleatorio

Para el modelo de bosque aleatorio se realizaron algunas corridas también. Los resultados vienen a continuación.

#### 2.5.2.2.1 Corrida 1

Call:

```
randomForest(formula = f, data = na.omit(train.set), mtry = 13, importance = TRUE)
              Type of random forest: classification
              Number of trees: 500
```

No. of variables tried at each split: 13

OOB estimate of error rate: 21.18%

Confusion matrix:

```
      0      1 class.error
0 15856  4496  0.2209119
1  3106 12427  0.1999614
```

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
tipo_archivo	-0.7757465	12.121565	12.538952	8.332841
operador	18.4522590	23.394092	26.383214	146.284167
tipo_formulario	23.3228035	37.242971	40.038945	266.955046
articulo_77bis	3.2498879	12.194952	12.409796	7.927658
edad_trabajador	6.1950192	175.303001	137.954163	2753.724909
sexo_trabajador	10.2831490	93.579236	88.132805	314.760749
num_dias_licencia	102.6382640	217.485557	211.556330	2420.648033
licencia_materna_sup	-0.7863704	13.459622	13.214951	9.901353
tipo_licencia	26.6768136	82.923992	85.904026	339.807195
recuperabilidad_laboral	5.2033483	31.439852	31.653791	105.803698
inicio_tramite_invalidez	10.8080391	19.299873	23.955416	47.441255
tipo_reposo	6.5500295	10.779580	13.451943	23.629751
jornada_reposo	7.4535486	13.019013	15.320772	35.189951
lugar_reposo	25.3159518	65.939595	65.099621	423.740591
tipo_profesional	-2.4235939	6.769575	2.518769	6.773137
licencia_modificada	2.9280005	12.883559	13.214161	9.933587
num_dias_incapacidad_aut	45.4649713	24.890094	36.418451	1897.876314
período	10.7056295	35.063349	42.552346	147.199675
num_dias_previos_autorizados	48.3679708	99.834935	139.984478	1487.501269
estado_resolucion	13.2772835	18.186938	18.213098	1142.568327
tipo_resolucion	0.0000000	0.000000	0.000000	0.000000
redictamen	0.3822952	11.975313	12.045096	7.327568
causa_rechazo	23.4805179	106.992693	86.316942	648.979674
tipo_reposo_autorizado	4.7334257	13.807240	15.061457	22.680719
jornada_reposo_autorizada	5.3395914	14.483605	15.469312	31.501407
derecho_a_subsidio	11.7787265	9.759392	16.735608	463.518383
region_empleador	69.4882734	203.219478	209.167420	1168.993156
ocupacion_trabajador	20.2500245	129.133966	115.939023	1340.030227
tipo_regimen_previsional	0.7523768	13.438614	13.683875	10.102349

calidad_trabajador	9.3623354	77.652275	73.661541	386.234081
tipo_entidad_pagadora_subsidio	-12.6376390	46.274204	36.237785	146.237785

```
my.prediction.rf      0      1
                    0 8129 2958
                    1 2965 9886
```

El modelo aplicado a un conjunto test (resumido en la matriz de confusión) entrega que un 23.03% de los reclamos están mal clasificados y que un 26.72% de los no reclamos también terminan mal clasificados.

### 2.5.2.2.2 Corrida 2

```
Call:
  randomForest(formula = f, data = na.omit(train.set), mtry = 13, importance = TRUE)
  Type of random forest: classification
  Number of trees: 500
  No. of variables tried at each split: 13
```

OOB estimate of error rate: 21.42%

Confusion matrix:

	0	1	class.error
0	15734	4489	0.2219750
1	3141	12262	0.2039213

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
tipo_archivo	-2.8695662	12.789227	12.854203	7.434324
operador	13.4863977	27.359321	27.767874	142.697686
tipo_formulario	25.0970656	36.352786	38.810144	271.105512
articulo_77bis	-3.6031043	11.391069	11.024216	7.565629
edad_trabajador	14.9463467	182.662792	164.570721	2762.216117
sexo_trabajador	8.8097040	91.174105	82.987240	336.486129
num_dias_licencia	103.1884422	189.603188	203.947727	2383.353948
licencia_materna_sup	-4.6798271	11.146940	11.286568	6.736388
tipo_licencia	34.2775400	90.990799	92.189229	307.970726
recuperabilidad_laboral	-2.5548413	27.477041	24.695319	95.353234
inicio_tramite_invalidez	7.6525777	26.131834	28.816424	44.759906
tipo_reposo	8.4691964	10.046686	13.190203	24.536874
jornada_reposo	10.8150626	10.608320	15.276279	33.775732
lugar_reposo	32.0114626	50.916729	61.353540	429.289223
tipo_profesional	-7.2719858	3.748116	-5.777847	5.711684
licencia_modificada	1.5927110	11.315038	12.086354	7.998578
num_dias_incapacidad_aut	47.7241724	25.752063	37.960394	1908.592087
período	3.8717297	39.038354	43.670853	142.661896
num_dias_previos_autorizados	46.2083882	94.860113	119.207604	1418.441896
estado_resolucion	16.5741352	16.821390	17.883697	1100.313637
tipo_resolucion	0.0000000	0.000000	0.000000	0.000000
redictamen	-1.1024728	11.773352	12.261759	6.538237
causa_rechazo	25.7477228	116.653091	105.676574	647.950513
tipo_reposo_autorizado	4.7392396	10.466535	12.151686	22.621307
jornada_reposo_autorizada	5.5579379	8.635899	10.870788	29.777347
derecho_a_subsidio	22.0449387	9.451466	17.847458	475.150436
region_empleador	65.8962750	174.922246	188.129766	1167.816970
ocupacion_trabajador	21.0705595	130.350398	119.154281	1327.499826
tipo_regimen_previsional	-3.1393983	10.937781	11.154165	8.851751
calidad_trabajador	11.7335244	78.746701	70.583971	400.147410
tipo_entidad_pagadora_subsidio	-0.9576982	47.465489	45.631302	147.539635

```
my.prediction.rf      0      1
                    0 8215 2836
                    1 2910 10006
```

El modelo aplicado a un conjunto test (resumido en la matriz de confusión) entrega que un 22.08% de los reclamos están mal clasificados y que un 26.15% de los no reclamos también terminan mal clasificados.

### 2.5.2.2.3 Corrida 3

```
Call:
  randomForest(formula = f, data = na.omit(train.set), mtry = 13,      importance = TRUE)
      Type of random forest: classification
      Number of trees: 500
No. of variables tried at each split: 13

      OOB estimate of error rate: 21.52%
Confusion matrix:
      0      1 class.error
0 15835 4553 0.2233176
1  3117 12135 0.2043666

      0      1 MeanDecreaseAccuracy MeanDecreaseGini
tipo_archivo      0.06689202 12.754562      12.741714      7.020761
operador      21.20258443 25.569049      28.428024      149.394918
tipo_formulario      25.06105151 33.885633      36.474029      253.534347
articulo_77bis      -1.71465663 11.969068      11.936681      8.498237
edad_trabajador      11.61181768 165.108753      139.151965      2747.024535
sexo_trabajador      13.75369842 91.049552      82.169978      336.187238
num_dias_licencia      113.53516264 199.337362      221.872341      2310.662545
licencia_materna_sup      -1.51429683 11.731467      11.510053      5.885644
tipo_licencia      28.88158896 86.295492      83.388353      317.621561
recuperabilidad_laboral      5.09599377 25.649198      27.256035      89.520811
inicio_tramite_invalidez      11.49563153 24.976432      29.559291      43.862663
tipo_reposo      5.30269652 8.487855      10.370399      25.533015
jornada_reposo      9.05340770 8.805669      12.767083      37.838059
lugar_reposo      27.65662174 63.973853      69.680812      420.472511
tipo_profesional      -0.37276188 5.372495      3.994564      7.137931
licencia_modificada      0.66732085 11.943893      12.550080      7.735288
num_dias_incapacidad_aut      45.34027695 20.430453      33.384257      1562.635694
periodo      1.15759300 47.631046      50.528166      149.123677
num_dias_previos_autorizados      37.55393903 92.506965      120.419552      1452.662620
estado_resolucion      16.99669377 22.767716      22.788824      1451.098419
tipo_resolucion      0.00000000 0.000000      0.000000      0.000000
redictamen      -0.87841359 12.274050      12.078693      7.742270
causa_rechazo      31.74180508 95.208763      83.692275      632.644430
tipo_reposo_autorizado      1.15796145 10.623834      10.212582      20.106869
jornada_reposo_autorizada      1.56364966 7.955369      7.970189      29.824622
derecho_a_subsidio      5.41675832 10.707988      15.124303      532.207262
region_empleador      76.80387348 169.316136      186.788986      1139.567200
ocupacion_trabajador      27.22376465 128.543600      122.625064      1307.579299
tipo_regimen_previsional      -0.82018415 11.976231      12.105284      9.109346
calidad_trabajador      8.10400511 77.792696      68.149892      407.762831
tipo_entidad_pagadora_subsidio      -8.81447668 46.682570      38.784904      150.039305

my.prediction.rf      0      1
      0 7988 2474
      1 2963 10539
```

El modelo aplicado a un conjunto test (resumido en la matriz de confusión) entrega que un 19.01% de los reclamos están mal clasificados y que un 27.06% de los no reclamos también terminan mal clasificados.

## 2.6 Conclusiones

A lo largo de este reporte hemos podido ver los intentos hechos para poder extraer un modelo clasificador a partir de los datos provistos. Respecto de los datos, éstos no presentan la mejor calidad posible puesto que hay varios faltantes.

La primera técnica empleada es de fácil utilización y permite de manera bien directa construir un modelo clasificador. Al utilizar esta técnica se descubrió que la falta de balanceo del conjunto de datos entrega clasificadores que por la vía de clasificar todas las licencias como No Reclamo tiene una tasa de éxito bastante alta. A continuación, se produjeron ejemplos con tasas de éxito de un 65% aproximadamente cuando el

conjunto de datos utilizado se basaba en una clasificación de licencia presentada utilizando sólo el RUT como regla de emparejamiento de casos en las tablas. Cuando finalmente se procedió a utilizar el folio de la licencia, fue posible conseguir modelos cuya precisión es del orden de 85% que es considerado bastante aceptable.

Las redes neuronales mostraron no ser implementables dado que tomaron demasiado tiempo en correr. Esto es a causa del número de variables que se pone peor en el caso de variables categóricas (que deben ser codificadas como dummies) y además a causa del número de casos de entrenamiento. Se recomienda no utilizar este tipo de modelos puesto que en la práctica no se pueden correr en tiempos razonables.

Los modelos de regresión si bien son fáciles de aplicar, no mejoran sustancialmente la calidad de la clasificación ya obtenida utilizando árboles y las diferencias podrían eventualmente responder a una razón estadística. Además, el uso de estos modelos requiere la codificación de las variables categóricas mediante el uso de variables dummy que sólo hacen que el problema sea más grande al aumentar el número de variables, hecho que a la larga trae consecuencias computacionales al aumentar el tamaño de la ya abultada base de datos. Se recomienda considerar el uso de este tipo de modelos como complementario a los modelos basados en árboles (en particular dada su consistencia).

Respecto de las variables a utilizar, modelos tales como los de regresión no permiten el uso de casos con faltantes en las variables. Debido a esto, es eventualmente problemático utilizar tantas variables en el modelo y deberá ser parte de un estudio futuro cuáles variables en definitiva deberán ser utilizadas. Esta decisión dependerá entre otras de la confiabilidad de la variable, su disponibilidad y eventualmente la atingencia. En los modelos de regresión intentados (Logit y Probit) se corrieron versiones de las regresiones con errores estándar robustos, esto es necesario puesto que muchos de los problemas que presentan los modelos de regresión impactan este error estándar que es necesario para calificar la importancia estadística de las variables. Usando estos errores robustos se observó que todas las variables mantienen su relevancia en ambos modelos intentados (logit y probit). Por otro lado, el uso de modelos de árboles de clasificación resulta ser mucho más conciso en el uso de variables para la obtención de una precisión comparable. Surge entonces la pregunta ¿Qué se quiere privilegiar? ¿Algunos pocos puntos porcentuales de precisión o economía descriptiva del fenómeno? Estas son preguntas para las cuales se deberá discutir y alcanzar consensos para decidir un camino a seguir en el futuro.

Lo que habría que considerar en una futura implementación a nivel de producción de cualquiera de los modelos son los requerimientos computacionales. En el presente reporte se han utilizado el paquete estadístico R y el software GRET. Si bien R puede hacer esencialmente lo mismo que puede hacer GRET, para la parte de regresiones se prefirió utilizar GRET puesto que este es más amistoso que R.

En el presente reporte, se ha demostrado sin dejar lugar a dudas, que es posible utilizar herramientas de aprendizaje supervisado para predecir la presentación de licencias con una tasa bastante aceptable de precisión y esto motiva su posterior

desarrollo con un nivel más acabado de detalle y considerando los aspectos de implementación en un ambiente de producción cuando la oportunidad así se presente.

Nuevos desarrollos basados en el uso de un nuevo set de datos trabajado por la SUSESO, otorgó resultados excelentes para identificar correctamente licencias rechazadas que se reclaman, desafortunadamente, esta precisión viene con un costo asociado que se traduce en un modelo conservador que tiende a sugerir más reclamos de los que hay, produciendo esto en particular más licencias rechazadas que no reclaman identificadas como que si lo harán (del orden del 50%). Esto es claramente un costo demasiado alto que será de una u otra manera potencialmente transferido a los usuarios finales por las aseguradoras (que deben proveer la información para confeccionar los expedientes, incrementando así sus costos de operación). Para solucionar este sobreajuste, se utilizaron métodos de conjunto (*ensemble*) que si bien disminuyen la precisión en la clasificación de las licencias rechazadas que reclaman, aumentan la precisión global del clasificador disminuyendo los errores en las licencias médicas rechazadas que no reclaman, previniendo así potenciales problemas futuros. Se recomienda que este sea el camino para seguir a futuro para crear modelos más robustos y estables en el tiempo.

## 3 Asignación Inteligente de Casos

### 3.1 Introducción

Una vez que se ha ingresado un reclamo a la SUSESO por medio de los formularios web o papel disponibles para ello, se inicia el accionamiento de una serie de tareas que son parte del Workflow necesario para poder resolver el reclamo. La cantidad de tiempo que toma el proceso completo depende entre otras cosas del tiempo que demora el expediente en constituirse, quien es asignado para resolver el caso y el tiempo que demora este profesional en preparar una resolución para su posterior visado y comunicación.

Es claro que, si la asignación es incorrecta, entonces el profesional escogido tomara más tiempo en producir un dictamen. Por otro lado, si sólo se decidiera asignar un reclamo al profesional que más eficientemente lo podría resolver, se podría llegar a situaciones en que algunos profesionales están saturados y otros con poca carga.

El objetivo de un modelo de asignación debiese ser entonces el de hacer el mejor uso posible de los recursos disponibles, de tal manera de poder minimizar el tiempo de resolución de un reclamo ingresado al sistema. Un punto no menor para considerar acá es las posibilidades que existen en la búsqueda de soluciones para este problema, existen a priori las siguientes aproximaciones al problema:

- Heurísticas Golosas
- Modelo de Programación Entera Mixta
- Metaheurísticas
- Programación Estocástica
- Simulación
- Métodos híbridos que combinan algunas de las anteriores

No es la intención entrar a describir en detalle cada una de las posibilidades, más bien el objetivo es dar a entender que existe una variada y amplia gama de opciones de modelamiento para tratar este problema. Es claro que, en una solución final estas opciones debiesen haber sido consideradas puesto que ellas influyen el tipo de solución que se puede obtener<sup>5</sup>.

Un último punto por destacar guarda relación con la naturaleza de la solución deseada y la definición última del problema. No es lo mismo entregar la mejor solución posible con los recursos disponibles (que podrían estar mal dimensionados) que hacer la pregunta relativa al tipo de recursos profesionales que se necesitan para poder resolver los casos que se estima ocurrirán. A grandes rasgos, un problema parece ser el dual del otro, esto en el sentido de que una formulación trata de ver la mejor forma de utilizar eficientemente el recurso profesional disponible mientras que la otra trata

---

<sup>5</sup> Este es un punto no menor puesto que no es lo mismo tener una solución óptima a un problema que tener una solución aproximada. Tampoco son menores los problemas asociados a los datos, si hay datos con incertidumbre considerable, algunas de las técnicas mencionadas anteriormente proveerán una mala solución al problema, por el contrario, cuando los datos son ciertos y fácilmente medibles no tiene sentido utilizar algunos de los métodos.

de determinar la forma más económica de proveer un determinado nivel de servicio. Adicionalmente, no es lo mismo resolver el problema de manera local que resolverlo de manera global, usualmente las soluciones locales son subóptimas en un contexto global, lo que puede llevar a pensar que un determinado método es bueno cuando en realidad no lo es.

Este reporte comienza con una revisión de un modelo actualmente propuesto para la asignación de reclamos a profesionales. Después continúa con una crítica a este modelo y en particular muestra con algunos contraejemplos posibles problemas que esta heurística presenta. Finalmente, este reporte discute las características que un modelo de asignación debiese tener, en particular en referencia a los datos disponibles y el tipo de pronósticos de demanda que pueden formarse a partir de la data existente. El reporte concluye con conclusiones en donde se destacan los puntos más importantes de este estudio.

