R E V I S T A

INGENIERIA DE SISTEMAS

Volumen XXIX Septiembre 2015

Publicada por el
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

R E V I S T A
INGENIERIA DE SISTEMAS

ISSN 0716 - 1174

Financiado parcialmente por el Instituto Sistemas Complejos de Ingeniería.

Representante legal: Fernando Ordóñez
Correo electrónico: ris@dii.uchile.cl
Diagramación: Cinthya Vergara

Dirección: República 701, Santiago, Chile.
Web URL: www.dii.uchile.cl/∼ris
Portada: Gabriella Fabbri

Carta Editorial Volumen XXIX

Nos es muy grato presentar este nuevo número de la Revista de Ingeniería de Sistemas (RIS) dedicado a temas de frontera en Investigación de Operaciones, Gestión y Tecnología. Queremos agradecer al Instituto Sistemas Complejos de Ingeniería (ISCI) por su colaboración para hacer posible esta publicación.

Este número contiene artículos de académicos y estudiantes de nuestro Departamento de Ingeniería Industrial (algunos de ellos incluso son consecuencia de trabajos finales de grado, tesis de magister o tesis de doctorado), de investigadores del ISCI, de funcionarios del Servicio de Impuestos Internos y de académicos de la Universidad de Buenos Aires.

Nuestro objetivo a través de esta publicación es contribuir a la generación y difusión de las tecnologías modernas de gestión y administración. La revista pretende destacar la importancia de generar conocimiento en estas áreas, orientado tanto a problemáticas nacionales como a la realidad de países de características similares de la región.

Estamos seguros de que los artículos publicados en esta oportunidad muestran formas de trabajo innovadoras que serán de gran utilidad e inspiración para todos los lectores, ya sean académicos o profesionales, por lo que esperamos que esta iniciativa tenga la recepción que creemos se merece.

Guillermo Durán
Editor

Richard Weber
Editor Asociado

Llamado a Presentar Trabajos

La Revista Ingeniería de Sistemas (RIS) busca constituir un canal de divulgación de los avances en las áreas de Gestión de Operaciones, Tecnologías de Información e Investigación Operativa, que incluya los mundos académico y empresarial. Son particularmente apropiados artículos orientados a la práctica de estas disciplinas, que estimulen su uso o den cuenta de aplicaciones innovadoras de ellas, especialmente en América Latina.

También son bienvenidos artículos con análisis del estado del arte en un campo particular y de la forma en que los avances en dicho campo se han utilizado en la práctica.

Se espera que los artículos estén escritos de manera que puedan ser leídos por personas no especialistas en el tema tratado. Se recomienda incluir una lista de lecturas sugeridas para que los lectores no especialistas puedan profundizar en el tema.

Formato del Manuscrito

Los autores deben enviar un archivo en formato PDF del manuscrito que desean someter a referato a:

Comité Editorial Revista Ingeniería de Sistemas,
Departamento de Ingeniería Industrial,
Universidad de Chile.
Santiago, Chile.
Email: ris@dii.uchile.cl

Los manuscritos deben estar formateados para hojas tamaño carta, a doble espacio, márgenes de 2,5 centímetros en todos los lados, deben incluir un resumen de no más de 150 palabras y su extensión no debe exceder las 20 hojas.

La primera hoja debe contener el título del trabajo, nombre y dirección de los autores (teléfono y correo electrónico del autor de contacto), y un resumen de no más de 150 palabras.

Referencias

Las referencias se deben citar en el cuerpo del texto usando el nombre del autor y el año de publicación, e.g., Morton (1998). Al final del artículo se debe incluir la lista en orden alfabético de las referencias citadas en el texto. Para referencias de revistas científicas el formato es el siguiente: Autor(es), Año de publicación. Título. Nombre completo de la revista , Volumen e.g.:

Kodialam, M. y H. Luss, 1998. Algorithms for Separable Nonlinear Resource Allocation Problems. Operations Research , 44(2), 272-284.

Para referencias de libros el formato es el siguiente: autor(es), año de publicación. Título. Editorial, Ciudad; e.g.:

Kleinrock, L., 1975. Queueing Systems . John Wiley, New York.

En caso de haber más de una referencia con el mismo autor y año de publicación, se debe usar “a”, “b”, etc. como sufijo del año de publicación para diferenciarlas.

Detalles en www.dii.uchile.cl/∼ris

Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería

Sergio Celis [*]
Luis Moreno [*]
Patricio Poblete [*]
Javier Villanueva [*]
Richard Weber [*]

Resumen

En la última década el avance de los sistemas de gestión docente y sistematización de datos en educación superior han motivado el uso de herramientas de la minería de datos para entender procesos de aprendizaje y los contextos en los cuales estos ocurren. En el mundo anglosajón, comunidades en torno al *learning analytics* o el *educational data mining* han surgido para desarrollar áreas de investigación e intervención en educación superior. En estas comunidades, un área de particular interés es la generación de modelos predictivos de deserción y rendimiento académico que permitan intervenciones de apoyo temprano a los estudiantes. En este artículo hacemos uso de herramientas de *learning analytics* para construir un modelo que predice la caída en causal de eliminación, por motivos académicos, en estudiantes de primer año del Plan Común de Ingeniería y Ciencias de la Universidad de Chile. El modelo clasifica correctamente a más del 86 % de los casos, con niveles bajos de error tipo II, y una precisión de 38 %. Dado que se usa información hasta el inicio del segundo semestre, el modelo permite desarrollar intervenciones focalizadas en aquellos estudiantes en mayor riesgo.

Palabras Clave: Modelo predictivo, Rendimiento académico, Learning Analytics, Educational Data Mining.

[*]Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile.

1. Introdución

El enorme crecimiento en la disponibilidad de datos ha generado recientemente muchas oportunidades de aplicar métodos para el análisis de estos datos. El área de la educación no es una excepción. Analizando los datos que se genera entorno a la educación permite descubrir nuevas oportunidades para mejorar la gestión docente.

En este artículo describimos la aplicación de minería de datos aplicada a datos académicos de los estudiantes de ingeniería y ciencias de la Universidad de Chile y mostramos cómo la gestión docente puede anticipar - y posiblemente evitar - efectos negativos, como por ejemplo la doble reprobación de un curso que termina en la eliminación de la carrera y que es el fenómeno estudiado en el presente trabajo.

En la Sección 2 del artículo describimos el estado-del-arte del área de *learning analytics*. La Sección 3 describe la situación actual en una escuela de ingeniería. En la Sección 4 mostramos la construcción del modelo predictivo. Los resultados de la aplicación de nuestro modelo presentamos en la Sección 5. La Sección 6 concluye este trabajo y muestra posibles trabajos futuros.