Es importante notar que la creación de un modelo para resolver este problema escapa al contexto en que se enmarca la presente consultoría. Por este motivo se discutirán opciones de modelamiento que de todas maneras deben ser probadas a futuro para poder decidir cuál de ellas es más efectiva para este importante problema presente en el workflow de reclamos presentados a la SUSESO.

### 3.2 Descripción de la Heurística Existente para el Problema

Actualmente hay una heurística que ha sido propuesta a la SUSESO para atacar el problema de asignación de reclamos a recursos profesionales para su dictaminación. El algoritmo propuesto puede ser visto en la Figura 7. A primera vista no es difícil notar que el algoritmo esencialmente lo que hace es chequear una serie de condiciones de factibilidad tales como disponibilidad de profesionales, si el reclamo es reconsideración o no, etc. De esta manera se filtra la base de datos de profesionales existentes (o se buscan externos) que podrían ser compatibles con el caso en cuestión. Muy importante en este paso es la categorización del reclamo en una de tres categorías correspondiente cada una a un nivel de complejidad distinto: baja, media y alta. Una vez que se han considerado todas las características del caso y se ha determinado el subconjunto de profesionales que podrían responder a este caso, se presentan estos profesionales en un listado ordenado desde el cual se escoge tomando en consideración la carga actual de todos los profesionales listados en este subconjunto.

Los datos que llaman la atención y que son mencionados de alguna forma en esta heurística son los siguientes:

- El tipo de especialidad requerida por un reclamo
- El tipo de complejidad que caracteriza a un reclamo
- El tipo de eficacia de un profesional para resolver un cierto tipo de reclamo
- La productividad de un profesional



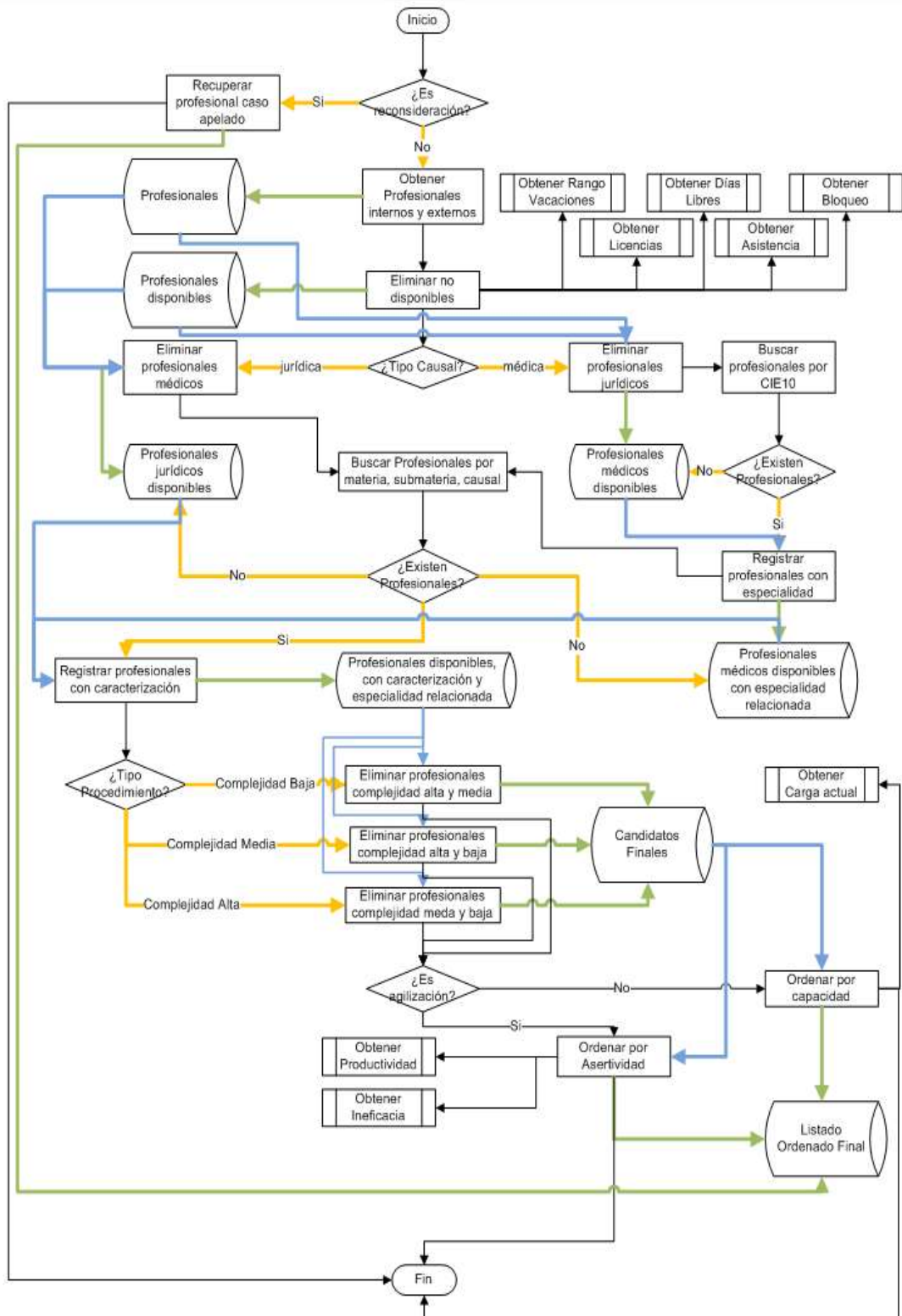


Figura 7: Diagrama de Flujo del Modelo Actualmente Propuesto

### 3.3 Críticas a la Heurística Existente

En esta sección se detallan los potenciales problemas que se pueden evidenciar del diagrama de flujo anterior. Algunos de estos puntos son propuestos para mayor discusión, otros para potencialmente influir en el desarrollo del sistema con miras a capturar información que podría resultar relevante para desarrollos posteriores.

#### 3.3.1 Medición de la Complejidad de un Reclamo

No es claro a partir del análisis del diagrama de flujo el cómo se determina la complejidad de un reclamo. *En ausencia de más información, se ha asumido que la clasificación de la complejidad de un caso es un procedimiento manual ingresado al workflow mediante la acción de un funcionario.*

Para partir hay que mencionar que hay varios factores que influyen en la complejidad de un caso, pero en esencia se podría decir que un caso es complejo en la medida de que cuesta más decidir sobre él. Esto pasa por dos factores esencialmente, primero el tiempo requerido para resolver, segundo es el nivel de entrenamiento/experiencia del profesional médico que debe resolver, eventualmente algunos casos requieren un panel de expertos lo que los hace bastante complejos. De cualquier forma, la clasificación de la complejidad de un caso debiese ser idealmente una decisión libre de emociones.

La forma adecuada de aproximar la complejidad de un caso sería a través de la cantidad de tiempo que demora en ser resuelto. Desafortunadamente no existe información en el workflow (actual) que indique la cantidad de tiempo efectiva utilizada en resolver un reclamo. Esta cantidad podría aproximarse por el tiempo que transcurre entre la asignación hecha al profesional y su resolución, sin embargo, este número se prevé contaminado por la sobrecarga laboral, la espera por información adicional y otros factores.

#### 3.3.2 Determinación de la Especialidad Requerida por un Reclamo

El código CIE10 debería indicar lo que se necesita, idealmente esto podría utilizarse automáticamente, sin embargo, la entidad aseguradora es quien coloca este código en la actualidad y hay evidencia que hacen dudar de la calidad de dicho código. Al mirar la información es fácil darse cuenta de que la familia CIE10 habitualmente está presente en la base de datos, sin embargo, la subfamilia ya es menos confiable y finalmente el dígito que es parte del código es muy poco confiable ya que muy pocas veces aparece.

Siendo ésta una característica importante no sólo para el modelo de asignación, sino que también para el modelo de sugerencia de dictamen, es de vital importancia que este código sea introducido en la base de datos con la mayor precisión posible. Si esto no es posible, un triage podría ser utilizado para clasificar cada caso a la entrada, lamentablemente esta opción encarece el workflow tanto en términos de tiempo como de recursos involucrados.

### 3.3.3 Determinación de la Eficacia de un Profesional

La eficacia de un profesional tiene que ver principalmente con la rapidez con la cual este profesional puede efectivamente resolver un caso, pero además también tiene que ver con la asertividad con la cual se obtiene este resultado. Por ejemplo, siempre es posible lanzar una moneda al aire y jugar al cara y sello para decidir el resultado de un caso, esto es extremadamente rápido, sin embargo, no es asertivo y se espera que la tasa de error sea alta.

Hay ciertos problemas asociados a esta información requerida por el modelo puesto que distintas estrategias de premios o bonificación por desempeño pueden cargar la balanza a uno u otro extremo. Si se bonifica la asertividad, entonces los profesionales preferirán tomar más tiempo para resolver un caso de tal forma de evitar futuros reclamos posteriores asociados al mismo caso. Si se bonifica la rapidez, entonces los profesionales podrían caer en situaciones donde los casos son resueltos rápidamente para efectos de poder mostrar indicadores abultados.

La eficacia eventualmente corresponde a una definición mucho más compleja que la que hemos discutido hasta el momento, y por lo tanto hay que tener mucho cuidado con la definición final que se utilice puesto que muy probablemente esta definición estará asociada a los incentivos futuros. Lamentablemente no hay una respuesta clara desde la literatura para poder medir eficacia en el contexto del problema en cuestión y deja la pregunta abierta para el futuro.

### 3.3.4 La Heurística es Golosa

El principal problema de la heurística existente es que ella es esencialmente golosa. Lo que ello significa es que la decisión que toma el modelo es caso a caso y esto no toma en consideración el impacto que esta decisión local tiene en el futuro. A modo de ejemplo, supongamos que tengo tres profesionales llamados A, B y C y que tengo sólo dos tipos de enfermedad: tipo 1 y tipo 2. Supongamos además que los médicos tienen las siguientes características:

Tabla 1: Ejemplo ficticio, factores en uso

Profesional	Carga Actual	Rapidez	Eficiencia 1	Eficiencia 2
A	4/1	0.3	0.8	0.5
B	2/6	0.5	0.3	0.5
C	0/3	0.4	0.1	1.3

Donde la carga actual indica el número de casos a cargo del profesional (no hemos incorporado en este ejemplo ficticio el nivel de complejidad de los casos), la rapidez es el número de casos que el profesional puede sacar por día de trabajo, por ejemplo 0.3 significa que, para terminar 5 casos, el profesional A necesita de 16.6 días aproximadamente. Finalmente, eficiencia 1 y eficiencia 2 son modificadores de rapidez, por ejemplo, en el mismo caso anterior, si la enfermedad o condición es de

tipo 1, el profesional A tomará  $16.6/0.8 = 20.75$  días para terminar esa carga de trabajo.

Supongamos además que el precio de contratar un profesional externo es de un 20% más caro que un profesional de planta y que en teoría en ningún momento del tiempo se puede tener una carga de trabajo más grande que la que la rapidez determina.

### 3.3.4.1 Ejemplo 1

Consideremos ahora el siguiente escenario de llegada de casos en que 1 significa enfermedad tipo 1 y 2 significa enfermedad tipo 2 y apliquemos la heurística. El resultado se resume en la siguiente tabla (tiempos de llegada entre reclamos exponencialmente distribuido y con media de 1.5 días):

Tabla 2: Ejemplo de heurística para un caso hipotético

DT	Seq1	xA	xB	xC	NA1	NB1	NC1	NA2	NB2	NC2	TAA	TAB	TAC
					4	2	0	1	6	3	23.33333	37.33333	5.769231
0.459394	2			1	0	0	0	0	0	1	22.87394	36.87394	7.232913
0.090461	1	1			1	0	0	0	0	0	26.95014	36.78348	7.142452
0.048176	1	1			1	0	0	0	0	0	31.06863	36.7353	7.094275
0.972033	1	1			1	0	0	0	0	0	34.26327	35.76327	6.122242
0.299665	2			1	0	0	0	0	0	1	33.9636	35.4636	7.745654
0.09463	1	1			1	0	0	0	0	0	38.03564	35.36897	7.651024
0.264043	2			1	0	0	0	0	0	1	37.7716	35.10493	9.310058
1.365664	2			1	0	0	0	0	0	1	36.40593	33.73927	9.86747
0.517207	1		1		0	1	0	0	0	0	35.88873	39.88873	9.350264
0.026735	1	1			1	0	0	0	0	0	40.02866	39.86199	9.323528
0.022842	2			1	0	0	0	0	0	1	40.00581	39.83915	11.22376
0.180033	1		1		0	1	0	0	0	0	39.82578	46.32578	11.04373
0.130173	1	1			1	0	0	0	0	0	43.86228	46.19561	10.91356
0.489146	2			1	0	0	0	0	0	1	43.37313	45.70646	12.34749
0.071184	1	1			1	0	0	0	0	0	47.46861	45.63528	12.2763
0.316288	1		1		0	1	0	0	0	0	47.15232	51.98566	11.96002
0.304166	2			1	0	0	0	0	0	1	46.84816	51.68149	13.57893
0.560697	2			1	0	0	0	0	0	1	46.28746	51.12079	14.94131
											37.653	41.12669	9.731274

La columna DT corresponde al tiempo transcurrido desde el último reclamo, por ejemplo, la primera fila indicando 0.459394 indica que dicho reclamo llega 0.459394 días después del último reclamo ingresado (acá se partió desde un tiempo  $t=0$  y los tiempos entre llegadas han sido asumidos exponencialmente distribuidos con media 1.5 días). La columna Seq1 indica el tipo de caso que llega (tipo 1 o tipo 2) y este ha sido generado al azar. Las columnas xA, xB y xC corresponden a la variable (binaria de asignación) e indican a que profesional se le asignarán (es a través de esta decisión que se simula la heurística). Las columnas NA1, NB1, NC1, NA2, NB2 y NC2 son columnas auxiliares que "traducen" la asignación hecha por las variables xA, xB y xC

con respecto a los casos tipo 1 o 2 para cada uno de los profesionales. Finalmente, las variables TAA, TAB y TAC llevan cuenta del tiempo utilizado para cada uno de los profesionales.

Para el caso de la segunda fila, con tiempo de llegada 0.459394, este es asignado al profesional C, este profesional al tener una utilización de tiempo igual al tiempo que ya tenía utilizado que corresponde a 5.769231, menos el tiempo transcurrido que asciende a 0.459394, lo que entrega 5.309837. Sin embargo, como el profesional ha sido asignado un caso tipo 2, entonces el tiempo debe actualizarse para considerar esta carga de trabajo, en este caso, la actualización se hace usando los factores descritos en la Tabla 1:

$$1/0.4/1.3 = 1.92307 \text{ (Para clarificar, esto ha sido calculado como } \frac{1}{\text{Rapidez/Eficiencia}} \text{)}$$

Lo que sumado el tiempo anteriormente calculado entrega 7.232907. Se procedió de esta misma manera para todas las llegadas. La heurística se aplicó viendo idoneidad y asignado a quien tuviese más tiempo disponible en función de su idoneidad.

Finalmente, en cada columna se puede observar la utilización promedio, que en este ejemplo alcanzan a 37.653 para el profesional A, 41.12669 para el profesional B y 9.731274 para el profesional C. Si se suman todos estos promedios, el total de tiempo utilizado en promedio para los tres profesionales es de 88.51096.

#### 3.3.4.2 Ejemplo 2

Así como en el ejemplo anterior, pero ahora considerando un nuevo escenario aleatorio de llegadas de casos, obtenemos lo contenido en la Tabla 3. Es necesario observar que los tiempos de llegada se han mantenido iguales y lo único que ha sido modificado es la secuencia de llegadas de casos al centro de reclamos. Los cálculos han sido hechos de la misma manera anteriormente descrita y la heurística se aplicó al igual que en el ejemplo anterior. La métrica de desempeño promedio global del sistema alcanza a 88.61555 unidades.

Hay que resaltar que estos dos ejemplos fueron confeccionados para poder dar a entender la naturaleza del problema de manera simple. Claramente es una situación idealizada y por ningún motivo constituye una prueba exhaustiva de la calidad de la heurística actual. La idea es entonces construir un modelo de optimización simple, pero que en la práctica pueda mostrar que una heurística del tipo golosa no provee una optimización global rigurosa.

La idea detrás del modelo de optimización será el minimizar la métrica de desempeño promedio global, puesto que esta se considera una medida razonable del tiempo puesto en resolver los casos, y por consiguiente si es que se minimiza esta cantidad, se está minimizando implícitamente el tiempo utilizado en cada caso (esto es un supuesto y claramente no es el objetivo de esta consultoría el tratar de encontrar el mejor modelo, solamente dentro del proceso de revisión se quiere mostrar evidencia

de que el tipo de heurística actual es mejorable y por lo tanto algo que puede hacerse en un futuro estudio).