2. Estado-del-Arte de Learning Analytics

A comienzos de este siglo, dos comunidades de investigación surgieron para usar herramientas matemáticas y computacionales para el análisis de datos educativos en educación superior: *Educational Data Mining* (EDM) y *Learning Analytics*. Ambas comunidades comparten el objetivo de usar la creciente recolección de datos en educación superior para mejorar los sistemas de evaluación, el entendimiento de los procesos educativos, y la priorización y diseño de intervenciones educativas [32]. Las diferencias entre ambas comunidades de investigación radica en los énfasis metodológicos y focos de investigación. En cuanto a metodología, mientras EDM privilegia el descubrimiento automatizado de patrones con poca intervención de juicio experto, *Learning Analytics* fortalece el juicio experto y testea hipótesis educacionales con ayuda de modelos de descubrimiento automático [4], tales como la selección de atributos [1]. Esto hace que el enfoque *Learning Analytics* sea más holístico y sistémico (p.ej. [27] que el enfoque basado en componentes individuales y la interacción entre ellos, característicos de la minería de datos [32]. En consecuencia, mode-

los generados por investigadores EDM son usualmente usados para desarrollar sistemas de tutoría inteligente, y los de *Learning Analytics* para apoyar la toma de decisiones de administrativos, profesores, y estudiantes. Sin embargo, ambas comunidades poseen límites porosos y múltiples convergencias entre ellas [32]. En lo que sigue, y sólo para efectos de este artículo, usaremos el concepto de *Learning Analytics* y lo entenderemos indistintamente a EDM.

Las tareas más comunes en *Learning Analytics* son clasificación, clustering, minería de textos, y visualización [22]. En cuanto a técnicas, las más usadas son árboles de decisión, redes neuronales, y redes Bayesianas [30]. Estás técnicas en contextos educacionales son frecuentemente complementadas con regresiones, correlaciones y otros técnicas estadísticas [30]. La principal fuente de datos para la investigación en *Learning Analytics* está en el uso de plataformas computacionales de aprendizaje, tales como sistemas de gestión de curso o CMS (*course management systems* en inglés) o sistemas de aprendizaje en línea [30], tales como los *Massive Open Online Courses* (MOOCs) [35]. Según Romero y Ventura [30] algunos de los problemas de investigación que concentran el interés de la comunidad de *Learning Analytics* son visualización de datos, retroalimentación a instructores, recomendaciones para estudiantes, predicciones de rendimiento de los estudiantes, modelos mentales de los estudiantes, y detección de comportamientos indeseados. En los últimos años su aplicación se ha extendido a otras áreas como el apoyo a metodologías activas basadas en problemas o proyectos [8], la toma de decisiones e intervenciones a nivel institucional (p.ej. [17], o el entendimiento de teorías del aprendizaje, tales como aprendizaje auto-regulado [29]).

Una contribución esencial del *Learning Analytics* a la línea clásica de teorías y modelos educacionales es que incorpora una nueva escala temporal a los procesos de aprendizaje. Si las teorías educacionales usan modelos invariantes en el tiempo o en largas etapas de desarrollo (en educación superior típicamente en semestres o años), las técnicas de minería de datos, son capaces de mostrar aprendizajes momento a momento [6]. Es decir, cambios en las capacidades de aprender, concentración y hasta estados de ánimo mientras el estudiante completa una evaluación en línea, trabaja en grupo, o interactúa con múltiples sistemas en el campus (p.ej. bibliotecas, unidades de tutoría académica). Más aún, esta información puede ser obtenida y procesada en tiempo real, permitiendo decisiones y acciones inmediatas o en el corto plazo [6]. De acuerdo a Berland et al. [8], *Learning Analytics* "permite una rigurosa, replicable, y precisa descripción del comportamiento de los estudiantes, así como también un análisis de cómo estos comportamientos interactúan con otros constructos de interés. El comportamiento de los estudiantes puede ser monitoreado en cuanto crece y cambia en el tiempo" (p. 211, traducción propia)

Otra emergente área de investigación en *Learning Analytics* es la combinación de datos institucionales con información proveniente del juicio humano. Esta combinación se realiza, por ejemplo, con instructores usando aplicaciones que evalúan el trabajo de los estudiantes en sala o talleres [6].

3. Retención y Rendimiento Académico en el Primer Año de Educación Superior

En las últimas décadas, la deserción en educación superior se ha transformado en un asunto prioritario de política educacional, tanto a nivel institucional como gubernamental. El impacto negativo de la deserción es relevante tanto porque los aranceles aumentan, como por el significado que socialmente ha adquirido la educación superior, entendida hoy como una instancia clave de desarrollo personal, social, económico y cultural. Se estima que en Chile la deserción al tercer año es cercana al 40 %, con una gran variabilidad según el tipo de institución (p.ej., universitaria, institutos profesionales, y centros de formación técnica) y áreas disciplinarias [31]. Por ejemplo, según Rolando et al. [28], en la cohorte del 2008 que ingresó a carreras profesionales, un 38 % desertó en el primer año en institutos profesionales, mientras que sólo un 14 % en universidades. De acuerdo al estudio Retención en Educación Superior con Perspectiva de Género [24] se evidencia que desde la cohorte 2007 hasta el año 2010 existió un aumento de la tasa de retención desde un 67 % a un 71 %, sin embargo, para el año 2013 ésta disminuyó a un 69 %. Según área disciplinar, la retención en los programas académicos en las áreas de tecnologías está entre las más bajas del país. En promedio, sólo un 65 % de los estudiantes permanece en sus programas luego del primer año [23].

La investigación de la persistencia y deserción tiene una larga historia en naciones con desarrollados sistemas de educación superior. Un gran número de estudios ha identificado los factores críticos que explican la persistencia y deserción. Parte importante de la complejidad a la que se ven enfrentados estos estudios, es la definición operacional de la deserción. Existen distintos tipos de deserción (p.ej., voluntaria o involuntaria; de transferencia o abandono), las cuales son registradas en diferentes tiempos (semanas, semestres, años), y que pueden ser transitorias o permanentes. Una discusión conceptual sobre las definiciones de la deserción en Chile puede ser consultada en [16].

Pascarella y Terenzini [26] revisaron más de tres décadas de este tipo de investigaciones, principalmente aquellas realizadas en Estados Unidos. Entre los resultados de su investigación, proponen un listado extenso de factores y

mecanismos que influyen en la persistencia y deserción, entre los que destacan características individuales de pre-ingreso y características institucionales. Las características individuales de pre-ingreso a la institución de educación superior tienen un consistente y estadísticamente significativo efecto en la persistencia. Al respecto, estudios previos han identificado la habilidad académica, status socioeconómico, grado de motivación, y expectativas de logro. Más aún, estas características académicas y sociales de los y las estudiantes tienen mayores efectos que las características institucionales en la persistencia y deserción. Otro factor importante es el ingreso retrasado a la educación superior, es decir, el tiempo que transcurre desde que el o la estudiante termina la educación secundaria, hasta que se matricula en alguna institución de educación superior. Estudios anteriores también muestran que deserciones previas tienen un efecto negativo en las chances de persistencia. Las características institucionales han recibido gran atención dado que pueden ser controladas por las instituciones y por políticas públicas. Aquellas características que han mostrado mayor impacto son la selectividad de las instituciones, incluso controlando por factores obvios como la habilidad académica de los y las estudiantes; su integración al campus y sus participaciones en actividades extracurriculares; actividades del primer año que introducen a estudiantes a la vida académica; becas para estudiantes de bajos ingresos; interacciones con profesores y profesoras fuera de la sala de clase; y la interacción entre pares. El rendimiento académico, las notas, es el mejor predictor de la persistencia, con un mayor efecto durante los dos primeros años de estudio. En relación a las diferencias disciplinarias, estudiantes en carreras de las ciencias, tecnologías, ingeniería y matemática (STEM en inglés), tienen una mayor tasa de deserción que estudiantes en otras disciplinas. Es importante mencionar que la mayoría de los factores ya discutidos interactúan con características sociodemográficas de los estudiantes, tales como etnicidad y sexo.

Algunos autores han propuesto constructos no observables para explicar la persistencia y deserción. Los modelos teóricos más influyentes son los modelos de deserción de Bean y el proceso de deserción de Tinto. Bean [7], basado en estudios de rotación organizacional, construye y testea un modelo de análisis de trayectorias causales para la deserción. En este modelo, la identificación de un o una estudiante con la institución, la certeza en la decisión de carrera, valores instrumentales (por ejemplo, creencia en que la educación es fundamental para conseguir un buen trabajo), y la intención de abandonar son factores que median los efectos de variables individuales, organizacionales y ambientales. Desde otra perspectiva teórica, los modelos de Tinto se basan en los estudios sobre suicidio de Durkheim y en los estudios de ritos de transición en sociedades tribales de Van Gennep. Tinto [34] extiende previos modelos de deserción,

proponiendo tres estados en la trayectoria de los y las estudiantes en educación superior: separación, transición, e incorporación, que son críticos en las decisiones de continuar o abandonar. Ambos, los modelos de Bean y Tinto, han sido consistentemente confirmados a través de estudios cuantitativos [9]. En la última década, investigadores han testeado estos modelos, analizando datos longitudinales y usando técnicas estadísticas más avanzadas. El estudio pionero de DesJardins, Ahlburg y McCall [15] basado en datos longitudinales de la University of Minnesota, arrojó que las variables definidas por estudios previos, tales como los descritos anteriormente, afectan la deserción, pero en magnitudes diferentes, según los años en la carrera. Por ejemplo, la locación de la residencia de origen tuvo un efecto significativo en la deserción en los tres primeros años de carrera, y la edad de ingreso sólo en los dos primeros. Numerosos estudios han continuado usando éstas y otras técnicas estadísticas para entender con mayor profundidad los fenómenos relacionados con la deserción (p.ej., [11, 12, 18, 20, 33]).

En Chile, también se han comenzado a testear estos modelos y a utilizar sofisticados métodos cuantitativos para entender la deserción en el sistema nacional de educación superior. Acuña [2] y Larroucau [19], basados en datos nacionales del sistema secundario y universitario chileno, confirman que el fenómeno de la deserción es multicausal y que las variables discutidas anteriormente tienen validez en el contexto local. Específicamente, Larroucau [19] encontró que en las características individuales de pre-ingreso, tales como el establecimiento de origen y el promedio de notas y ranking en la enseñanza media, eran mejores predictores de la deserción que el puntaje PSU. Mizala, Hernández, y Makovec [25] estiman la probabilidad de deserción en las carreras de pedagogía. Sus resultados confirman a la habilidad académica (medida por puntaje PSU) como uno de los factores más influyentes en la deserción, efecto que sería moderado por el quintil socioeconómico del estudiante. Díaz [13] y Celis [10] calcularon modelos de duración con datos de las carreras de ingeniería de la Universidad Católica de la Santísima Concepción y la Universidad de Chile, respectivamente. Ambos estudios muestran que el tipo de establecimiento de educación media impacta en la deserción. Celis [10] muestra que estudiantes provenientes de colegios particulares tienen menores tasas de deserción en los últimos años de la carrera que aquellos provenientes de la educación pública. Díaz [13] encontró que a mayor puntaje en la PSU y a mayor ingreso familiar, menores son las chances de deserción. El estudio mostrado en [31] usa la técnica de *propensity score matching* para estudiar el impacto de los créditos y becas en la persistencia. Los resultados sugieren una asociación positiva de los créditos y becas de excelencia en la persistencia. El Centro de Microdatos de la Universidad de Chile [14], mediante una

encuesta, determinó que las principales causas de deserción en el primer año universitario se deben a problemas vocacionales (p.ej., no quedar en la carrera de preferencia), situación económica familiar, y rendimiento académico. Recientemente, herramientas estadísticas tradicionales de la minería de datos también han comenzado a usarse para analizar la deserción y otras variables educacionales [21] (ver [3] para un caso aplicado en una universidad chilena).

En resumen sabemos que hay factores previos al ingreso, características individuales, y condiciones de vida y académicas que influeyen en la retención de primer año. Además sabemos que las carreras de ingeniería y ciencias tienen promedios altos de deserción en el primer año universitario. Sin embargo, muchas de estas investigaciones se han realizado en naciones con sistemas de educación superior desarrollados. Más investigación es necesaria para entender el fenómeno de la deserción en Chile, en especial en carreras de ciencia e ingeniería. Aquí es donde *Learning Analytics* brinda oportunidades no solo para entender empíricamente la deserción, sino que también para generar modelos predictivos que permitan generar alertas tempranas e intervenciones que le brinden apoyo oportuno a estudiantes en riesgo de deserción o de insuficientes desempeños académicos. A continuación se presenta un modelo predictivo desarrollado para detectar bajos rendimientos académicos en el primer año del Plan Común de las carreras de ingeniería y ciencias de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile.