Tabla 3: Nuevo ejemplo, secuencia aleatoria distinta, mismos tiempos de llegada

DT	Seq1	xA	xB	xC	NA1	NB1	NC1	NA2	NB2	NC2	TAA	TAB	TAC
					4	2	0	1	6	3	23.33333	37.33333	5.769231
0.459394	1	1			1	0	0	0	0	0	27.04061	36.87394	5.309836
0.090461	1	1			1	0	0	0	0	0	31.11681	36.78348	5.219375
0.048176	1	1			1	0	0	0	0	0	35.2353	36.7353	5.171199
0.972033	2			1	0	0	0	0	0	1	34.26327	35.76327	6.122242
0.299665	2			1	0	0	0	0	0	1	33.9636	35.4636	7.745654
0.09463	2			1	0	0	0	0	0	1	33.86897	35.36897	9.5741
0.264043	1	1			1	0	0	0	0	0	37.7716	35.10493	9.310058
1.365664	1		1		0	1	0	0	0	0	36.40593	40.40593	7.944393
0.517207	1	1			1	0	0	0	0	0	40.05539	39.88873	7.427187
0.026735	2			1	0	0	0	0	0	1	40.02866	39.86199	9.323528
0.022842	2			1	0	0	0	0	0	1	40.00581	39.83915	11.22376
0.180033	1		1		0	1	0	0	0	0	39.82578	46.32578	11.04373
0.130173	1	1			1	0	0	0	0	0	43.86228	46.19561	10.91356
0.489146	2			1	0	0	0	0	0	1	43.37313	45.70646	12.34749
0.071184	1	1			1	0	0	0	0	0	47.46861	45.63528	12.2763
0.316288	2			1	0	0	0	0	0	1	47.15232	45.31899	13.88309
0.304166	2			1	0	0	0	0	0	1	46.84816	45.01482	15.502
0.560697	1		1		0	1	0	0	0	0	46.28746	51.12079	14.94131
											38.3109	40.77581	9.528845

### 3.3.4.3 Un modelo de optimización muy simple

El modelo de optimización sugerido es el siguiente:

$$\begin{aligned}
 & \min \mu_A + \mu_B + \mu_C \\
 & \text{s. a.} \\
 & x_{At} + x_{Bt} + x_{Ct} \leq 1 \quad \forall t \in T \\
 & T_{Ait} = T_{Ait-1} + \sum_i t_i(x_{At}, x_{Bt}, x_{Ct}) \quad \forall i, \forall t \in T \\
 & T_{Ait} \leq K \quad \forall i \in \{A, B, C\} \\
 & x_{it} \in \{0, 1\} \quad \forall i \in \{A, B, C\}
 \end{aligned}$$

La función objetivo corresponde a la métrica de desempeño global utilizada anteriormente (la utilización total promedio de los profesionales), claramente no es posible asignar un caso a más de un profesional (restricción 2), la tercera restricción ajusta el tiempo utilizado en función de la asignación (de la misma manera que describimos anteriormente y que hace que el problema sea no lineal), la cuarta restricción establece una cota superior para la utilización de un profesional y finalmente las variables de decisión son binarias. El modelo se implementó en EXCEL

y se utilizó SOLVER para resolverlo. Al ser el tamaño pequeño no hubo problemas de licencias. Se espera que en un caso realista se deba contar con una herramienta de optimización más idónea.

### 3.3.4.3.1 Primer ejemplo con modelo de optimización

*Los resultados se presentan en la*

Tabla 4

Tabla 4: Optimización para ejemplo 1

DT	Seq1	xA	xB	xC	NA1	NB1	NC1	NA2	NB2	NC2	TAA	TAB	TAC
					4	2	0	1	6	3	23.33333	37.33333	5.769231
0.459394	2			1	0	0	0	0	0	1	22.87394	36.87394	7.232913
0.090461	1	1			1	0	0	0	0	0	26.95014	36.78348	7.142452
0.048176	1	1			1	0	0	0	0	0	31.06863	36.7353	7.094275
0.972033	1	1			1	0	0	0	0	0	34.26327	35.76327	6.122242
0.299665	2			1	0	0	0	0	0	1	33.9636	35.4636	7.745654
0.09463	1	1			1	0	0	0	0	0	38.03564	35.36897	7.651024
0.264043	2			1	0	0	0	0	0	1	37.7716	35.10493	9.310058
1.365664	2			1	0	0	0	0	0	1	36.40593	33.73927	9.86747
0.517207	1	1			1	0	0	0	0	0	40.05539	33.22206	9.350264
0.026735	1	1			1	0	0	0	0	0	44.19532	33.19532	9.323528
0.022842	2			1	0	0	0	0	0	1	44.17248	33.17248	11.22376
0.180033	1	1			1	0	0	0	0	0	48.15911	32.99245	11.04373
0.130173	1	1			1	0	0	0	0	0	52.19561	32.86228	10.91356
0.489146	2			1	0	0	0	0	0	1	51.70646	32.37313	12.34749
0.071184	1	1			1	0	0	0	0	0	55.80195	32.30195	12.2763
0.316288	1	1			1	0	0	0	0	0	59.65232	31.98566	11.96002
0.304166	2			1	0	0	0	0	0	1	59.34816	31.68149	13.57893
0.560697	2			1	0	0	0	0	0	1	58.78746	31.12079	14.94131
											42.03897	34.10914	9.731274

El desempeño global para este caso es de 85.87938 lo que es una mejora del orden de un 3% con respecto al caso base.

### 3.3.4.3.2 Segundo ejemplo con modelo de optimización

Los resultados se presentan en la Tabla 5.

Tabla 5: Modelo de optimización para el segundo ejemplo

DT	Seq1	xA	xB	xC	NA1	NB1	NC1	NA2	NB2	NC2	TAA	TAB	TAC
					4	2	0	1	6	3	23.33333	37.33333	5.769231
0.459394	1	1			1	0	0	0	0	0	27.04061	36.87394	5.309836
0.090461	1	1			1	0	0	0	0	0	31.11681	36.78348	5.219375
0.048176	1	1			1	0	0	0	0	0	35.2353	36.7353	5.171199
0.972033	2			1	0	0	0	0	0	1	34.26327	35.76327	6.122242
0.299665	2			1	0	0	0	0	0	1	33.9636	35.4636	7.745654
0.09463	2			1	0	0	0	0	0	1	33.86897	35.36897	9.5741
0.264043	1	1			1	0	0	0	0	0	37.7716	35.10493	9.310058
1.365664	1	1			1	0	0	0	0	0	40.5726	33.73927	7.944393
0.517207	1	1			1	0	0	0	0	0	44.22206	33.22206	7.427187
0.026735	2			1	0	0	0	0	0	1	44.19532	33.19532	9.323528
0.022842	2			1	0	0	0	0	0	1	44.17248	33.17248	11.22376
0.180033	1	1			1	0	0	0	0	0	48.15911	32.99245	11.04373
0.130173	1	1			1	0	0	0	0	0	52.19561	32.86228	10.91356
0.489146	2			1	0	0	0	0	0	1	51.70646	32.37313	12.34749
0.071184	1	1			1	0	0	0	0	0	55.80195	32.30195	12.2763
0.316288	2			1	0	0	0	0	0	1	55.48566	31.98566	13.88309
0.304166	2			1	0	0	0	0	0	1	55.18149	31.68149	15.502
0.560697	1	1			1	0	0	0	0	0	58.78746	31.12079	14.94131
											42.47756	34.10914	9.528845

El desempeño global para este caso es de 86.11555 lo que es una mejora del orden de un 3% con respecto al caso base.

### 3.3.5 Algunos comentarios adicionales

En el contexto del caso artificial analizado previamente, es posible ver que un modelo de esa naturaleza puede ser utilizado para estimar falencias en el servicio de atención de casos por parte de profesionales.

Es sabido que un funcionario no puede trabajar más que una cierta cantidad de horas en un determinado mes, si se genera un perfil de demanda compatible con la historia para un mes dado, y en presencia de la información actual de la capacidad resolutive (horas hombres con perfiles específicos), al incorporar una restricción de capacidad en el modelo, si este produce infactibilidad entonces eso se puede interpretar como una posibilidad más que un problema, esto pues si se agregan posibles funcionarios al listado (externos) la solución de la nueva instancia del modelo puede mostrar una nueva solución en que se obtenga factibilidad en términos de no asignar más casos que la cantidad de horas disponibles para ese mes.

Es claro que una decisión de esta naturaleza, que eventualmente involucra muchas variables de decisión escapa al dominio de lo razonable para un ser humano y que rápidamente se vuelve compleja. Lo otro importante es que la heurística actual no incorpora costos de operación del sistema lo que en efecto puede cambiar la decisión.



### 3.4 Características Deseables de un Modelo de Asignación

#### 3.4.1 Función Objetivo Razonable

La función objetivo debe medir la calidad de la solución, sin embargo, ésta debe considerar la naturaleza de los datos, el problema y lo que se persigue. En el caso particular de la heurística existente, hay un alto riesgo de generación de óptimos locales puesto que la heurística actual no considera pronósticos de demanda y por lo tanto tiene una visión acotada. Si se cambia el objetivo por uno de utilizar los recursos eficientemente en una cierta ventana de tiempo de tal forma de minimizar los tiempos de resolución

#### 3.4.2 Debiese considerar el costo de operar el sistema

Este costo no es evidente en la heurística existente. Se desliza de alguna manera la idea puesto que está la posibilidad de contratar externos, sin embargo, se estima que esto tiene un costo por hora más caro que tener profesionales de planta. Lo otro es que un modelo de asignación debiese considerar el costo de reprocesamiento de casos debido a malas decisiones (lo que en la heurística actual se denomina asertividad).

#### 3.4.3 Debiese ayudar a determinar el dimensionamiento requerido para entregar el servicio

Una vez que se tiene un modelo para el cual los recursos disponibles se conocen, se puede en dicho caso ver cuál es la mejor manera de responder a los requerimientos (estimados). En función de la respuesta del modelo, este podría considerar variables de decisión que permitan considerar la contratación de personal adicional o la subcontratación de externos. En cualquier caso, para dichas contrataciones no es posible contar con un historial preciso de habilidades, por lo que se cree que un buen modelo podría además entregar dichos requerimientos para poder ir a buscar el recurso humano adecuado (algo así como se necesita un profesional cuyo perfil esté alineado con una lista de dolencias)

#### 3.4.4 Debiese incorporar demanda futura esperada en la decisión

Como se ha mencionado antes, si bien los casos llegan discretamente al sistema, el modelo de decisión no puede perder de vista el resultado global de la operación del sistema. Uno de los problemas tradicionales de las heurísticas golosas es que ellas optimizan localmente y pierden de vista el global. En el caso del problema en cuestión, el global tiene que ver con el uso del modelo para poder planificar el uso de recurso en presencia de requerimientos que entran al sistema, pero el ideal sería el poder hacer esto sin perder de vista lo que se espera ocurra en tiempos a venir. Por ejemplo, para un día dado se podrían tener 10 ingresos de reclamos por rechazos de licencias por depresión y la heurística eventualmente podría tratar de copar el recurso que es considerado más eficiente en el diagnóstico de dichos casos, sin embargo, si al mismo tiempo se percibe que hay profesionales que a pesar de no ser tan eficientes pueden invertir tiempo en algunos de estos casos, esta decisión de asignación permite no sobrecargar parte del recurso especializado y dejarlo así disponible para enfrentar un

alza de demanda proyectada en las semanas próximas. Ciertamente esto no es posible con todas las dolencias, los casos traumatológicos probablemente deban ser asignados a un traumatólogo a menos claro de que el caso no sea tan complejo y por lo tanto pueda ser abordado por cualquier profesional con formación general que le permita decidir (tal vez no tan rápidamente como lo haría un traumatólogo).

#### 3.4.5 Debiese incorporar el concepto de nivel de servicio

Lo que la heurística actual no hace para nada es el considerar la incertidumbre de la demanda futura. En presencia de incertidumbre, las decisiones tomadas hoy tienen repercusiones en el futuro, pero con cierta probabilidad. No hay forma de asegurar lo que ocurrirá en el futuro, sin embargo, se puede estimar que tan probable es que algo pase dado los datos históricos. Una forma alternativa de ver el problema de optimización que hemos mencionado a lo largo de este reporte consiste en considerar el nivel de servicio más que la contabilidad exacta del tiempo que toma resolver un caso. Desde este punto de vista, el modelo podría ser capaz de evaluar para instancias de un futuro que aún no ocurre pero que han sido generadas basadas en información histórica y predicción de demanda, lo que ocurriría con la dotación inicial y por lo tanto calcular a partir de dicha colección de instancias el desempeño, para así determinar la probabilidad de que o bien ciertos recursos serán sobre-utilizados, o bien los tiempos de respuesta para ciertas enfermedades serán en promedio mayores a las deseadas y a partir de ahí actuar acordemente para modificar la disponibilidad de recurso humano para evaluación. Esta opción es implementable cuando se tiene un modelo de optimización para el caso determinístico y a partir de ahí se puede probar en escenarios equiprobables para generar soluciones robustas al problema.

### 3.5 Conclusiones

En este capítulo se analizó a fondo una heurística propuesta para asignar casos al personal existente. Entre otras cosas se ha identificado que la heurística es golosa, pero no sólo eso, sino que además hay problemas con la obtención de información esencial para la heurística, a saber: la eficacia y asertividad de un evaluador, la productividad de un profesional y finalmente la determinación de la especialidad requerida por un reclamo, que por su parte incide en la complejidad, puesto que hay falencias en la introducción de los códigos CIE10 por parte de las entidades aseguradoras.

Es interesante concluir del ejercicio de análisis de la heurística existente que, si bien esta es golosa en esencia, y que por lo mismo generará soluciones subóptimas para el problema en consideración, en los casos artificiales en que se probó su desempeño no es malo cuando se compara con un modelo de optimización siendo la diferencia del orden de un 3%. Cabe destacar que el ejemplo que se probó en una sobresimplificación de la realidad y que este desempeño no necesariamente se trasladaría al caso real. No obstante, se concluye que la heurística propuesta es un buen punto de partida y que sólo se puede mejorar a futuro partiendo de estos cimientos.

Finalmente, se entrega un listado de condiciones que según la opinión del autor son deseables en un algoritmo de asignación de casos. Un algoritmo de esta naturaleza podrá ser desarrollado siempre y cuando se pueda contar con datos apropiados y cuando se tenga una mejor noción del proceso de atención, pero visto desde la perspectiva de un negocio. Es importante notar que se observan buenas posibilidades de mejorar el modelo existente como parte de un proceso de mejora continua en el futuro, esto de la mano de una mejor recolección de información podría generar buenos ahorros de tiempos en el proceso y una buena planificación de recursos acordes con la realidad del sistema.

## 4 Propuesta Automatizada de Resolución de Casos

### 4.1 Introducción

Una vez que el caso ha sido asignado a un profesional que lo estudiará, es importante poder proveer asistencia a este profesional para efectos de poder proponer un diagnóstico de manera eficaz. Con este propósito es que se ha planteado este estudio en que se pretende analizar estrategias basadas en datos existentes para poder ver si es posible configurar un “asistente inteligente” que permita ayudar al profesional a tomar una decisión.

La racionalidad detrás de la idea es la siguiente: si en el pasado se ha resuelto consistentemente de una manera para casos con caracterización similar, nada hace sospechar que esto no debiese seguir siendo así y por lo tanto sobre la base de dicha regularidad es posible aliviar la carga del profesional por la vía de hacer esta regularidad explícita.

Desde este punto de vista es que se propone un método que básicamente trata de discriminar los posibles casos correspondientes a una cierta sintomatología (tal y como está descrita utilizando la nomenclatura CIE10) y así ver basado en otros atributos adicionales que tipo de resolución es más probable. Es menester notar que los casos se caracterizan por varios atributos adicionales a la sintomatología, sin embargo, dicha sintomatología es un primer filtro necesario que aplicar para poder así sólo acceder a los casos relacionados al caso en cuestión sobre el cual se desea pronunciamiento.

La idea principal de este reporte es poder mostrar los experimentos que se han llevado a cabo para poder estudiar el tipo de algoritmos/métodos que podrían utilizarse en la sugerencia al profesional de un dictamen para resolver el caso.

Finalmente, es necesario hacer notar que este estudio separará los tipos de licencias en dos grandes grupos: origen común y maternal. El motivo de esta decisión se basa en que las licencias maternas presentan una problemática distinta a las licencias comunes por cuanto el diagnóstico no corresponde a quien presenta la licencia, sino que a su hijo(a), por lo tanto, fue acordado con la contraparte técnica de la SUSESO que sería preferible evitar potenciales confusiones y por lo tanto percibido como más adecuado el realizar esta separación.

### 4.2 Breve Descripción de los Datos a Utilizar

La sábana de datos provista para esta etapa del estudio es una sábana consistente en registros provenientes de dos tablas que han unidas mediante el uso de una clave común a ambas tablas (en el más puro estilo de una consulta SQL utilizando cláusulas JOIN). Las tablas que fueron unidas corresponden a las sábanas SILMSIL y expedientes\_workflow en donde sólo los campos más relevantes para el análisis (tal y como ha sido definido por la SUSESO) de ambas tablas fueron incluidos. Adicionalmente se agregaron los campos modificada, reclamada, resultado y

reconsideración que son indicadores de si el respectivo folio ha sufrido modificaciones, ha sido reclamada a la SUSESO, el resultado de dicho reclamo (acepta, rechaza o sin pronunciamiento) y finalmente si el folio corresponde a una reconsideración.