4. Construcción del modelo predictivo

4.1. Situación actual en la Facultad de Ciencias Físicas y Matemáticas (FCFM)

La FCFM es una unidad académica altamente selectiva, con una alta producción científica y sofisticados sistemas de gestión docente relativo al contexto regional y latinoamericano. La población estudiantil es cercana a los 4.900 estudiantes de pregrado, seleccionados del 3 % superior de la enseñanza media de acuerdo al la Prueba Nacional de Selección Universitaria (PSU). La FCFM la componen además cerca de 1.200 estudiantes de postgrado y 220 profesores de jornada completa, de los cuales un 97 % posee un grado de doctor. La FCFM ofrece 9 carreras de ingeniería, geología y tres licenciaturas científicas. Todos los estudiantes de pregrado ingresan a un Plan Común de dos años de duración. Actualmente, el primer año está estructurado en dos semestres. En el primer semestre los estudiantes son asignados en siete secciones con similares

capacidades académicas según ranking de ingreso. Todos los estudiantes tienen los mismos ramos en el primer semestre: introducción al cálculo, introducción al álgebra, introducción a la física newtoniana, introducción a la ingeniería, química, y herramientas computacionales para ingeniería y ciencias. En total, la carga académica suma 30 SCT (Sistema de Créditos Transferibles), lo que equivale a 50 horas de trabajo semanal durante 15 semanas. En general, los estudiantes aprueban el 85 % de los cursos inscritos en primer año. En las últimas dos décadas, la FCFM ha venido realizando sostenidos esfuerzos para mejorar las tasas de retención y el rendimiento académico de los estudiantes. Por ejemplo, el 2007 se realizó un cambio curricular que implicó un giro hacia estrategias de enseñanza centradas en el estudiante, además de importantes mejoras de infraestructura y el lanzamiento de nuevas unidades de apoyo docente y al estudiante. Actualmente las tasas de retención de primer año son cercanas al 95 %. Pese a que este indicador es muy superior a las carreras de ingeniería y tecnología a nivel nacional (en torno al 65 %), la FCFM está empeñada en seguir mejorando esta tasa, consciente de la gran calidad académica de los estudiantes que recibe y de que el pequeño grupo que no persiste luego del primer año representa un desafío particular. El estudio aquí descrito se circunscribe en estos esfuerzos. Así, el objetivo de esta investigación es usar la información personal y académica disponible de los estudiantes para detectar estudiantes en riesgos de abandonar el plan de estudios. Para tales efectos se usó información histórica para generar y calibrar un modelo predictivo que permitiese la instalación de un sistema de alertas tempranas que le de soporte a los estudiantes que más lo necesiten. Para la construcción del modelo predictivo se usaron datos de las cohortes de ingreso 2010, 2011, 2012, 2013, y 2014. A continuación se presenta el modelo predictivo en sí, discutiendo la variable dependiente, las variables independientes consideradas, y la construcción del modelo.

4.2. Variable dependiente

En la primera fase del estudio se decidió acotar la variable dependiente a la doble reprobación de al menos un curso del primer semestre. Esta definición se justifica en dos ideas importantes. Primero, la doble reprobación de un curso es causal de eliminación de los estudiantes, que a la vez afecta negativamente las tasas de retención de primer año. Aunque un alumno en causal de eliminación puede elevar una solicitud especial para rendir un curso por tercera vez y proseguir en la Escuela, estas solicitudes requieren un esfuerzo no menor en la gestión docente y un porcentaje importante de estos alumnos igual termina eliminado de la Facultad. La segunda razón tiene un

argumento metodológico. La deserción en un lugar como la FCFM, así como en otras escuelas de ingeniería, es multidimensional y diversa. Tal como se indicó en la revisión de la literatura sobre deserción, las razones van desde lo económico (p.ej., falta de financiamiento) pasando por crisis vocacionales, situaciones excepcionales, hasta rendimiento académico. Así focalizarse en las causas académicas (las cuales no están necesariamente disociadas del resto), permite darle mayor precisión al modelo, al menos conceptualmente. La Tabla 1 muestra la distribución de la reprobación para las poblaciones estudiadas. Dado que reprobar al menos un ramo es condición necesaria para la reprobación de un ramo por segunda vez, la población de estudiantes considerada para esta investigación se reduce a entre 195 a 255 estudiantes por cohorte, que son los que reprobaron por lo menos un curso en su primer semestre.

Tabla 1: Reprobación y Doble Reprobación en Primer Año

Año Ingreso	Cohorte Ingreso [1]	Al menos 1 curso reprobado 1^{er} semestre	Doble reprobación 2^{do} semestre[2]
2010	687	195 (28 %)	43 (24 %)
2011	720	220 (31 %)	26 (14 %)
2012	704	213 (30 %)	41 (21 %)
2013	700	255 (36 %)	26 (11 %)
2014	762	216 (28 %)	27 (14 %)
Total	3.573	1.099 (31 %)	163 (17 %)

(1) Número de estudiantes que se mantuvieron activos durante el primer semestre.
(2) El porcentaje corresponde a estudiantes que reprobaron por segunda vez algún ramo de primer semestre sobre el total de estudiantes que reprobaron al menos un ramo de primer semestre y se mantuvieron activos durante el segundo.

4.3. Variables independientes

Las variables independientes (o atributos) consideradas fueron seleccionadas basado en la revisión de la literatura y la información disponible. Así las variables independientes se dividen en tres grupos: características individuales, variables de pre-ingreso y variables de rendimiento académico. En cuanto a características individuales sólo incluimos género, tiempo desde el egreso de la enseñanza media y región de procedencia. En variables de pre-ingreso usamos tipo de establecimiento de enseñanza media (i.e., particular, subvencionado, público emblemático y público no emblemático), experiencias previas en educación superior, puntajes en la PSU, vía de ingreso (i.e., PSU o ingresos

especiales), ranking y promedio de notas en la enseñanza media. Finalmente se construyeron otras once variables (continuas, ordinales y binarias) basadas en información detallada sobre las notas parciales de los estudiantes en los dos primeros semestres de la población objetivo. Dentro de aquellas variables podemos mencionar ratio de créditos aprobados versus reprobados, variación de notas del primer al segundo semestre tanto en ramos aprobados como reprobados, y diferencias con la nota mínima de aprobación, la cual en este caso es 4, dónde 1 es la mínima y 7 la máxima.

La decisión de cuánta información académica incluir en el modelo merece mayor discusión. En nuestro caso, la doble reprobación ocurre al final del segundo semestre del primer año. Mientras antes en el año se detecten aquellos estudiantes en riesgo de doble reprobación, mejor, ya que existiría mayor tiempo de intervención y reacción por parte del estudiante. Por otro lado el tiempo le otorga mayor información al modelo predictivo lo que aumenta su precisión. En un extremo, si usamos información academica de todo el primer año se logra una predicción perfecta, es decir sin errores de clasificación. En un primer momento nos propusimos estimar el modelo sólo con información del primer semestre. Los resultados no fueron satisfactorios ya que si bien nuestras predicciones fueron sustantivamente mejores que el azar, se obtuvo un alto número de errores del tipo I (falso negativos) y del tipo II (falso positivos) (estos resultados pueden ser solicitados a los autores). El mejor escenario siguiente se esquematiza en la Figura 1. Tradicionalmente, las asignaturas de primer año del Plan Común realizan tres pruebas parciales (localmente conocidas como controles) y un examen final. Tal como muestra la figura, el modelo predictivo final fue construido con información recolectada hasta la primera ronda de los controles 1 del segundo semestre. Esto deja varias semanas (un 75 % del semestre) para intervenir y dos controles y el examen final para recuperarse.

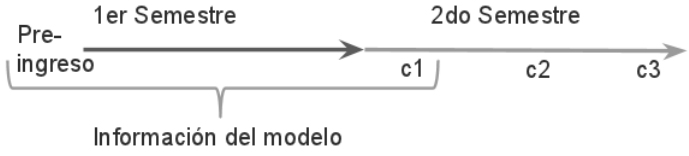


Figura 1: Tiempo de captura de información académica para el modelo

4.4. Construcción del Modelo

En cuanto al modelo predictivo, se utilizó un modelo de regresión logística en combinación con una metodología de selección de atributos, debido a la simplicidad de interpretación y utilización ampliamente aceptada. Selección de atributos puede ser considerada parte de la fase de pre-procesamiento o de minería de datos, su objetivo es encontrar el subconjunto de atributos con mayor valor predictivo, evitando así utilizar variables que agreguen ruido en la fase de entrenamiento, mejorando la predicción y acelerando así el proceso de adaptación de los modelos. Entre los enfoques más utilizados se encuentran:

- Forward Feature Selection (FS): Se comienza sin atributos en el modelo, se agregan una a una las variables y se evalúa bajo cierta métrica el desempeño de agregar cada variable, eligiéndose, de ellas, la que mejore más el desempeño (si es que hubiese mejora). El proceso se repite hasta que ninguna variable mejora el rendimiento al ser agregada.
- Backward Elimination (BE): En este enfoque se comienza con todos los atributos, luego se evalúa la eliminación de cada variable, eliminándose efectivamente la variable con mayor aumento de desempeño al ser eliminada (si es que alguna lo mejora). El proceso se repite hasta que ninguna mejora sea posible.

La metodología propuesta de selección de atributos consiste en una mezcla entre FS y BE en conjunto a una selección por frecuencia. En particular, se comienza realizando el proceso FS, tras agregar un atributo, se realiza el proceso BE. Esto con el fin de eliminar atributos ya agregados que posean mayor ruido, es decir, se pueden haber incluido nuevos atributos que, en conjunto con algunos de los atributos previamente agregados, mejoran la predicción y eliminan ruido de un atributo ya agregado.

La metodología híbrida entre FS y BE se realiza mediante validación cruzada (Cross−Validation) con cierta cantidad de conjuntos, los que van variando entre entrenamiento y validación [5]. Esto con el fin de obtener resultados representativos en cuanto al valor predictivo de cada atributo, evitando así posibles sobreajustes.

Debido a la reducida cantidad de datos que se posee en comparación a la cantidad de atributos, se combina toda la metodología previamente propuesta con una selección por frecuencia, es decir, se realiza un gran número de veces la selección de atributos y se lleva conteo de los atributos seleccionados. Finalmente se consideran como atributos seleccionados aquellos que posean una cantidad de selecciones mayor a un umbral previamente determinado.

Una vez ocurrida la selección de atributos, se utiliza un modelo de regresión logística el cual se ajusta a variables dependientes del tipo binaria (Long, 1997). La Ecuación 1 más abajo describe la función logística, donde Y representa la variable dependiente, en este caso la doble reprobación, $X_1, ..., X_n$ las variables independientes seleccionadas mediante el proceso de selección de atributos, β_0 el parámetro constante, y $\beta_1, ..., \beta_n$ los parámetros del modelo. Al estimar aquellos parámetros es posible realizar predicciones acerca de la doble reprobación basado en las variables independientes.

$$ln(\frac{Y}{1-Y}) \quad = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{1}$$

Para estimar los parámetros se utilizó la información recopilada para las cohortes de ingreso de 2010 a 2013. Luego, se usó el modelo obtenido para predecir el comportamiento del universo objetivo de la cohorte de ingreso 2014. En otras palabras, el modelo fue entrenado con las cohortes 2010-2013 y puesto a prueba con la información obtenida para la cohorte de ingreso 2014. El poder predictivo del modelo fue evaluado mediante dos reconocidos indicadores *recall* y *precision* (ver las ecuaciones más abajo), donde TP=true positive, FP = false positive, y FN=false negative , las respectivas tasas.

$$Recall \quad = \quad \frac{TP}{TP + FN} \tag{2}$$

$$Precision \quad = \quad \frac{TP}{TP + FP} \tag{3}$$

Se puede interpretar el *recall* como la tasa de los verdaderos positivos, es decir la tasa de los positivos que el modelo detecta como positivo, mientras la *precision* es la tasa de los predichos como positivo que realmente son positivo.

5. Resultados

El proceso de selcción de atributos arrojó siete variables independientes: género, tipo de establecimiento de enseñanza media, ratio de créditos reprobados, promedio controles 1 del 2do semestre menor promedio final del 1er semestre en cursos reprobados (etiqueta: C1IRS2 < FIRS1, tipo: variable binaria), diferencia con nota 4.0 del peor promedio actual en cursos reprobados en 2do semestre (C1IRS2 - 4.0, variable continua), promedio controles 1 del 2do semestre menor promedio final del 1er semestre en cursos no reprobados (C1INRS2 < C1INRS1, variable binaria), y el peor promedio controles 1 del

2do semestre menor que el peor promedio final del 1er semestre en cursos no reprobados (minC1INRS2 < C1INRS1, variable binaria). Para el caso del tipo de establecimiento de enseñanza media, partimos distinguiendo entre establecimiento público emblemático y público no emblemáticos. Luego que esta diferenciación no produjo cambios estadísticamente significativos, decidimos mantener esta variable en las tradicionales tres categorías: privado, subvencionado, y público.

La Tabla 2 muestra los resultados de la regresión logística. El test de *likelihood ratio* indica que el modelo se ajusta de buena manera a los datos (LR Chi-cuadrado = 201,62, p < 0,001). Es decir, las variables independientes tienen poder explicativo sobre el evento de doble-reprobar una asignatura. La variable que tiene el mayor poder explicativo es sin duda el ratio de los créditos inscritos reprobados. Este resultado no debiese sorprendernos. A mayor número de cursos reprobados en el primer semestre, mayor son las probabilidades de reprobar un curso por segunda vez. El poder explicativo de esta variable es tal, que se podría aplicar la heurística: si un alumno que reprueba dos o más cursos en su primer semestre, tendrá altas probabilidades de volver a reprobar al menos uno de ellosen el segundo semester. Por ejemplo, un estudiante que reprueba álgebra, cálculo, y física tiene aproximadamente cinco veces más probabilidades de doble reprobación que un estudiante que sólo reprueba una de esas asignaturas. Dos otras variables mostraron una relación estadísticamente significativa con la doble reprobación. Una de ellas es género. Un estudiante hombre tiene 88 % más probabilidades de doblereprobar que una mujer, ceteris paribus. La otra variable significativa es la diferencia entre el promedio de los primeros controles de los cursos ya reprobados y la nota de reprobación 4.0. Esto indica que aquellos estudiantes que superen la nota de aprobación en los primeros controles tienen menores probabilidades de volver a reprobar una asignatura que aquellos que no. Por ejemplo, un estudiante con promedio 3.0 en los controles 1 de las asignaturas reprobadas tiene 31 % más probabilidades de reprobar que aquel con nota 4.0, ceteris paribus.