En la sábana de datos hay varios campos, varios de ellos ya utilizados previamente en los otros modelamientos (ver Etapa 1). Entre los campos disponibles en la sábana de datos, uno de los más importantes para efectos del presente estudio corresponde al campo `codigo_diagnostico` en el cual se documenta el código CIE10 para la sintomatología presentada por el paciente. El código CIE10 utilizado tiene el siguiente formato:

$$\alpha N. M$$

donde:

- $\alpha$  corresponde a una letra que indica la familia
- $N$  corresponde a un dígito que indica la subfamilia
- $M$  corresponde a un dígito que indica el diagnóstico

A modo de ejemplo, F41.1 significa<sup>6</sup> trastorno de ansiedad generalizada, siendo F trastornos mentales y del comportamiento, 41 correspondiente a otros trastornos de ansiedad y el 1 correspondiente a generalizada. El código F41.9 por otro lado significa trastorno de ansiedad, no especificado. El dígito de diagnóstico no es consistente a lo largo de la tabla, y las subfamilias obviamente dependerán de cuál sea la familia. Para efectos de poder trabajar con este dato en particular, se propone la creación de tres nuevas columnas, una correspondiente a la familia, otra correspondiente a la subfamilia y finalmente una correspondiente al diagnóstico.

Una vez hecho este "desacoplamiento" del código CIE10, ha sido posible observar las estadísticas básicas asociadas a la presencia de este código en la sábana:

- 160.361 licencias presentan el campo familia
- 160.288 licencias presentan el campo subfamilia
- 74.991 licencias presentan el campo diagnóstico

Ha sido acordado con la contraparte técnica de la SUSESO que aquellos registros que no contengan un código de diagnóstico serán eliminados del análisis puesto que no aportan en el desarrollo de este. También es razonable ver que muy probablemente el código correspondiente al diagnóstico es de poco uso al parecer.

Para poder separar los datos, se codificaron las variables adicionales `origen_comun` y `maternal` que son variables indicadoras cuyo valor es 1 en caso de ser una licencia de origen común (tipo de licencia igual a 1 o 7) o maternal (tipo de licencia igual a 3 o 4), en cualquier otro caso estas variables valen cero. Los datos asociados a licencia de origen común corresponden a 156.051 observaciones y las restantes 6.845 corresponden a las denominadas licencias maternas.

---

<sup>6</sup> Se ha utilizado un documento encontrado en la web (<http://cie10.org/Descargas/CIE10-ES%20MAS.pdf>) donde se provee una tabla exhaustiva con todas las posibles descripciones. No se ha provisto como parte de la documentación la subfamilias y diagnósticos utilizados por la SUSESO, por lo demás no es información relevante al momento de correr los algoritmos y la idea de buscar este documento fue más que nada el demostrar la forma en que se codifican las enfermedades.

Lo siguiente que se procedió a hacer fue eliminar los registros con NA (auto-codificado por R), con esto los registros por tipo de licencia corresponden a 153.232 licencias de origen común y 6795 licencias maternas, es decir un total de 160.027 registros (98.23% del total). Dentro de los registros válidos (i.e., aquellos para los cuales hay algo en el campo `codigo_diagnostico`) Aún quedan 328 registros para los cuales el diagnóstico registrado es \*\*\*\*\* , y 73 registros en que el código es 0. Estos registros no han sido eliminados de la sábana de datos.

Las frecuencias de los tipos de enfermedades para licencias de origen común se pueden observar en la Figura 8

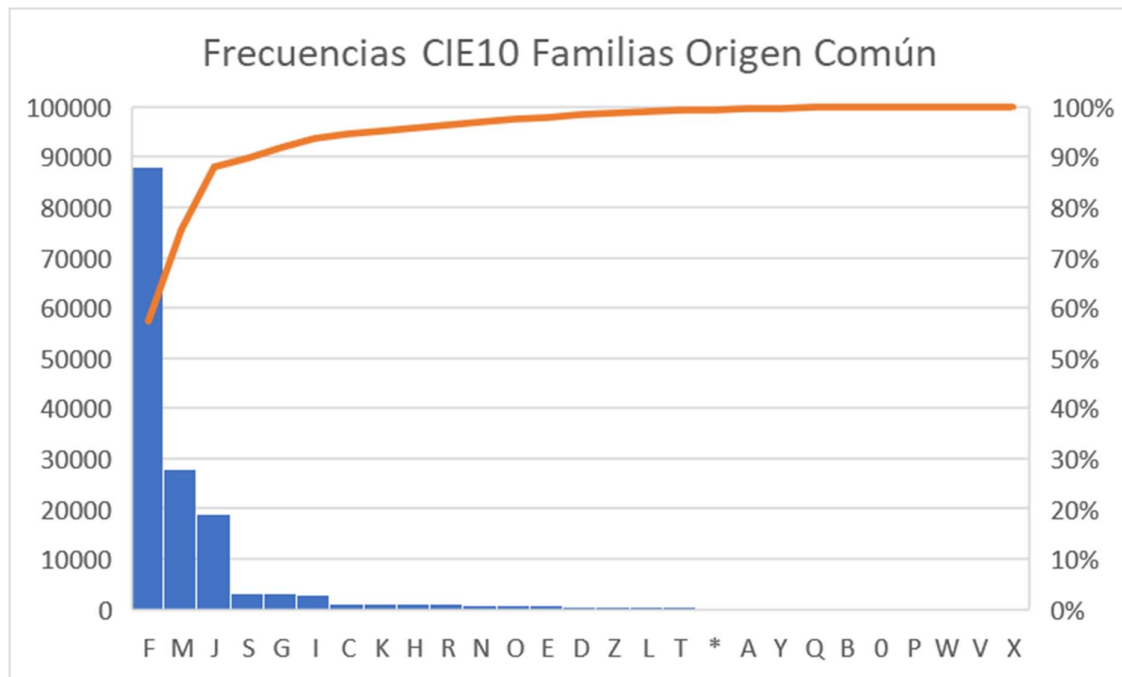


Figura 8: Frecuencias para cada familia CIE10 en el caso de licencias de origen común

Es notable la prevalencia de los casos relativos a trastornos mentales y del comportamiento (F), enfermedades del sistema osteomuscular y del tejido conectivo (M) y finalmente enfermedades del sistema respiratorio (J), todas ellas en conjunto acaparando aproximadamente un 87.89% de todos los casos existentes en la base de datos. La condición que sigue: Traumatismos, Envenenamientos y algunas otras consecuencias de causa externa (S + T), sólo representa un 2.33% del total de casos. Parece razonable entonces ocupar dos estrategias distintas para lidiar con la brutal diferencia de las frecuencias observadas.

Para efectos de observar diferencias en los grupos en que el estudio será dividido, la misma estadística se calculó para licencias maternas, los resultados se pueden apreciar en el siguiente gráfico:

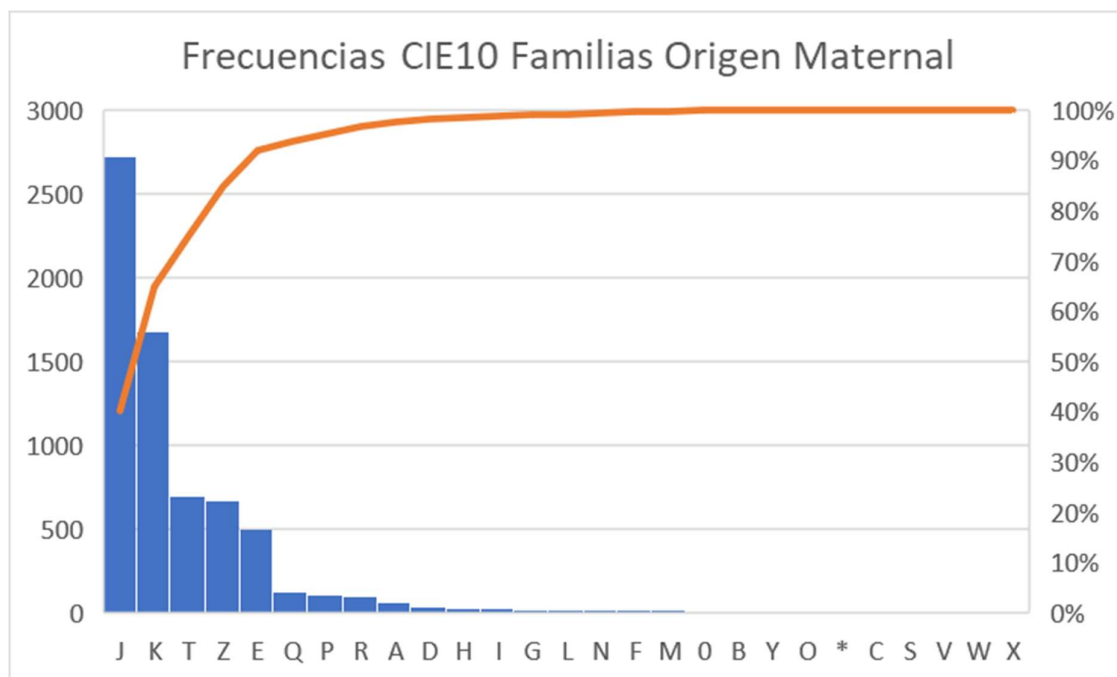


Figura 9: Frecuencias para cada familia CIE10 en licencias de origen materno

En este caso la prevalencia de las licencias se da por las causas: enfermedades del sistema respiratorio (J), enfermedades del aparato digestivo (K), traumatismos, envenenamientos y algunas otras consecuencias de causa externa (S +T), factores que influyen en el estado de salud y contacto con los servicios de salud (Z) y enfermedades endocrinas, nutricionales y metabólicas (E). Todas estas causas comprometen un 91.95% de los casos totales.

### 4.3 La Metodología Propuesta

Los datos provistos, como fue mencionado anteriormente corresponden a una sábana de datos que ha sido "pegada" horizontalmente. Lo que se hará con esta sábana es primero filtrar por familia y subfamilia de tal forma que sólo queden disponibles los casos que están relacionados con el caso que se presenta en cuestión. A partir de ahí, se puede construir un árbol clasificador basado en los datos disponibles (si es que hay muy pocos casos para la familia/subfamilia) entonces se procederá a utilizar sólo la familia, y posteriormente se aplicará el árbol clasificador para estudiar la clasificación propuesta para el caso presentado.

Lamentablemente, esta metodología no podría ser aplicada para los casos en que no haya demasiadas observaciones disponibles. Hemos vistos de los cálculos de frecuencia anteriores que, para el caso de licencias de origen común, la gran mayoría de los códigos de familia CIE10 poseen muy pocos casos como para que valga la pena intentar un árbol clasificador. En dicho caso, es preferible construir una simple probabilidad basado en frecuencias para ver que sugerencia entregar a partir de los datos, eventualmente se podría clasificar en acogimiento probable (70% o más de

acogimientos), acogimiento improbable (30% o menos de acogimientos) y finalmente acogimiento indefinido (el porcentaje de acogimientos mayor que 30% y menor que 70%).

#### 4.4 Aplicación de la Metodología a Licencias de Origen Común

##### 4.4.1 Utilizando solo las Familias CIE10 correspondientes a F, M y J

Como los números de casos son significativos, se ha escogido para poder entrenar un modelo, el uso de un árbol de decisión. La idea es ver la efectividad de un modelo de ese tipo para cada familia, entrenando en cada caso un modelo particular basado en los datos existentes. En todas las subsecciones que siguen se ha utilizado un 80% de los casos como entrenamiento para después ver el comportamiento en el 20% de casos que han sido separados (al azar) como conjunto de test.

###### 4.4.1.1 Familia CIE10 F

Para el caso de esta familia, los resultados son los siguientes:

```
Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "derecho_a_subsidio" "redictamen"
Number of terminal nodes: 3
Residual mean deviance: 1.39 = 54760 / 39390
Misclassification error rate: 0.3952 = 15568 / 39393

node), split, n, deviance, yval, (yprob)
* denotes terminal node

1) root 39393 58210 Acepta ( 0.498363 0.492499 0.009139 )
 2) derecho_a_subsidio: A 2491 1783 Acepta ( 0.905660 0.076676 0.017664 ) *
 3) derecho_a_subsidio: B,C 36902 54260 Rechaza ( 0.470869 0.520568 0.008563 )
    6) redictamen: *,1 20279 29960 Acepta ( 0.552739 0.436412 0.010849 ) *
    7) redictamen: 2 16623 23020 Rechaza ( 0.370992 0.623233 0.005775 ) *
```

La matriz de confusión para este caso es:

prediction	Acepta	Rechaza	S/P
Acepta	5136	2642	94
Rechaza	3261	6377	67

Es decir, se consigue un porcentaje de clasificación del orden del 65% para ambos resultados (dada la similitud de los tamaños de los subconjuntos de Aceptación y Rechazo no se balanceó el conjunto de datos). Si por el contrario hubiésemos sugerido utilizando frecuencias, la frecuencia mayoritaria sería rechazo lo que origina un error de clasificación del orden del 48% (el tamaño del subconjunto de aceptación).

###### 4.4.1.2 Familia CIE10 M

Al correr el modelo, los resultados que se obtienen son los siguientes:



```

Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "modificada"
Number of terminal nodes: 2
Residual mean deviance: 1.382 = 14370 / 10400
Misclassification error rate: 0.4545 = 4727 / 10401

```

```

node), split, n, deviance, yval, (yprob)
* denotes terminal node

```

```

1) root 10401 15170.0 Acepta ( 0.545524 0.446496 0.007980 )
 2) modificada: 0 1087 755.6 Acepta ( 0.908004 0.077277 0.014719 ) *
 3) modificada: 1 9314 13610.0 Acepta ( 0.503221 0.489586 0.007193 ) *

```

La matriz de confusión viene dada por:

prediction	Acepta	Rechaza	S/P
Acepta	3044	2468	42

En este caso, el árbol clasificador responde de la misma manera que una respuesta basada en frecuencias.

#### 4.4.1.3 Familia CIE10 J

Al correr el modelo, se obtienen los siguientes resultados:

```

Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "causa_rechazo" "actividad_laboral_trabajador" "region_empleador"
[4] "tipo_entidad_pagadora_subsidio" "ocupacion_trabajador" "edad_trabajador"
[7] "tipo_formulario" "num_dias_licencia" "inicio_tramite_invalidez"
Number of terminal nodes: 22
Residual mean deviance: 0.7909 = 476.9 / 603
Misclassification error rate: 0.1776 = 111 / 625

```

```

node), split, n, deviance, yval, (yprob)
* denotes terminal node

```

```

1) root 625 885.900 Acepta ( 0.608000 0.384000 0.008000 )
 2) causa_rechazo: *,3 153 137.700 Acepta ( 0.849673 0.143791 0.006536 )
   4) actividad_laboral_trabajador: **,3,4,6,7 38 58.330 Acepta ( 0.605263 0.368421 0.026316 ) *
   5) actividad_laboral_trabajador: 0,1,2,5,8,9 115 58.080 Acepta ( 0.930435 0.069565 0.000000 )
   10) region_empleador: **,1,13 66 48.750 Acepta ( 0.878788 0.121212 0.000000 ) *
   11) region_empleador: 12,14,15,2,5,6,7,8,9 49 0.000 Acepta ( 1.000000 0.000000 0.000000 ) *
 3) causa_rechazo: 1,2,4,5 472 692.700 Acepta ( 0.529661 0.461864 0.008475 )
   6) tipo_entidad_pagadora_subsidio: A,C 174 227.100 Rechaza ( 0.310345 0.683908 0.005747 )
   12) actividad_laboral_trabajador: 0,2,3,6,8,9 108 158.600 Rechaza ( 0.444444 0.546296 0.009259 )
   24) actividad_laboral_trabajador: 0,3 7 5.742 Acepta ( 0.857143 0.000000 0.142857 ) *
   25) actividad_laboral_trabajador: 2,6,8,9 101 137.100 Rechaza ( 0.415842 0.584158 0.000000 )
   50) ocupacion_trabajador: 11,14,16,17,18,19 95 126.100 Rechaza ( 0.378947 0.621053 0.000000 )
   100) region_empleador: **,1,6,9 25 13.940 Rechaza ( 0.080000 0.920000 0.000000 ) *
   101) region_empleador: 10,12,13,4,5,8 70 96.980 Rechaza ( 0.485714 0.514286 0.000000 )
   202) actividad_laboral_trabajador: 2,8 15 11.780 Rechaza ( 0.133333 0.866667 0.000000 ) *
   203) actividad_laboral_trabajador: 6,9 55 74.770 Acepta ( 0.581818 0.418182 0.000000 )