Si bien, el resto de las variables independientes seleccionadas no son estadísticamente significativas en el modelo, el signo de los coeficientes se comporta dentro de lo esperado y es consistente con la literatura nacional. Por ejemplo, estudiantes proveniente de establecimientos de enseñanza media particular o subvencionada tienen menores probabilidades de doble reprobación que aquellos provenientes de establecimientos municipales. Los coeficientes de las tres variables binarias de rendimiento académico que no son estadísticamente significativas para el modelo también se comportan en el sentido esperado. Si el promedio de los primeros controles del segundo semestre en las asignaturas cursadas por segunda vez es menor que el promedio final de las

asignaturas reprobadas en primer semestre (C1IRS2 < FIRS1), existe una mayor inclinación a doble reprobación. Lo mismo sucede si los estudiantes bajan sus calificaciones en los controles de los ramos no reprobados en el segundo semestre con respecto al primero (C1INRS2 < C1INRS1 y minC1INRS2 < C1INRS1). Esto último es interesante ya que el modelo considera también el desempeño en aquellas asignaturas aprobadas y cursadas por primera vez.

Etiqueta	Coef.	Std. Err.	Odd Ratio
Género (hombre)	0,63**	0,30	1,88
colegio particular	-1,63	3,00	0,19
colegio subvencionado	-2,24	3,00	0,10
ratio creditos reprobados	4,41***	0,62	82,41
C1IRS2 < FIRS1	0,13	0,40	1,14
C1IRS2 - 4.0	-0,38**	0,13	0,69
C1INRS2 < C1INRS1	0,19	0,67	1,21
minC1INRS2 < C1INRS1	0,53	0,63	1,70
_cons	-3,77	3,01	
Log likelihood = -252,63			
Df = 8			
LR chi2(8) = 201,62 ***			
*p<0,1, **p<0,05, ***p<0,01			

Tabla 2: Resultado de Regresión Logística: Doble Reprobación en Primer Año (n=830)

Como se señaló en la sección anterior, el modelo fue estimado sólo con datos de las cohortes 2010-2013. Los datos de la cohorte de ingreso 2014 fueron usados para probar el poder predictivo del modelo. A cada estudiante ingresado el 2014 y con al menos una asignatura reprobada en el primer semestre, se le calculó una probabilidad de doble reprobación con información obtenida hasta la primera ronda de controles del segundo semestre. La Figura 2 muestra el resultado de esa simulación. En el eje horizontal se ubican todos los estudiantes, desde aquellos con la más alta probabilidad de doble reprobación hasta aquellos con más baja probabilidad. Dado que la variable dependiente es binaria, es necesario fijar un umbral o un porcentaje dónde aquellos con probabilidad mayor se les asigna el valor 1, la doble reprobación, y aquellos con probabilidad menor al umbral se les asigna cero o no doble reprobación. El umbral de alumnos predichos en la figura representa ese punto. El umbral fue decidido empíricamente como aquel valor de umbral donde se interceptan las curvas de sensitividad y especificidad (i.e., dónde se optimiza la correcta clasificación de casos positivos y negativos), en este caso 19 %.

Los colores representan la doble reprobación real de los estudiantes al final del segundo semestre del año 2014. Aquellas columnas en color rojo (oscuro)

representan los estudiantes que efectivamente presentaron una doble reprobación, aquellos con color verde (claro) los que no. En la figura, se evidencia que todos los que efectivamente presentaron doble reprobación, con la excepción de dos, son efectivamente predichos por el modelo. De hecho, el *recall* es 0,86, es decir el modelo clasifica correctamente a 12 de los 14 casos positivos, es decir un 86 % de las veces. Este resultado es sobresaliente en el contexto de predicciones sobre enseñanza y aprendizaje y da respaldo para generar intervenciones tempranas de apoyo. Sin embargo, al mismo tiempo el modelo clasifica incorrectamente casos negativos (i.e., falsos positivos). En total, el modelo predice 32 casos como positivos de los cuales solamente 12 son realmente positivos, lo cual da una *precision* de 37,5 %. Sin embargo, el número de falsos positivos es tolerable para el tipo de intervenciones y decisiones a tomar en base a los resultados del modelo.

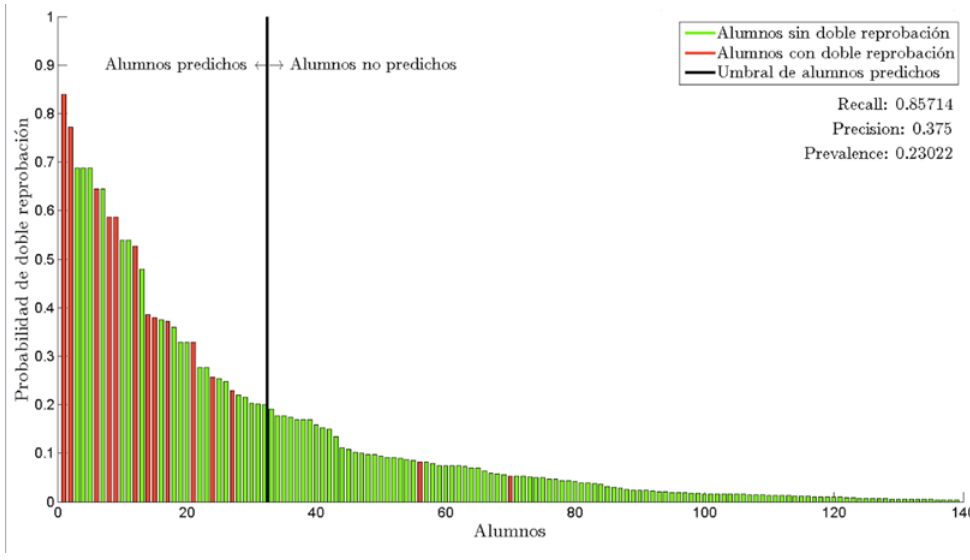


Figura 2: El Poder Predictivo del Modelo

6. Conclusiones y trabajo futuro

Este estudio tiene el objetivo de mostrar cómo herramientas de *learning analytics* pueden ser usadas para generar modelos predictivos que sirvan para apoyar a aquellos estudiantes en riesgo de deserción o de insuficientes desempeños académicos. Para tales efectos usamos datos institucionales y académicos históricos de cinco cohortes de ingreso al Plan Común de Ingeniería y Ciencias de la Universidad de Chile. Con estos datos se generó un modelo con

un alto poder prodictivo. El objetivo se centró en predecir tempranamente a aquellos estudiantes con riesgo de reprobar un mismo curso por segunda vez, lo cual los deja automáticamente en causal de eliminación. Los resultados del modelo son notables de acuerdo a su poder predictivo. Estos resultados sirvieron de base para que los directivos de la Escuela de Ingeniería y Ciencia apoyaran una serie de intervenciones de apoyo, desde comunicaciones personalizadas a los estudiantes y reforzamientos periódicos y tutorías académicas para alumnos en riesgo. Estas intervenciones se aplicarán a partir del semestre de primavera 2015.