```

```

406) edad_trabajador < 42 25 32.670 Rechaza ( 0.360000 0.640000 0.000000 ) *
407) edad_trabajador > 42 30 32.600 Acepta ( 0.766667 0.233333 0.000000 ) *
51) ocupacion_trabajador: 12,13 6 0.000 Acepta ( 1.000000 0.000000 0.000000 ) *
13) actividad_laboral_trabajador: 1,5 66 40.210 Rechaza ( 0.090909 0.909091 0.000000 )
26) edad_trabajador < 60 15 20.190 Rechaza ( 0.400000 0.600000 0.000000 ) *
27) edad_trabajador > 60 51 0.000 Rechaza ( 0.000000 1.000000 0.000000 ) *
7) tipo_entidad_pagadora_subsidio: B,D 298 410.000 Acepta ( 0.657718 0.332215 0.010067 )
14) region_empleador: **,13,14,2,3,5,6,7,8 274 360.200 Acepta ( 0.700730 0.288321 0.010949 )
28) tipo_formulario: 1,3 120 124.500 Acepta ( 0.825000 0.158333 0.016667 )
56) num_dias_licencia < 14.5 75 94.930 Acepta ( 0.733333 0.253333 0.013333 ) *
57) num_dias_licencia > 14.5 45 9.591 Acepta ( 0.977778 0.000000 0.022222 ) *
29) tipo_formulario: 2 154 217.000 Acepta ( 0.603896 0.389610 0.006494 )
58) ocupacion_trabajador: 12,14,15,16,17 94 108.900 Acepta ( 0.734043 0.265957 0.000000 )
116) inicio_tramite_invalidez: 1 9 6.279 Rechaza ( 0.111111 0.888889 0.000000 ) *
117) inicio_tramite_invalidez: 2 85 85.070 Acepta ( 0.800000 0.200000 0.000000 )
234) edad_trabajador < 40.5 20 27.530 Acepta ( 0.550000 0.450000 0.000000 ) *
235) edad_trabajador > 40.5 65 48.490 Acepta ( 0.876923 0.123077 0.000000 )
470) region_empleador: **,14,3,5,6,7 32 0.000 Acepta ( 1.000000 0.000000 0.000000 ) *
471) region_empleador: 13,2,8 33 36.550 Acepta ( 0.757576 0.242424 0.000000 ) *
59) ocupacion_trabajador: 13,19 60 89.900 Rechaza ( 0.400000 0.583333 0.016667 )
118) actividad_laboral_trabajador: **,3,8 27 25.870 Rechaza ( 0.185185 0.814815 0.000000 )
236) num_dias_licencia < 8.5 12 16.300 Rechaza ( 0.416667 0.583333 0.000000 ) *
237) num_dias_licencia > 8.5 15 0.000 Rechaza ( 0.000000 1.000000 0.000000 ) *
119) actividad_laboral_trabajador: 1,5,6,7,9 33 52.190 Acepta ( 0.575758 0.393939 0.030303 ) *
15) region_empleador: 10,12,15,9 24 21.630 Rechaza ( 0.166667 0.833333 0.000000 )
30) edad_trabajador < 44.5 17 0.000 Rechaza ( 0.000000 1.000000 0.000000 ) *
31) edad_trabajador > 44.5 7 9.561 Acepta ( 0.571429 0.428571 0.000000 ) *

```

La matriz de confusión viene dada por:

prediction	Acepta	Rechaza	S/P
Acepta	651	802	8
Rechaza	899	1439	6

Aquí podemos percibir que la clasificación correcta en el conjunto de datos test es del orden de 44.5% para los "Acepta" y de un "61.4%" para los "Rechaza". Si hubiésemos utilizado la regla de sugerir basado sólo en frecuencias, entonces la clasificación proveería un método con un error del 40% en el conjunto de test (todo se clasificaría como "Rechaza" por ser esta mayoría). Es notable que en este caso el árbol de decisión se aprecia muy bueno para el conjunto de entrenamiento, pero bastante malo para el conjunto de test. Se probaron otros ejemplos eligiendo otros conjuntos de entrenamiento y de test y los resultados consistentemente muestran este comportamiento.

#### 4.4.2 Utilizando solo las Familias CIE10 correspondientes al complemento de la subsección 4.4.1

En este ejemplo, lo que se hizo fue filtrar la base de datos eliminando para ello los códigos de diagnóstico F, M y J para las licencias de origen común. A partir de ahí se pueden calcular las probabilidades basados en frecuencias.

Tabla 6: Cálculo de probabilidades familias CIE 10 poco relevantes

Familia	Acepta	Rechaza	SP	% Acepta
*	180	144	3	0.55
0	39	29	0	0.57
A	111	87	2	0.56
B	118	26	12	0.76
C	805	340	6	0.70
D	314	229	8	0.57
E	454	242	5	0.65
G	1739	1261	20	0.58
H	580	349	14	0.62
I	1961	822	47	0.69
K	527	452	6	0.54
L	321	170	4	0.65
N	585	307	11	0.65
O	576	251	8	0.69
P	10	4	0	0.71
Q	123	43	1	0.74
R	529	396	9	0.57
S	2058	1065	28	0.65
T	226	196	6	0.53
V	2	0	0	1.00
W	3	6	0	0.33
X	1	0	0	1.00
Y	123	54	3	0.68
Z	309	186	9	0.61

#### 4.4.3 Utilizando Codificación de Variables Dummies

Para efectos de poder utilizar una noción de distancia, se ha decidido codificar los factores como variables dummies y crear un Dataframe a partir de ellas. Se dejaron las variables familia, subfamilia y diagnóstico en la Sabana para poder filtrar. La idea es definir una noción de distancia, al tener todas las variables una cuantificación numérica, esto debiese ser posible puesto que las columnas sólo tienen valor 0 o 1. El codificador de variables dummies se logró mediante el uso de un código en R que permite crear nuevas columnas para las variables basado en los niveles de dicha variable categórica, desafortunadamente esto implica también crear las columnas correspondientes a NA que no contienen valor alguno y por lo tanto son eliminadas del análisis. El código para la creación de variables dummies a partir de factores es el siguiente:

```

for(i in names(SabanaComun)){
  if (is.factor(SabanaComun[[ i ]])) {
    for(level in unique(SabanaComun[[ i ]])){
      SabanaComun[paste(i, level, sep = "_")] <- ifelse(SabanaComun[[ i ]] == level,
1, 0)
    }
  }
}

```

Algunas otras variables numéricas tales como `edad_trabajador`, `num_dias_incapacidad_aut`, etc. fueron tomadas directamente de la sábana de datos original.

```

SabanaComunWDummies <- SabanaComun[,c("familia", "subfamilia", "diagnostico", "edad_trabajador",
"num_dias_licencia", "num_dias_incapacidad_aut", "num_dias_previos_autorizados", "modificada",
"reclamada", "reconsideracion", "origen_comun", "maternal", "tipo_formulario_2", "tipo_formulario_1",
"tipo_formulario_3", "tipo_formulario_*", "articulo_77bis_2", "articulo_77bis_*", "sexo_trabajador_1",
"sexo_trabajador_2", "tipo_licencia_1", "tipo_licencia_7", "recuperabilidad_laboral_1",
"recuperabilidad_laboral_2", "inicio_tramite_invalidez_1", "inicio_tramite_invalidez_2",
"inicio_tramite_invalidez_*", "inicio_tramite_invalidez_0", "tipo_reposo_1", "tipo_reposo_2",
"tipo_reposo_3", "jornada_reposo_A", "jornada_reposo_*", "jornada_reposo_B", "jornada_reposo_C",
"lugar_reposo_1", "lugar_reposo_3", "lugar_reposo_2", "lugar_reposo_*", "licencia_modificada_2",
"licencia_modificada_1", "licencia_modificada_*", "licencia_modificada_3", "tipo_licencia_medica_1",
"tipo_licencia_medica_7", "tipo_licencia_medica_*", "tipo_licencia_medica_6",
"tipo_licencia_medica_2", "tipo_licencia_medica_4", "periodo_2", "periodo_1", "periodo_*",
"periodo_3", "estado_resolucion_1", "estado_resolucion_2", "estado_resolucion_*",
"estado_resolucion_5", "estado_resolucion_4", "estado_resolucion_3", "tipo_resolucion_1",
"tipo_resolucion_*", "tipo_resolucion_2", "redictamen_2", "redictamen_1", "redictamen_*",
"causa_rechazo_*", "causa_rechazo_5", "causa_rechazo_1", "causa_rechazo_4", "causa_rechazo_2",
"causa_rechazo_3", "tipo_reposo_autorizado_1", "tipo_reposo_autorizado_2", "tipo_reposo_autorizado_3",
"jornada_reposo_autorizada_A", "jornada_reposo_autorizada_*", "jornada_reposo_autorizada_B",
"jornada_reposo_autorizada_C", "derecho_a_subsidio_A", "derecho_a_subsidio_C", "derecho_a_subsidio_B",
"derecho_a_subsidio_*", "region_empleador_8", "region_empleador_9", "region_empleador_13",
"region_empleador_5", "region_empleador_*", "region_empleador_2", "region_empleador_6",
"region_empleador_NULL", "region_empleador_7", "region_empleador_12", "region_empleador_4",
"region_empleador_1", "region_empleador_10", "region_empleador_11", "region_empleador_15",
"region_empleador_3", "region_empleador_14", "actividad_laboral_trabajador_2",
"actividad_laboral_trabajador_8", "actividad_laboral_trabajador_5", "actividad_laboral_trabajador_9",
"actividad_laboral_trabajador_7", "actividad_laboral_trabajador_6", "actividad_laboral_trabajador_0",
"actividad_laboral_trabajador_*", "actividad_laboral_trabajador_1", "actividad_laboral_trabajador_3",
"actividad_laboral_trabajador_4", "actividad_laboral_trabajador_14", "ocupacion_trabajador_19",
"ocupacion_trabajador_16", "ocupacion_trabajador_17", "ocupacion_trabajador_15",
"ocupacion_trabajador_14", "ocupacion_trabajador_13", "ocupacion_trabajador_12",
"ocupacion_trabajador_18", "ocupacion_trabajador_11", "tipo_regimen_previsional_2",
"tipo_regimen_previsional_1", "tipo_regimen_previsional_*", "calidad_trabajador_1",
"calidad_trabajador_3", "calidad_trabajador_2", "calidad_trabajador_4",
"tipo_entidad_pagadora_subsidio_C", "tipo_entidad_pagadora_subsidio_D",
"tipo_entidad_pagadora_subsidio_A", "tipo_entidad_pagadora_subsidio_B",
"tipo_entidad_pagadora_subsidio_*", "tipo_entidad_pagadora_subsidio_E", "modificada_0",
"modificada_1", "resultado_Acepta", "resultado_Rechaza", "resultado_S/P")]

```

También, se cambiaron todos los NA existentes en las distintas columnas por el valor cero, de esta manera se evita que los registros causen problemas al momento de calcular distancias. Para efectos de poder correr el algoritmo, se dividieron los datos en dos conjuntos: un conjunto de entrenamiento y un conjunto de prueba. El conjunto de prueba corresponde al 2% del total de registros, seleccionados al azar.

El algoritmo esencialmente lo que hace es primero filtrar todos los registros del conjunto de entrenamiento para identificar aquellos que poseen la misma familia que el elemento para el cual se desea consultar. A partir de ahí, se calculan los vecinos más cercanos y se establece como condición que la distancia máxima sea menor que una cierta tolerancia (no vale la pena comparar con vectores tan distintos). Una vez

que se determinan los vecinos, que por este mecanismo son bastante similares al caso en cuestión, se procede a calcular la tasa de aceptación y la tasa de rechazo. Con estas tasas, se decide de manera mayoritaria la sugerencia. Si no hay casos que satisfagan la condición de estar a una distancia menor que la tolerancia, entonces se considera que dichos casos no poseen vecinos, y por lo tanto no se puede sugerir nada.

Se hizo una corrida basado en estos principios y se obtuvo solo un 47% de las clasificaciones predichas por este método siendo correctas. Es bastante claro que no existe una lógica demasiado evidente detrás de las asignaciones y esto lleva a pensar que pueden existir otros factores no tomados en cuenta en estos momentos que podría ser influyentes. Entre otros, se espera que distintos médicos resuelvan consistentemente de manera distinta los casos que se les presentan, y por este motivo sería interesante contar con un identificador del médico que ha resuelto en cada caso para así poder entre otras cosas ser capaz de caracterizar individualmente a cada médico y su respuesta frente a casos que se les presentan. Por otro lado, también sería deseable el poder contar con información relativa al número de licencias previas asociadas al mismo RUT en el sistema y en último lugar el número de licencias asociadas al mismo RUT en el mismo expediente. De discusiones con la contraparte técnica de la SUSESO, queda claro que la forma en que los casos son decididos tiende a considerar otros factores como los mencionados previamente, junto a ello se debe sumar el sesgo personal que cada médico posee y que de momento no es posible incorporar puesto que esta información es inexistente en la base de datos.

A pesar de este aparente fracaso, un factor a destacar es que la metodología propuesta permite generar un conjunto de casos "cercaños" que pueden ser presentados al profesional a cargo del caso para permitirle tener mejores antecedentes respecto del cómo poder orientar su fallo. Este es un producto en sí mismo que eventualmente puede evolucionar a futuro en un mejor producto de datos, que permitirá asistir de mejor manera al profesional en su trabajo diario.

Actualmente la base de datos no registra tiempos asociados a la resolución de casos, sin embargo, y a partir de discusiones con la contraparte técnica de la SUSESO, se piensa que, a partir de la variabilidad en las respuestas generadas de la manera anteriormente descrita, podría encontrarse el germen de un mecanismo de estimación automática de la dificultad de un caso. Por ejemplo, si un caso dado tiene un 90% de vecinos alineados con un resultado, no es difícil entender que este caso debería ser más fácil de decidir que uno que frente a similares condiciones de distancia en el espacio de casos presenta un Split 50-50 a favor o en contra. La variabilidad asociada a los vecinos de un caso dado debería ser una componente de la medida de la dificultad, la otra componente siendo si el caso se acepta o se rechaza. La racionalidad de lo último tiene que ver con el hecho de que los reclamos que son fáciles de acoger no generan potenciales consecuencias judiciales a posteriori, en el caso de reclamos rechazados, siempre hay una posibilidad de que el caso vuelva a ser presentado después de no acogerlo y por ende se estima que se trataría de colocar más esfuerzo para evitar esta situación.

Para ilustrar la propuesta anterior se propone la siguiente tabla:

Tabla 7: Propuesta conceptual de estimación de esfuerzo basado en naturaleza de casos cercanos

Tipo Resolución		Acoge	No Acoge	Sin Pronunciamiento
Tipo Vecindad				
	Inexistente	Yellow	Yellow	Yellow
	Concentrado (80+)	Green	Green	Yellow
	Medio Concentrado (80-)	Green	Yellow	Red
	Variable (50-50)	Red	Red	Red

Los colores significan de manera bien básica los tres niveles que se podrían presentar siendo el verde un caso fácil o rutinario, el rojo representando un caso complejo y el amarillo denotando un caso de mediana complejidad. Una de las gracias de un sistema propuesto de esta forma es que es posible entrenarlo continuamente y hacerlo evolucionar, aún más, si se obtiene un modelo basado en cada profesional, donde los tiempos además están claramente definidos, es posible entonces hacer una caracterización más útil del esfuerzo asociado a la resolución de los casos que es un insumo bastante importante al momento de decidir su asignación.

#### 4.5 Aplicación de la Metodología a Licencias de Origen Maternal

Dado el bajo número de casos que son considerados maternales y en virtud de las conclusiones obtenibles a partir del ejemplo anterior, se propone tener mejor claridad respecto de los datos adicionales que podrían conseguirse para poder ver si esto permite predecir el diagnóstico de mejor manera.

#### 4.6 Observaciones Adicionales

El proceso requerido para poder correr la heurística necesita la transformación de todas las variables categóricas en variables dummies, lo que a su vez es requisito esencial para poder medir distancias con la métrica propuesta. Dado lo importante de este requisito, se propone que las variables categóricas sean codificadas como variables dummies para así poder facilitar análisis futuros. Otros algoritmos requieren de la misma manera el uso de variables dummies, así que tener los datos disponibles en este formato abren las puertas a otras posibilidades de modelamiento futuro.

#### 4.7 Conclusiones

Del análisis de los datos, es bastante claro que la Ley de Pareto nuevamente se cumple en el sentido de que con a lo sumo un 20% de los casos posibles se acumula un 80% de la frecuencia. En el caso particular de los datos que tenemos a nuestro alcance, es importante apreciar una diferencia estructural en cuanto al tipo de licencia médica: los motivos difieren entre licencias de origen común y maternales.

Llama la atención de que en ambos casos la razón prevalente común en ambos grupos sean las enfermedades respiratorias, sin embargo, esto se podría deber a la eventual prevalencia de enfermedades respiratorias en los valles de Chile Central (Santiago, Rancagua, Temuco) que por un tema de contaminación tienden a producir un incremento de casos de enfermedad respiratoria con clara estacionalidad manifiesta todos los años.

El análisis de los distintos tipos de metodologías que se pueden utilizar para “predecir” si se acoge o no un caso indica que no hay mucha lógica al momento de decidir. Por un lado, si se trata de buscar los casos más similares, no es posible encontrar una clara tendencia en bastante de los casos, esto entrega un clasificador con una tasa de clasificación correcta del orden de 47% lo que de hecho es sólo ligeramente más alto que la utilización de árboles de decisión, pero aún en la opinión del autor deficiente.

El uso de árboles de decisión requiere que los conjuntos para estudio sean lo suficientemente grandes lo que impone una restricción importante dada la naturaleza muy concentrada de los casos en términos de familias de enfermedades tanto para las licencias de origen común como para las licencias maternas.