Además el estudio abrió puertas no sólo para generar un modelo con aplicaciones prácticas, sino que también para ganar en entendimiento y generar nuevas preguntas acerca del fenómeno de la deserción y rendimiento académico. Un resultado interesante y consistente con previos estudios (p.ej., [10]), es que las estudiantes mujeres exhiben un mejor rendimiento académico que los hombres. Este resultado requiere mayor examinación, ya que otorgaría luces para la promoción de la mujer en disciplinas científicas e ingenieriles. Actualmente, tenemos en marcha un estudio que usa métodos mixtos de investigación para entender la experiencia de las estudiantes mujeres en primer año, no solo de aquellas en riesgo de caer en causal de eliminación sino que a través de todo el espectro de rendimiento académico en primer año.

Otro estudio interesante que surge a partir de este trabajo es el de aprender desde los estudiantes error tipo I, es decir aquellos a los cuales el modelo les asigna una alta probabilidad de doble reprobación pero que terminan aprobando todas sus asignaturas. ¿Qué antecedentes personales, prácticas de estudio, y actitudes determinaron el desempeño mejor de los esperado? Es una pregunta desde la cual se pueden guiar futuras políticas de intervención o la simple promoción de estrategias efectivas de estudio. En este caso, también estamos conduciendo una investigación que indaga en la experiencia de estos estudiantes para entender desde sus perspectivas cómo fue el proceso de aprendizaje durante sus primeros años de Plan Común.

En futuras investigaciones nuestro equipo está realizando esfuerzo para traer variables desde otros tipos de experiencias académicas dentro del modelo. Un área atractiva y de mayor tradición en la emergente área del *learning analytics* es aquella que estudia el uso de los sistemas en línea de gestión docente o CMS. En la FCFM el CMS local no solamente es usado para acceder información docente sino para generar discusiones, debates y otro tipo de interacciones no académicas. De algún modo entendemos que el uso de los estudiantes de estas plataforma es un indicador de su compromiso con sus planes de estudio y la vida universitaria. Estas variables nos aportarían nuevas dimensiones al análisis.

En resumen, aquí demostramos que con herramientas sencillas de *learning analytics* es posible generar modelos predictivos que permitan robustecer las decisiones curriculares y de intervención a nivel docente y administrativo. Además este estudio contribuye al entendimiento del despeño académico de los estudiantes de ingeniería y ciencias en universidades nacionales, áreas disciplinares que debiesen tener a su alcance las capacidades de usar inteligencia de datos para aprender más y ser más eficaces en los procesos educativos.

Agradecimientos: Este trabajo fue financiado por el proyecto Basal "Diseño de un sistema de información para el monitoreo, evaluación y mejoramiento continuo de la docencia de pregrado"(UCH1298) y el Instituto Sistemas Complejos de Ingeniería (ICM: P-05-004-F, CONICYT: FB016).

Referencias

[1] A. Acharya y D. Sinha. Application of feature selection methods in educational data mining. *Journal of Computer Applications*, 103(2):34–38, 2014.

[2] C. Acuña Veliz. Acceso y deserción en la educación superior: caso aplicado a Chile. *Tesis de Magíster. Universidad de Chile, Santiago, Chile*, 2012.

[3] F. Angulo y E. Sergio. Modelo para la automatización del proceso de determinación de riesgo de deserción en alumnos universitarios. *Tesis de Magíster. Universidad de Chile, Santiago, Chile*, 2012.

[4] P. Baepler y C. J. Murdoch. Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, pages 1–9, 2010.

[5] B. Baesens. Analytics in a big data world. *John Wiley and Sons*, 2014.

[6] R. S. Baker. Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent Systems*, pages 78–82, 2014.

[7] J. P. Bean. Student attrition, intentions, and confidence: Interaction effects in a path model. *Research in Higher Education*, 14(4):291–320, 1982.

[8] M. Berland, R. S. Baker, y P. Blikstein. Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, 19(1-2):205—220, 2014.

[9] A. F. Cabrera, A. Nora, y M. B. Castañeda. College persistence: Structural equations modeling test of an integrated model of student retention. *Journal of Higher Education*, 64(2):123–139, 1993.

[10] S. Celis. Student attrition and student time-to-degree at a selective engineering school in Chile. *Documento interno, Escuela de Ingeniería y Ciencias, Universidad de Chile*, 2012.

[11] R. Chen. Institutional characteristics and college student dropout risks: A Multilevel Event History Analysis. *Research in Higher Education*, 53(5):487–505, 2012.

[12] R. Chen y S. L. DesJardins. Exploring the effects of financial aid on the gap in student dropout risks by income level. *Research in Higher Education*, 49(1):1–18, 2007.

[13] C.J. Díaz. Factores de deserción estudiantil en ingeniería: Una aplicación de modelos de duración. *Información Tecnológica*, 20(5):129–145, 2000.

[14] Centro de Microdatos. Estudio sobre causas de la deserción universitaria. *Departamento de Economía, Universidad de Chile*, 2008.

[15] S. L. DesJardins, D. A. Ahlburg, y B. P. McCall. An event history model of student departure. *Economics of Education Review*, 18(1):375–390, 1999.

[16] E. Himmel. Modelos de análisis de la deserción estudiantil en la educación superior. *Calidad de la Educación*, 17:91–107, 2002.

[17] P. Jia y T. Maloney. Using predictive modelling to identify students at risk of poor university outcomes. *Higher Education*, 70:127–149, 2014.

[18] I. Johnson. Enrollment, persistence and graduation of in-state students at a public research university: Does high school matter? *Research in Higher Education*, 49(8):76–793, 2008.

[19] T. Larroucau. Estudio de los factores determinantes de la deserción en el sistema universitario chileno. *Tesis de Magíster. Universidad de Chile, Santiago, Chile*, 2013.

[20] S. A. Lesik. Do developmental mathematics programs have a causal impact on student retention? an application of discrete-time survival and regression discontinuity analysis. *Research in Higher Education*, 48(5):583–608, 2007.

[21] J. Luan, T. Kumar, S. Sujitparapitaya, y T. Bohannon. Exploring and Mining Data. in: R.D. Howard, G.W. McLaughlin, W.E. Knight (eds.). *The Handbook of Institutional Research. San Francisco, CA: Jossey-Bass*, pages 478–501, 2012.

[22] T. Martin y B. Sherin. Learning analytics and computational techniques for detecting and evaluating patterns in learning: An introduction to the special issue. *Journal of the Learning Sciences*, 22(4):511–520, 2013.