Alternativas basadas en el uso de información adicional podrían entregar mejores luces respecto de la viabilidad de construir un mejor asistente de diagnóstico. Se propone que en particular se agregue al menos los campos “médico a cargo del caso” y también el número de folios previos relacionados al RUT del reclamante del caso en cuestión. De esta manera es posible agregar un nuevo nivel de filtrado basado en el médico y también ver el impacto de un historial de licencias en el sistema.

Se cree que avanzar en obtener mejor información permitirá no sólo mejorar los modelos de asistencia en el diagnóstico, sino que además, como subproducto, permitir estimar de mucho mejor manera el esfuerzo asociado a casos de distinta complejidad.

## 4.8 Apéndice

### 4.8.1 Familias de Diagnóstico CIE10

- A, B: Ciertas enfermedades infecciosas y parasitarias
- C, D: Neoplasias; Enfermedades de la sangre y de los órganos hematopoyéticos y otros trastornos que afectan el mecanismo de la inmunidad
- E: Enfermedades endocrinas, nutricionales y metabólicas
- F: Trastornos mentales y del comportamiento
- G: Enfermedades del sistema nervioso
- H: Enfermedades del ojo y sus anexos; Enfermedades del oído y de la apófisis mastoides
- I: Enfermedades del sistema circulatorio
- J: Enfermedades del sistema respiratorio
- K: Enfermedades del aparato digestivo
- L: Enfermedades de la piel y el tejido subcutáneo

- M: Enfermedades del sistema osteomuscular y del tejido conectivo
- N: Enfermedades del aparato genitourinario
- O: Embarazo, parto y puerperio
- P: Ciertas afecciones originadas en el periodo perinatal
- Q: Malformaciones congénitas, deformidades y anomalías cromosómicas
- R: Síntomas, signos y hallazgos anormales clínicos y de laboratorio, no clasificados en otra parte
- S, T: Traumatismos, envenenamientos y algunas otras consecuencias de causa externa
- V, W, X, Y: Causas externas de morbilidad y de mortalidad
- Z: Factores que influyen en el estado de salud y contacto con los servicios de salud (exceptuando lo que está en otros)
- U: Códigos para situaciones especiales
- Otros códigos creados adicionales a CIE10
- 1650= Reposo Prenatal
  - Z02.70= Reposo Prenatal
- 645= Ampliación Reposo Prenatal
  - Z02.71 = Reposo Postnatal
- 650= Reposo Post Natal
  - Z02.72= Ampliación Reposo Prenatal



## 5 Predicción y Clasificación de Comportamiento Post-Atención

### 5.1 Introducción

Parte no menor de la carga de trabajo de la SUSESO en cuanto a reclamos se refiere guarda relación con las solicitudes de revisión de casos ya dictaminados. Este flujo circulante de casos se estima en la actualidad en torno al 17% de los casos presentados lo que no deja de ser menor puesto que este número asciende a más o menos 9.000 casos anuales, lo que representa un costo significativo. Si bien no es posible eliminar las revisiones a un 100%, se establece la hipótesis de que podría ser posible predecir con algún grado de confiabilidad los casos que si son rechazados entonces serán presentados nuevamente como revisión, de esta manera se puede hacer una idea de la demanda derivada de servicios, número que permite entre otras cosas poder planificar los recursos necesarios para poder hacer frente a los reclamos que se presentan año a año a la SUSESO, de manera fluida y dentro de los plazos esperados.

De manera básica, la primera pregunta que el tomador de decisiones se hace guarda relación con la posibilidad de poder establecer un modelo predictivo para las reconsideraciones. Es decir, del universo de licencias reclamadas, tratar de estimar proactivamente cuales de dichos reclamos presentarán reconsideración si es que no se acogen (notar que se hace un supuesto implícito de que un reclamo acogido no presentará reconsideración). Si es que existe algún tipo de lógica en el comportamiento de los usuarios, entonces se cree que este comportamiento debiese ser observable en los datos y permitiría caracterizar un caso en función de su propensión a presentar reconsideración. Hay que notar en todo caso que, puesto que no hay una adecuada caracterización socioeconómica de los pacientes, se hace difícil poder discriminar ciertas características que se estiman a priori pueden influir notablemente en la decisión de presentar una reconsideración: nivel de ingresos, nivel de estudios, comportamiento previo (si alguien ha presentado reconsideración una vez o más entonces debería ser más probable que lo haga de nuevo), etc.

La siguiente pregunta que uno podría querer formular en un trabajo de esta naturaleza guarda relación con la probabilidad de que un reclamo no acogido termine judicializado. Lamentablemente, para poder responder esta pregunta es necesario tener información que hoy día no está presente en las sábanas de datos puesto que no se puede establecer a ciencia cierta qué casos fueron judicializados. Si bien es cierto que la información respecto de los folios judicializados es parte del expediente del proceso judicial, no es posible de momento conectar con esa información de manera clara y unívoca, por lo tanto, en esta etapa se debe abandonar este objetivo dejando planteada la inquietud respecto del poder conseguir esta información a futuro en un formato ameno al análisis.

Finalmente, otra pregunta interesante que surge en el contexto de esta etapa es ver si es que es posible utilizar los datos para apoyar efectivamente las labores fiscalizadoras. Se utilizará un ejemplo simple basado en datos reales provistos por la

SUSESO para mostrar el tipo de problemas en que se pueden caer al utilizar malos indicadores. Un trabajo de esta naturaleza es un trabajo que evoluciona en el tiempo puesto que requiere responder preguntas que surgen de las operaciones y el día a día. Es claro observar que cuando un factor problemático ha sido controlado, el foco de atención pasa al siguiente problema que no se podía ver antes puesto que estaba de alguna manera cubierto por otros problemas más importantes. Usualmente este es el tipo de trabajo que cae dentro del dominio de lo que se conoce como analytics, es intensivo en trabajo y requiere mantención constante, por lo tanto es necesario tomar decisiones arquitecturales para poder implementar una solución software que permita crear análisis y reportes en la medida de que ellos se necesiten.

## 5.2 Predictor de Reconsideraciones

Para esta parte la pregunta fundamental que se quiere responder es ver si es que es posible predecir cuales reclamos no acogidos presentarán reconsideraciones. Para ello se correrá un modelo de árbol clasificador sobre la base de datos "horizontal" (que es una tabla conjunta formada a partir de la sábana SILMSIL y expedientes\_workflow). Recordemos que esta sábana posee una variable indicadora de si el expediente en cuestión es una reconsideración y aún más, es posible a partir de ella (mirando el número de folio) determinar la cantidad de veces que un expediente ha presentado reconsideraciones, sin embargo, no es claro determinar cuales folios en particular han sido presentados a reconsideración, más aún, el número de folio está codificado para preservar confidencialidad de los casos en los últimos archivos provistos por la SUSESO.

Nos concentraremos entonces en el caso (aparentemente más simple) de determinar a partir de aquellos reclamos que no son acogidos (es decir con un valor de Rechaza en el campo Resultado de la sábana) la probabilidad de que ellos presenten revisión (esto medido a través del campo Reconsideracion agregado a la sábana por la SUSESO). Los datos son leídos como es habitual del archivo e ingresados al sistema R. A partir de ahí se procede a filtrar los datos de acuerdo con la descripción previa, es decir, primero se dejan solamente los registros para los cuales los reclamos no son acogidos (rechazados en el lenguaje de la sábana de datos) y a partir de este subconjunto se procederá a correr un árbol clasificador para ver si es que es posible encontrar una regla que permita predecir con cierta decencia quiénes presentarán reconsideración.

Los resultados del árbol de clasificación (después de obtener un conjunto balanceado) son:

```
Classification tree:
tree(formula = f, data = train.set)
Variables actually used in tree construction:
[1] "licencia_modificada"
Number of terminal nodes: 2
Residual mean deviance: 1.361 = 34250 / 25160
Misclassification error rate: 0.4402 = 11078 / 25164
```

El árbol de decisión que se obtiene es el siguiente:

```
node), split, n, deviance, yval, (yprob)
  * denotes terminal node

1) root 25164 34710 0 ( 0.5411 0.4589 )
  2) licencia_modificada: *,1 10497 13930 0 ( 0.6209 0.3791 ) *
  3) licencia_modificada: 2 14667 20320 Si ( 0.4840 0.5160 ) *
```

Finalmente la matriz de confusión resultante es:

```
prediction    0    Si
              0 1991 1351
              Si 4990 5478
```

Lo que confirma la mala calidad del modelo puesto que en el conjunto de entrenamiento el error es del orden de un 44% y en el conjunto de test se obtienen resultados similares. Una nueva corrida muestra una matriz de confusión como la que sigue:

```
prediction    0    Si
              0 1942 1361
              Si 5028 5479
```

Lo que muestra que el modelo no varía mucho frente a instancias individuales y mantiene su mala calidad de clasificación.

Sobre la base de estos resultados se puede ver que no es posible predecir de manera razonable quien presentará reconsideraciones a la SUSESO basados solamente en las variables actualmente en la base de datos. Eventualmente con otras variables adicionales se podría establecer una relación causal mejor. Se asume que si el monto en juego es elevado, entonces deberían haber más incentivos para apelar a una decisión de no acogimiento de un reclamo por parte de SUSESO.

### 5.3 Ayuda a la Fiscalización

La existencia de una sábana de datos como la que posee actualmente la SUSESO permite no sólo mirar la información con fines predictivos, sino que además puede ser utilizada para realizar análisis más globales en torno al comportamiento de individuos dentro del sistema que podrían alertar de ciertas posibles irregularidades. El listado de las cosas que se podrían hacer podría llegar a ser bastante extenso y en cierta medida depende del tipo de preguntas que el fiscalizador quiera hacer. Por nombrar sólo una posible aplicación se podrían establecer indicadores estadísticos para ver si es que el comportamiento de un médico particular difiere sustancialmente del comportamiento promedio o esperable en promedio. La base de dicha caracterización tiene que ver con el comportamiento estadístico de los datos, por ejemplo, si sabemos que los valores "normales" de otorgamiento de licencias por semana por cada médico sigue una distribución normal con una cierta media y una cierta varianza, entonces es

ciertamente posible poder establecer un intervalo de confianza en torno al valor medio y a partir de ahí concentrar esfuerzos en aquellos profesionales que escapan el límite superior de dicho intervalo de confianza.

Para ser un poco más concretos, miraremos un ejemplo del tipo de análisis que puede realizarse de manera más o menos rutinaria con los datos, es posible mirar el desempeño de los médicos de la SUSESO durante los años 2014 – 2017 en términos de la cantidad de casos asignados en el período y los resultados particulares de la evaluación de reclamos presentados a esos médicos. Este tipo de información se puede mirar de manera global (es decir durante todo el período) o bien en períodos a determinar (por ejemplo, año a año, trimestre a trimestre, mes a mes, etc.). Para efectos de simplificar la ilustración, se procederá a estudiar el comportamiento de la variable Resultado de la sábana de datos, pero considerando a cada médico de la SUSESO por separado. Por condiciones de confidencialidad se utiliza un identificador en vez del nombre del médico. Lo otro es que hay que considerar que no necesariamente todos los médicos han trabajado igualmente durante el período en consideración.

Si miramos todos los datos durante el período en consideración se puede concluir que de un total de 159.677 casos para los cuales hubo un reclamo, 78.489 de ellos fueron acogidos, 79.744 fueron no acogidos y 1.414 casos terminaron sin pronunciamiento. Es decir, a este nivel global se puede concluir que básicamente el servicio entrega un resultado distribuido 50%-50% para acoger y no acoger reclamos. Uno podría quedarse con este número en la cabeza, sin embargo, como veremos en la siguiente subsección, los mismos números vistos de otra manera cuentan otra historia.

#### 5.3.1 Primer Ejemplo: Comportamiento global Médicos de la SUSESO en el período 2014 – 2017

Se filtró la sábana de datos y se creó una tabla de contingencia basados en los registros otorgados por la SUSESO. Esta base de datos esencialmente contabiliza los reclamos acogidos, no acogidos y sin pronunciamiento para cada profesional. Dicho detalle se encuentra en la Tabla 10 para su análisis más en detalle.

Si se calcula el porcentaje para cada doctor en términos de la cantidad de reclamos acogidos, se puede observar grandes variaciones entre médicos. Hay algunos médicos que tienen un 100% de aceptación o rechazo, hay médicos con muy pocos casos y otros médicos con muchos (llama la atención los médicos identificados con el número 38, 43, 51 y 72). En particular, hay un médico con 17.755 casos asignados durante el período 2014 – 2017, si notamos los años bajo consideración para esta sábana de datos, esto corresponde a 192 semanas estimadas de trabajo (un año son 52 semanas, considerando tres semanas de vacaciones y una semana de enfermedad por año), esto considerando 5 días de trabajo por semana (es decir sin considerar feriados) entrega 960 días laborales. Cuando se observan los casos resueltos en el período, se puede concluir que en promedio este profesional dictamina 18.5 casos diariamente. Esto a su vez implica dos casos por hora, es decir un caso cada media hora y esto sin considerar reuniones, colación idas al baño y otras típicas pausas que un profesional

habitualmente tiene en su jornada. Claramente un análisis de este tipo se puede entregar información errónea si es que no se consideran otras fuentes adicionales. En el caso recién descrito, hay que conocer entre otras cosas el tipo de jornada laboral que la profesional tiene, que tipo de casos se le han asignado, etc.

El promedio de los porcentajes de aceptación de reclamos para los médicos es de un 65%. Claramente esta historia es distinta de la historia que cuenta el global anteriormente mencionado. La diferencia entre los números globales y el promedio de los indicadores individuales tiene que ver con el volumen tratado por los profesionales. No es lo mismo tener un 80% de aprobación cuando se han visto 300 casos que tener un 40% de aprobación cuando se han visto 3000 casos. Para este ejemplo particular el promedio será de 60% sin embargo el porcentaje real de aprobación de casos es del 44% aproximadamente.

### 5.3.2 Segundo Ejemplo: Comportamiento Anual de Médicos de la SUSESO en el período 2014 – 2017

El mismo conjunto de datos anterior fue ahora filtrado no sólo por médico a cargo del caso, sino que además por año. En este caso los números tienden a contar otra historia aún más interesante que la anterior. Este resultado se resume en la Figura 10.

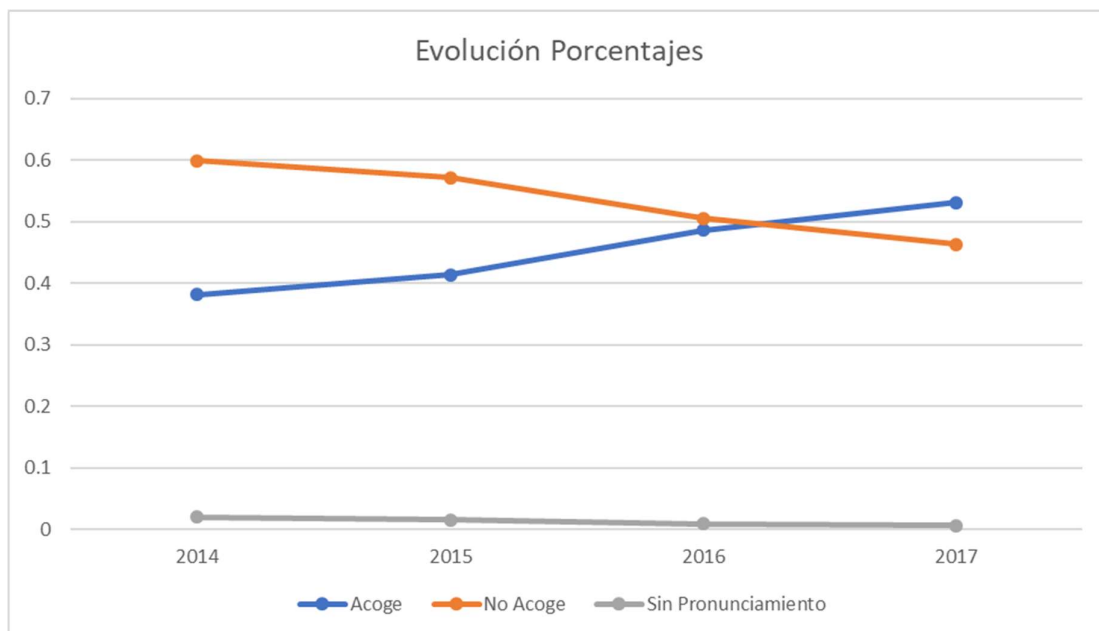


Figura 10: Evolución de Porcentajes de Aprobación/Rechazo

Se puede ver que en general ha habido un cambio sostenido en los porcentajes de aprobación/rechazo de reclamos. Adicionalmente el número de reclamos presentados cada año ha también experimentado una evolución tal y como muestra la Figura 11.

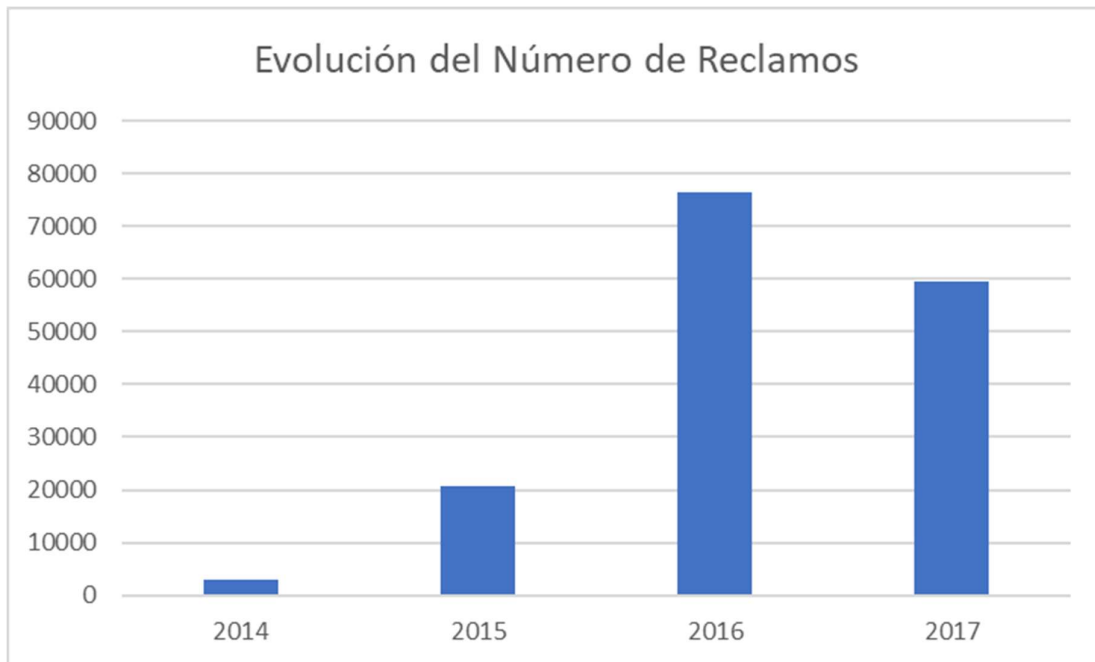


Figura 11: Evolución del Número de Reclamos

Ciertamente este nivel de desagregación de los datos entrega más información que el nivel anterior, permite entender de mejor manera el fenómeno.

Volviendo al caso del médico con 17.755 casos asignados en cuatro años, podemos ver ahora su comportamiento desagregado por año, esto entrega la siguiente información:

Tabla 8: Ejemplo de productividad diaria estimada por año para un médico

Año	Acepta	Rechaza	S/P	TOTAL	Productividad diaria Estimada
2014	44	19	0	63	0.26
2015	190	272	0	462	1.93
2016	3295	4815	18	8128	33.87
2017	3879	5213	10	9102	37.93

Es claro que hubo un cambio de circunstancias puesto que de otra manera no es posible explicar el cambio notable que existe entre los dos primeros y los dos últimos años. Es posible que este profesional haya pasado de jornada parcial a completa y que se le hayan consistentemente asignado casos fáciles de resolver. En particular notamos que durante el 2017 este profesional fue capaz de procesar del orden de 4 casos por hora, es decir, un caso cada 15 minutos (nuevamente sin descansos ni quiebres de ningún tipo).

### 5.3.3 Comentarios Adicionales

Como otro detalle importante, se puede argumentar que el número de acogimientos de reclamos presentados a la SUSESO es una medida del error en el que incurren las aseguradoras. Es importante notar aquí que un error del orden del 50% puede ser considerado alto, pero es importante recordar el universo total del fenómeno, las licencias que presentan reclamo corresponden básicamente a un 10% del total de licencias, por lo tanto, un 50% de este 10% es básicamente un 5% de error que para cualquier tipo de análisis es más que aceptable puesto que implica que en el 95% de los casos las aseguradoras han hecho bien su trabajo. Conviene también recordar que estas licencias reclamadas son sólo una muestra de las licencias inicialmente rechazadas, varias de ellas son revertidas en los procesos intermedios de los cuales no se sabe nada.

El ejemplo anteriormente desarrollado, con la base de datos más nueva y con menos problemas con que se cuenta, ilustra el tipo de cosas que se pueden hacer con los datos. Sólo resta pensar cuáles serían las preguntas más típicas que se podría hacer un fiscalizador. En particular se vienen a la mente dos preguntas que la SUSESO como fiscalizador quisiera preguntar. La primera de ellas guarda relación con el número de licencias rechazadas por las aseguradoras. La segunda de ellas guarda relación con el número de licencias que un doctor en particular puede estar emitiendo.

Para la primera pregunta, lo que se puede hacer es considerar no sólo el número absoluto de casos, sino que también el volumen que le corresponde a cada aseguradora dentro del mercado (participación de mercado). Cuando se posee la participación de mercado de cada aseguradora, los números absolutos pueden ser transformados en tasas de aceptación/rechazo que permiten comparación entre distintas aseguradoras. A partir de ahí, se puede monitorear si es que las tasas de aceptación/rechazo mantienen consistencia en el tiempo. El típico análisis que después se puede realizar es ANOVA, que permite comparar el desempeño de variables categóricas. Hay que notar sin embargo que un test de este tipo no indica cuáles son los grupos que presentan diferencias en caso de que las haya y usualmente se debe proceder a realizar tests con pares de variables para determinar cuáles son estadísticamente iguales y cuáles no.

Otra cosa importante que observar es que las fluctuaciones que existen en los datos tienden a desaparecer cuando se consideran ventanas de tiempo mayores. Este hecho es muy conocido y debe ser considerado al momento de planificar experimentos para responder preguntas. La fluctuación que se observa por ejemplo entre observaciones diarias dista del tipo de fluctuación que se observa entre meses.

Por ejemplo, para efectos de ilustrar el tipo de resultados que un método como ANOVA trae, consideremos el caso anterior, y estudiemos si es que hay diferencias estadísticamente significativas entre grupos (acepta, rechaza, S/P). Si corremos el análisis ANOVA, veremos que el resultado es como sigue para 2017:

Anova: Single  
Factor

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Acepta	73	31633	433.3288	655970
Rechaza	73	27574	377.726	658783.5
S/P	73	370	5.068493	111.4814

#### ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	7917389	2	3958695	9.03217	0.000171	3.037667
Within Groups	94670279	216	438288.3			
Total	1.03E+08	218				

Este resultado indica que la hipótesis nula del test (no hay diferencias entre grupos) no puede ser apoyada por la evidencia presentada, por lo que no se puede aceptar esta hipótesis como cierta (notar el P-value de 0.000171). Es el resultado esperable puesto que el número de casos que no tienen pronunciamiento es sustancialmente distinto a los números de casos de aceptaciones y rechazos para todos los médicos. Cabe hacer notar que ANOVA indica si es que hay diferencias estadísticamente significativas entre grupos considerando además la varianza de los datos (o que tanto traslapan las distribuciones). Si se corre el mismo análisis, considerando sólo las categorías Acepta y Rechaza, se obtiene lo siguiente:

Anova: Single Factor

#### SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Acepta	73	31633	433.3288	655970
Rechaza	73	27574	377.726	658783.5

#### ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	112845.8	1	112845.8	0.171661	0.679257	3.906849
Within Groups	94662253	144	657376.8			
Total	94775098	145				

Este resultado indica que las dos categorías tienen comportamientos similares y que desde un punto de vista estadístico estos grupos son indistinguibles en su comportamiento. Esta simple ilustración muestra que aplicar métodos a diestra y siniestra no es tan simple como parece y más que explotar el conjunto de datos hasta la muerte vale más la pena concentrarse en preguntas para las cuales se puedan diseñar tests. Finalmente, y sólo por completitud, podemos ver que si corremos un



test de diferencia de medias para el 2017 entre los grupos Acepta y Rechaza se obtiene:

t-Test: Paired Two Sample for Means

	<i>Acepta</i>	<i>Rechaza</i>
Mean	433.3287671	377.7260274
Variance	655970.0015	658783.5072
Observations	73	73
Pearson Correlation	0.862012531	
Hypothesized Mean Difference	0	
df	72	
t Stat	1.115353467	
P(T<=t) one-tail	0.13420395	
t Critical one-tail	1.666293696	
P(T<=t) two-tail	0.268407899	
t Critical two-tail	1.993463567	

El resultado de este test indica claramente que las medias de los grupos son estadísticamente equivalentes (considerando para ello la varianza en torno a los datos), por lo tanto, desde un punto de vista estadístico, la cantidad de reclamos aceptados y rechazados con iguales.

Para la segunda pregunta, no es difícil ver que simples consultas en la base de datos pueden entregar información valiosa. Por ejemplo, cuando se agregan los registros por el número de licencias otorgadas por los médicos es posible ver lo siguiente (se han mostrado sólo las primeras filas de esta consulta):

Tabla 9: Muestra de reporte de médicos que emiten licencias

	2014	2015	2016	2017	TOTAL
RODRIGUEZ ALISTER ELIZABETH	23	88	213	251	575
LOEWENWATER QUILHOT IVES	0	1	200	286	487
VARGAS URREA JUAN	3	55	228	192	478
FERNANDO OYARZUN HERNANDEZ	13	36	257	171	477
MUNOZ RIQUELME OSCAR	0	51	307	77	435
ALVARADO HURTADO JOSE LUIS	22	63	225	103	413
GUZMAN ROJAS VICTOR	0	23	183	188	394
CUADRADO CHIRIBOGA ALI IVAN	0	102	237	41	380
OYARZUN HERNANDEZ FERNANDO	25	82	223	47	377
ESPEJO SAN CRISTOBAL CLAUDIO	8	39	137	173	357
SANCHEZ ESTEVES JUSTO XAVIER	0	15	91	246	352
SCHIAPPACASSE SAIEG MARCELO CLAUDIO INSC.	11	37	261	19	328
ROZAS MUNOZ CESAR	4	50	189	83	326
BRAVO SUDZUKI FRANCISCO	0	32	130	163	325
APPELT VEGA HERMANN	0	57	196	62	315
MU<U+FFFD>OZ RIQUELME OSCAR EDUARDO	0	41	195	79	315
ARENAS CONTRERAS LUIS	25	35	185	57	302
MARRO FREITTE PATRICIO	0	51	206	33	290
ZUNIGA HERRANZ FELIPE FERNANDO	0	9	126	153	288

En este reporte sólo se ha consultado por las licencias otorgadas por médicos y que han sido rechazadas y han presentado reclamos. Es claramente posible ver que los médicos que emiten licencias en general lo hacen año a año y si se presentan diferencias entre los dos primeros años y los dos últimos es sólo por las diferencias de volumen en los reclamos. Esta misma técnica puede ser utilizada en la base de datos con el total de licencias y establecer un corte que permita ir a fiscalizar a los médicos que estén sobre ese corte.

## 5.4 Conclusiones

El trabajo hecho en este capítulo muestra claramente que no es posible predecir con claridad quien presentará una reconsideración de un reclamo no acogido por la SUSESO. Básicamente, al nivel de lo que se puede hacer en estos momentos, se puede considerar la tasa histórica de reconsideraciones y estimar la carga de trabajo futura basado en este número. Por ejemplo, para el período 2014 – 2017 este porcentaje es de un 21% aproximadamente, pero decidir caso a caso se hace imposible sin una variable que permita separar los casos de manera adecuada. Una posible línea de acción futura podría consistir en tratar de incorporar los montos en juego a la base de datos y ver a partir de ahí si es que la clasificación mejora sobre la base de este factor adicional (que se asume a priori importante).

Una conclusión obvia guarda relación con el trabajo futuro necesario para poder obtener datos que permitan responder preguntas de manera más expedita. En particular, uno de los datos faltantes como parte del proceso guarda relación con los casos que se judicializan. Esta información existe, pero en otro lugar y no está disponible de manera adecuada en la base actual y por lo tanto no permite responder a esta pregunta.

Otra conclusión importante del análisis llevado a cabo tiene que ver con la forma en que se utilizan los indicadores que pueden ser determinados a partir de los datos. Hay que tener cuidado en la forma en que estos se definen y las preguntas que ellos tratan de responder puesto que ellos pueden ser una gran herramienta al momento de proceder a planificar fiscalizaciones apropiadas. La forma usual de poder mostrar estos indicadores es a través de un dashboard donde se muestran simultáneamente todos los indicadores y que a su vez permiten tener una idea global de la operación del sistema. Como los indicadores están asociados a preguntas que se desean responder de la operación del sistema y como estas preguntas cambian en el tiempo, es preferible que se utilicen herramientas del tipo Tableau de Microsoft o Spotfire de TIBCO, este tipo de herramientas facilitan de gran manera la confección de reportes y son una excelente forma de visualizar las anomalías, lo que a su vez puede ayudar a direccionar los esfuerzos fiscalizadores.

## 5.5 Apéndice

### 5.5.1 Agrupamiento de casos por médico

En esta subsección se detallan los resultados de la base de datos para cada médico en el período 2014 – 2017.

Tabla 10: Detalle de los resultados de reclamo por médico

Id Doctor	Acepta	Rechaza	S/P	TOTAL
1	602	66	2	670
2	8476	6387	201	15064
3	3344	3492	23	6859
4	160	635	1	796
5	2619	3743	40	6402
6	11	0	0	11
7	22	1	0	23
8	356	287	2	645
9	14	0	0	14
10	616	364	16	996
11	44	0	0	44
12	1684	223	11	1918
13	24	6	0	30
14	2	4	0	6
15	2182	568	19	2769
16	4600	3958	90	8648
17	2	0	0	2
18	0	4	0	4
19	14	0	0	14
20	190	57	0	247
21	819	519	38	1376
22	24	4	0	28
23	4	1	0	5
24	11	0	0	11
25	25	16	0	41
26	3	6	2	11
27	2853	3742	18	6613
28	0	0	1	1
29	722	2036	23	2781
30	6	4	0	10
31	21	6	0	27
32	383	250	29	662
33	740	561	57	1358
34	3850	4889	44	8783
35	1	0	0	1
36	2168	5112	21	7301
37	0	3	0	3
38	4980	8003	77	13060
39	2404	3097	21	5522
40	984	803	71	1858

41	3	0	0	3
42	1919	201	29	2149
43	7408	10319	28	17755
44	15	6	0	21
45	26	0	6	32
46	285	268	64	617
47	3	0	0	3
48	694	581	0	1275
49	26	0	0	26
50	11	0	0	11
51	6069	3941	111	10121
52	182	73	8	263
53	6	0	0	6
54	2	0	0	2
55	0	2	0	2
56	2	0	0	2
57	85	115	0	200
58	23	0	0	23
59	4025	4043	85	8153
60	2350	5224	47	7621
61	538	366	53	957
62	18	6	0	24
63	1689	1469	8	3166
64	0	0	2	2
65	55	0	0	55
66	1	0	0	1
67	47	0	0	47
68	1	0	3	4
69	17	0	0	17
70	43	0	0	43
71	7119	3786	161	11066
72	28	12	0	40
73	839	515	2	1356

## 6 Conclusiones

En este trabajo se abordó la problemática de utilizar de la mejor manera posible bases de datos existentes para responder cuatro preguntas, que por completitud formulamos nuevamente acá:

- ¿Es posible predecir cuál licencia rechazada presentará un reclamo a la SUSESO? (Caracterización de demanda)
- ¿Cuál es la mejor manera de asignar ese reclamo a un profesional cuando este sea presentado a la SUSESO? (Caracterización de oferta y optimización del uso de la oferta disponible)
- ¿Es posible sugerir un diagnóstico a un caso presentado a la SUSESO? (Reducción de la utilización de recursos, modificación de la función de oferta)
- ¿Cuáles casos no acogidos por SUSESO presentarán nuevos reclamos? (Predicción de demanda secundaria o flujo circulante)

Conviene recordar nuevamente acá que todo este trabajo gira en torno a la promesa de reducir los actuales 120 días promedio que son requeridos para dictaminar un reclamo a 30 días en el 95% de los casos. Claramente la forma de conseguir una reducción considerable pasa por una adecuada caracterización de la oferta, la demanda y finalmente un uso eficiente de los recursos disponibles. No es difícil observar que, si en un año típico hay del orden de 50.000 reclamos presentados a la SUSESO, entonces siempre es posible poner suficiente dinero para contratar una cantidad grande de profesionales que podrían ser asignados a cada caso en la medida de que estos lleguen y así poder conseguir su resolución en el mínimo tiempo posible. Lamentablemente, una estrategia de esta naturaleza no es eficiente, no es escalable y últimamente no utiliza medios tecnológicos disponibles en la actualidad para hacer uso eficiente del conocimiento existente y contenido en las bases de datos de la operación del sistema.

Desde esta perspectiva es que surge este proyecto, la idea central de él es explorar hasta que punto es posible utilizar los datos existentes para poder tomar decisiones y responder preguntas relativas al proceso y por otro lado el poder ver las limitaciones de los datos actuales y de qué forma dichas limitaciones podrían ser reducidas a futuro. Todo esto dentro del contexto del uso de herramientas de modelamiento matemático modernas, que permitan aprendizaje y ajuste de algoritmos, pertenecientes genéricamente al área de Business Intelligence (BI) o Data Science.

Gran parte del trabajo desarrollado no es visible en los entregables de este proyecto puesto que involucró mantener una comunicación fluida, abierta y constante con la contraparte técnica de la SUSESO. Todo el flujo actual de los procesos, la legislación existente, las particularidades de los distintos casos, las diferencias entre las licencias de origen común y los accidentes laborales por nombrar sólo algunos son una pequeña muestra de la complejidad que el equipo consultor tuvo que enfrentar. Bajo la misma óptica, esto también entrega una aproximación probablemente deficiente y subvalorada de la inmensa cantidad de trabajo que esta consultoría trajo a la contraparte técnica de la SUSESO. La calidad de este trabajo, en aquellas secciones donde esta es evidente es en gran medida gracias al esfuerzo de este grupo de

profesionales cuya dedicación y compromiso es admirable, las faltas y omisiones aún presentes en el mismo son muy probablemente culpa del equipo consultor y de antemano se pide disculpas por dichos errores involuntarios.

Un trabajo de esta naturaleza no siempre produce los resultados esperados puesto que hay un riesgo inherente al intentar cosas que nadie ha intentado previamente. En teoría los datos existen, se ingresan y algunos se miden, sin embargo, dependiendo de cuáles sean los controles asociados al ingreso de ellos es posible encontrarse en que la calidad de la información existente no permite producir un análisis que produzca los resultados esperados. Por otro lado, las complejidades del comportamiento humano no se pueden modelar de manera precisa, más aún, como algunos ejemplos muestran, en el contexto de esta consultoría es posible ver ejemplos donde los resultados desafían lo lógicamente esperable, a pesar de que la información es más confiable, lo que pone más de manifiesto lo difícil de tratar de encontrar regularidad en el comportamiento humano. Se espera que medidas tendientes a caracterizar apropiadamente ciertos aspectos del desempeño profesional de los evaluadores de casos pueda ayudar a eliminar parte de la incertidumbre actual, sin embargo, es claro que siempre existirán elementos humanos no medibles por lo que ningún modelo podrá ser capaz de obtener predicciones con un 100% de precisión. Un monitoreo continuo, modelos que se entrenan constantemente, que se actualizan en función de nueva información y un adecuado sistema de reconocimiento y premiación de desempeño pueden ser las estrategias que permitan hacer menos inciertos los comportamientos y tener un sistema más predecible.

La consultoría ha sido capaz de demostrar y concluir los logros resumidos en la Tabla 11.

## 6.1 Sugerencias para mejoras futuras

### 6.1.1 Mejora de las relaciones entre tablas de distintas bases de datos

Una constante en el desarrollo del trabajo fue el problema que resultó de tratar de conectar tablas de datos provenientes de distintos orígenes. En teoría hay campos que pertenecen a ambas tablas, por ejemplo, el RUT de los reclamantes, sin embargo, dicha relación no es uno a uno, sino que uno a varios, es decir, para un reclamante pueden existir varios reclamos. Por otro lado, en la base expedientes\_workflow, es difícil actualmente identificar que licencia en particular ha sido reclamada. Es posible utilizar ciertas heurísticas para definir cuáles lo han sido (y de hecho ese trabajo ha sido hecho por la contraparte técnica de la SUSESO), pero en general este trabajo no produce resultados 100% correctos y por lo tanto esta es un área donde es vital mejorar para que la calidad del análisis mejore puesto que se puede estar realizando el análisis y corriendo algoritmos en licencias que efectivamente no presentaron reclamos o viceversa.

Tabla 11: Resumen de Conclusiones y Recomendaciones

	Predicción y clasificación proactiva de reclamantes	Asignación Inteligente de Casos	Propuesta automatizada de Resolución de Casos	Predicción y clasificación de comportamiento post-atención
Conclusiones/Resultados	<ul style="list-style-type: none"> <li>• Para licencias médicas que presentan reclamos fue posible conseguir una clasificación correcta del orden de 85%</li> <li>• Las Redes Neuronales probaron no ser adecuadas para el tipo de datos que se posee dado el gran número de variables</li> <li>• Los modelos de regresión si bien corren son de difícil interpretación</li> <li>• Los modelos de ensemble permiten obtener buenos resultados en términos de error para las licencias que presentan reclamos, sin embargo tienen un error mayor en las que no reclaman</li> <li>• Para los accidentes laborales, la situación es totalmente opuesta, la calidad de los datos es mejor, lamentablemente no parece haber una lógica subyacente que permita una mejor clasificación</li> </ul>	<ul style="list-style-type: none"> <li>• Se demuestra que el uso de heurísticas no permite obtener el óptimo global</li> <li>• La diferencia entre la heurística propuesta y el modelo de optimización es de un 3% para el caso en estudio (no necesariamente generalizable)</li> <li>• La heurística depende de parámetros e información que se estima difícil de obtener</li> </ul>	<ul style="list-style-type: none"> <li>• Aparentemente no existe regularidad en los dictámenes, experimentos hechos con una regla de mayoría simple para casos vecinos entrega un porcentaje de error del orden de un 50%</li> <li>• El nivel de acuerdo entre diagnósticos para casos similares se propone como una medida de la complejidad del caso, así como también el tipo de resultado (acoge-rechaza)</li> </ul>	<ul style="list-style-type: none"> <li>• Los resultados de un árbol clasificador sobre el conjunto de datos entregan un pésimo clasificador con una tasa de acierto en torno al 50%</li> <li>• Mediante el uso de un ejemplo simple, se ilustró el uso de ciertas técnicas para resumir información y se resaltaron las diferencias que se obtienen al usar distintas escalas</li> <li>• No es posible responder a la pregunta de casos que se judicializan, la información no está disponible</li> </ul>
Recomendaciones	<p>Se recomienda utilizar modelos de ensemble, actualizar dichos modelos de manera frecuente y monitorear constantemente su desempeño</p>	<p>Tanto la heurística como el modelo de optimización requieren datos que no necesariamente están disponibles o son medibles. Se recomienda concentrar los esfuerzos primero en obtener los datos para posteriormente mejorar la heurística y perseguir optimalidad</p>	<ul style="list-style-type: none"> <li>• Se recomienda presentar al médico a cargo del caso resoluciones previas que presenten similitud con el caso actual, de esta manera se puede facilitar el trabajo</li> <li>• Se recomienda utilizar una medida de la complejidad de un caso de acuerdo al esquema propuesto basado en dificultad percibida y resultado probable</li> <li>• Se requiere que las variables categóricas sean codificadas utilizando variables dummies, varios análisis se beneficiarían de esta medida</li> </ul>	<ul style="list-style-type: none"> <li>• Se sugiere agregar información extra a la sábana de datos (monto en juego) para ver si esto mejora la clasificación</li> <li>• Se recomienda el uso de herramientas de Data Analytics tales como Tableau o Spotfire que permiten visualizar de manera fácil indicadores</li> <li>• Se sugiere concentrar los esfuerzos en definir las preguntas a responder para poder conseguir los datos adecuados</li> </ul>

### 6.1.2 Enlace a bases de datos proveniente de otros servicios

Uno de los problemas presentes en la base de datos y que no han permitido realizar un análisis más adecuado tiene que ver con la falta de datos socioeconómicos de

quienes presentan sus reclamos. En particular nos referimos acá a la ausencia de información relativa a los ingresos reales de quienes reclaman. En un mundo ideal, se esperaría que para una adecuada caracterización socioeconómica se pueda conectar con la base de datos del SII para así poder tener una idea de los montos involucrados en los reclamos.

Las particularidades de la legislación chilena, donde existe un tope imponible, hacen que la información relativa a ingresos sea difícil de aproximar si es que se mide por cotizaciones previsionales o inclusive mediante el uso de las cotizaciones de los planes de salud (al ser un porcentaje del sueldo hasta el tope imponible más eventualmente contribuciones voluntarias). La presentación de reclamos tiene ciertamente una componente económica detrás que en el presente trabajo no ha podido ser explorada puesto que no se cuenta con dicha información. Se sospecha que o bien los sueldos muy bajos o los sueldos muy altos tienen fuertes incentivos para colocar reclamos, cada caso por motivos diferentes puesto que los individuos que ganan un bajo sueldo necesitan dicho ingreso para una mínima subsistencia y por otro lado los que ganan demasiado pierden un monto considerable. Integrar las bases de datos actuales con otras fuentes de información puede entregar nuevas luces respecto de las motivaciones de quienes presentan reclamos y permitir entonces una mejor caracterización del fenómeno que permitirá implementar predictores más precisos.

Este mismo requerimiento también surge en el contexto de tratar de estudiar los casos que se judicializan. De momento si bien se sabe que la información existe, esta no está disponible como parte de las sábanas de datos a las cuales se tiene acceso.

### 6.1.3 Creación de una Línea de Tiempo de cada caso

En la actualidad los casos se ingresan a la SUSESO después de pasado un tiempo indeterminado que transcurre entre que la licencia es rechazada por la entidad aseguradora y el ingreso del reclamo. De acuerdo con el flujo actual del proceso de reclamo, cada rechazo viene acompañado de apelaciones a la entidad aseguradora, la COMPIN en caso de que la aseguradora sea privada y la apelación no sea fructífera y finalmente, en última instancia la SUSESO recibe el reclamo. Estadísticas recolectadas previamente en otras consultorías indican que este tiempo asciende al orden de algunos meses.

Lamentablemente, en la actualidad no sólo no se conocen de manera apropiada los tiempos que toman las etapas intermedias, sino que además se tiene el problema de que no se conocen los resultados de las apelaciones intermedias. Todo esto contribuye, a la luz del trabajo hecho en esta consultoría, a que sea dificultoso establecer una relación de causalidad más clara entre las licencias rechazadas y las reclamadas. Por este motivo es que es posible producir resultados de clasificación que giran en torno al 80% de acierto para las licencias que se reclaman, pero esto introduciendo el costo de clasificar de manera incorrecta licencias que no lo serán. Si se sigue esta regla, se crearán expedientes para licencias que nunca serán reclamadas en la SUSESO que eventualmente ya han sido acogidas en cuanto a sus reclamos en



instancias intermedias, esto último con el consiguiente costo que involucra el solicitar información a las entidades aseguradoras para la conformación de expedientes.

Se sugiere lo obvio que podría sugerirse en este punto y que no es más que tratar de obtener información relativa a las instancias intermedias del proceso para poder así retroalimentar a los algoritmos predictivos y tener la oportunidad de poder incrementar su precisión no sólo para los casos que efectivamente terminan en reclamo, sino que aquellos que no.

#### 6.1.4 Mejoras en la clasificación de las licencias de acuerdo con el Código de Diagnóstico

Los códigos CIE10 son fundamentales a la hora de poder planificar no sólo la asignación de trabajo, sino que además la composición del panel de especialistas evaluadores (tanto internos como externos). Llama la atención que en la actualidad dicho código presenta falencias en el sentido de que la categorización al parecer sólo está decentemente establecida al nivel de la familia, siendo de menor calidad en la subfamilia y finalmente el código correspondiente al padecimiento particular dentro de una familia-subfamilia es casi inexistente.

En la actualidad los códigos CIE10 son provistos por el servicio de salud o Isapre según corresponda (bajo el nombre código de diagnóstico en el formulario de licencia médica en la sección B). Sin embargo, se percibe un gran problema de calidad dada la información faltante y se sugiere que se establezca alguna medida que permita controlar de mejor manera el ingreso del código CIE10 en las bases de datos (un ejemplo de dicha medida es no permitir que un formulario sea ingresado sin que se registren la familia, subfamilia y código de enfermedad en el campo código de diagnóstico).

#### 6.1.5 Respecto del concepto de Expediente

Uno de los problemas observados con el proceso actual guarda relación con el concepto de expediente. En opinión del equipo consultor, si bien el concepto de expediente tiene cierta lógica desde un punto de vista médico, este podría predisponer a los médicos a evaluar de una cierta forma. Más aún, surge la legítima pregunta de si un médico debiese saber a qué paciente específico se refiere un caso dado puesto que puede sesgar su juicio. En particular, en la implementación actual los reclamos se presentan sobre distintos folios que son parte de un expediente lo que dificulta la determinación de cuáles casos específicos (asociados a un número de folio) se están reclamando. Esto dificulta el trabajo de identificar la verdadera cantidad de veces que un folio determinado es reclamado. Se sugiere para el futuro poder manejar folios separadamente y en otra tabla poder crear expedientes como una colección de folios. También se sugiere evitar el ruido que introduce el eventual sesgo de reclamos múltiples asociados al mismo RUT por la vía de no mostrar el historial asociado a un individuo.

#### 6.1.6 Creación de un modelo económico para los reclamos

Una de las preguntas centrales que se hiciera el equipo consultor durante la ejecución de los trabajos guarda relación con la racionalidad económica del proceso. Es decir, ¿Hasta qué punto vale la pena utilizar recursos en la gestión y resolución de casos si es que lo que se debe pagar por ellos es de hecho menor que el costo en el cual se incurre? Si bien no es un objetivo de este estudio, si es preciso entender que eventualmente los montos reclamados son bajos en comparación al costo involucrado en el proceso de reclamo y por lo mismo se podría considerar una resolución de tipo rápida para ciertos casos que cumplan con ciertas condiciones: jurisprudencia consistente (es decir casos similares acogidos con anterioridad), por lo tanto fáciles de resolver, y que los montos involucrados sean considerados menores a un cierto nivel a determinar en función de un modelo económico. Se entiende que los que deben pagar son las entidades aseguradoras, sin embargo, se podría considerar un subsidio a ese pago sobre la base de los principios establecidos anteriormente y de esta forma disminuir la carga de trabajo por casos de bajo impacto.

#### 6.1.7 Estrategias tendientes a eliminar posibles sesgos en los Dictámenes

Como se mencionó anteriormente, en la actualidad las revisiones son asignadas al mismo médico y esto podría presentar problemas de sesgo puesto que se estaría actuando como juez y parte en este proceso. Una posible forma de evitar esto sería asignar aleatoriamente revisiones de casos ya dictaminados a revisores externos para ver si es que la decisión a la que se llega es la misma o no. Esta inquietud surge principalmente a raíz de la falta de coherencia lógica observada en casos similares cuando se trata del resultado del reclamo (recordar que en el fondo la respuesta era similar a lanzar una moneda al aire con un 50% de errores de clasificación). Otra opción ya sugerida es hacer los casos anónimos, es decir, no permitir al médico que revisa ver antecedentes relativos a la persona que presenta el reclamo ni al doctor que emite la licencia. Se cree que si hay doctores que tienden a emitir demasiadas licencias, esto debe estudiarse de manera separada puesto que de lo contrario se puede caer en el caso de que se rechacen licencias que son válidamente emitidas solo a raíz de un sesgo proveniente o bien del paciente o del médico tratante.

#### 6.1.8 Caracterización de la fuerza de trabajo

Hay varias preguntas para las cuales aún no existe respuesta ¿Cuáles son los tiempos que efectivamente se necesitan para dictaminar? ¿Es posible caracterizar a cada profesional por separado en función de sus habilidades en distintos casos? ¿Cómo incorporar esto en la determinación de la dificultad de un caso? (observar acá que un caso fácil para un profesional puede ser difícil para otro).

Como ya se discutió previamente, la caracterización de la fuerza de trabajo está influenciada por dos factores básicos: rapidez y asertividad. Estos dos factores son contrapuestos y por lo tanto no es fácil balancearlos en una única definición, lamentablemente mejorar un objetivo conlleva empeorar el otro. Es de central importancia para la operación futura del sistema el poder contar con dicha caracterización, no sólo es una ayuda invaluable al momento de tomar decisiones que afectan la rapidez de la respuesta, sino que además permiten efectivamente planificar la adquisición de recurso humano para satisfacer demanda futura.

### 6.1.9 Codificación de variables categóricas como variables dummies

Muchas de las variables son categóricas y habitualmente hay que hacer alguna transformación o explícitamente decir que las variables son categóricas para poder utilizarlas en ciertos algoritmos. Lamentablemente, algunos algoritmos tienen limitaciones en cuanto a la cantidad de niveles que una variable categórica puede tener (por ejemplo, en el caso de árboles clasificadores en R estos no pueden ser más que 32 niveles). Otros algoritmos tales como Redes Neuronales, no aceptan variables categóricas y en tal caso hay que redefinir las variables categóricas usando variables dummies que esencialmente consisten de un uno si es que dicho registro toma el valor de esa categoría o cero si es que no. Como regla general, se esperaría que las variables categóricas sean especificadas mediante dummies, esto permite correr algoritmos sin tediosos procesamientos intermedios. Por ejemplo, para poder correr nuevamente el algoritmo propuesto en la etapa 3 con un conjunto de datos nuevo, hay que hacer una transformación nuevamente de todas las variables categóricas a dummies (para poder medir distancias) lo que indica la necesidad de poder contar con los datos en este formato para así poder facilitar análisis futuros.

### 6.1.10 Codificación de valores faltantes como algún valor por defecto

Parte importante de los problemas con los datos tuvo que ver con la no existencia de valores. Dichos valores cuando se importan a paquetes como R u otros se codifican como NA (Not Applicable). Cuando se corren los algoritmos, usualmente se eliminan los registros que contienen algún NA en ellos, esto afecta seriamente la disponibilidad de datos y no es posible pasar a producción sin tener una solución más definitiva a este problema.