[23] SIES Ministerio de Educación. Retención de primer año en educación superior. programas de pregrado, 2014.

[24] SIES Ministerio de Educación. Retención en educación superior con perspectiva de género, 2014.

[25] A. Mizala, T. Hernández, y M. Makovec. Determinantes de la elección y deserción en la carrera de pedagogía. *Proyecto FONIDE N° F511059*, 2011.

[26] E. T. Pascarella y P. T. Terenzini. How college affects students. *San Francisco, CA: Jossey-Bass*, 2, 2005.

[27] R. Pea. The learning analytics workgroup: A report on building the field of learning analytics for personalized learning at scale. 2014.

[28] R. Rolando, J. Salamanca, A. Lara, y C. Blanco. Deserción y reingreso a la educación superior en Chile: Análisis de la cohorte 2008. *SIES, Ministerio de Educación*, 2012.

[29] I. Roll y P. H. Winne. Understanding, evaluating and supporting self-regulated learning using learning analytics. *Journal of Leaning Analytics*, 2(1):7–12, 2015.

[30] C. Romero y S. Ventura. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[31] V. Santelices, X. Catalán, C. Horn, y D. Kruger. Determinantes de deserción en la educación superior chilena, con Énfasis en efecto de becas y créditos. *Proyecto FONIDE N° F611103*, 2013.

[32] G. Siemens y R. S. J. Baker. Learning analytics and educational data mining : Towards communication and collaboration. 2012.

[33] L. D. Singell y G. R. Waddell. Modeling retention at a large public university: Can at-risk students be identified early enough to treat? *Research in Higher Education*, 51(6):546–572, 2010.

[34] V. Tinto. Stages of student departure: Reflections on the longitudinal character of student leaving. *Journal of Higher Education*, 59(4):438–455, 1988.

[35] C. Ye y G. Biswas. Early prediction of student dropout and performance in moocs using higher granularity temporal information. *Journal of Learning Analytics*, 1(3):169–172, 2014.

Una arquitectura de software para la provisión continua de servicios de salud en ambientes ubicuos: Aplicaciones de Semantic Web y BPM

Matías Echeverría [*]
Ángel Jiménez-Molina [*]
Sebastián Ríos [*]

Resumen

Los pacientes con enfermedades crónicas requieren una atención médica continua, personalizada y anticipativa. El paradigma de la computación ubicua es un enfoque que permite hacer realidad esta visión. Sin embargo, el uso de servicios de salud ubicuos e información contextual del paciente es bastante limitado. Para abordar este desafío, este artículo propone un framework basado en Web semántica para la provisión continua de servicios ubicuos, donde las necesidades de los pacientes se representan por medio de procesos de negocios. Además, se propone un modelo de descripción semántica de los procesos, servicios, información médica y contexto, con el fin de facilitar la selección de los servicios adecuados para los procesos de negocio. El framework se evalúa a través de un caso de estudio para enfermedades crónicas respiratorias que hace uso de datos de pacientes reales, y con un estudio de usabilidad aplicado a profesionales de salud de un hospital público pediátrico.

Palabras Clave: Ubiquitous Health, Semantic Web, Context-Awareness, Web Services, Business Process Management.

[*]Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile.

1. Introducción

Las enfermedades crónicas son un desafío para los sistemas públicos de salud de todo el mundo. Se estima que el costo de los tratamientos requeridos podría alcanzar el 80 % del presupuesto de salud en muchos países [10]. Además, estas enfermedades suelen ser de larga duración y de progresión lenta. Sólo en 2008, 36 millones de personas murieron a causa de condiciones crónicas en el mundo.

Las enfermedades crónicas más comunes son, dentro de otras, las patologías cardiovasculares, el stress y la depresión, la diabetes e hipertensión, la demencia, los problemas de obesidad, la apnea del sueño obstructiva y las complicaciones respiratorias [2, 4, 10].

Varios países han coincidido en la necesidad de hacer frente a este desafío por medio de un servicio de salud más preventivo, personalizado y anticipativo. Ésto requiere que los pacientes crónicos sean protagonistas de su salud, asumiendo una mayor responsabilidad para mantenerse en estados estables y así no colapsar la red de atención pública. Lo anterior implica hacer realidad la visión de una provisión de atención médica continua al paciente, entregando el servicio de asistencia sanitaria en todas partes y en cualquier momento [1].

El paradigma de la computación ubicua (CU) es un enfoque que busca hacer realidad visiones como la anterior. Sus principios consisten en la provisión invisible y poco invasiva de servicios al usuario en momentos oportunos. Ejemplo de ello en el dominio de la salud ubicua (u-health) es el servicio de monitoreo remoto de pacientes, en que por medio de biosensores y dispositivos adosados al cuerpo de una persona es posible capturar y analizar en tiempo real sus signos vitales independientemente de su ubicación geográfica. Estos datos constituyen información de contexto que permite predecir riesgos de crisis antes que éstas ocurran, alarmando a tiempo al paciente, su cuidador y al conjunto de actores de la red de salud que tienen relación con el usuario (médicos, paramédicos, enfermeras, secretarias, etc.).

Este artículo propone un framework basado en Web Semántica para la provisión continua de servicios u-health a pacientes con enfermedades crónicas. Este framework está centrado en el paciente, es decir, los servicios u-health se proporcionan desde la perspectiva de la persona y en menor grado de la tecnología. Esto se realiza representando los servicios u-health que requieren los pacientes mediante procesos de negocio. Ejemplo de tales procesos son los procedimientos clínicos definidos por la autoridad sanitaria, principalmente los Ministerios de Salud, para diagnosticar, tratar o controlar la patología del paciente. Tales procedimientos por lo general carecen de una representación en

procesos y por lo tanto difícilmente son reutilizados. Además, en caso de existir apoyo tecnológico a las actividades de tales procedimientos, por lo general es adhoc y poco sofisticado. Lo anterior limita la reusabilidad de aplicaciones tecnológicas, la calidad de los diagnósticos y la anticipación a crisis.

El framework se estructura en tres niveles de abstracción: una capa de procesos de negocio, una capa mediadora compuesta por un conjunto de actividades coordinadas, y una capa de servicios Web. Este framework se complementa con un modelo semántico que describe los procesos de negocio, sus actividades, los servicios Web y la información de contexto médico y ambiental. Contar con un modelo semántico de este tipo facilita la creación de diversas instancias de aplicaciones en función de la información contextual. La razón es que este tipo de modelos fomenta en el proceso y en los desarrolladores de servicios la reutilización e integración de una diversidad de recursos.

Los beneficios del framework incluyen la personalización de aplicaciones en función del estado del paciente y de la información de contexto; la independencia de dispositivos y servicios Web específicos, toda vez que lo relevante es la funcionalidad de éstos, pudiendo ser reemplazados en cada instanciación de un proceso en una aplicación; la reusabilidad de procesos y servicios Web para diferentes pacientes y aplicaciones respectivamente; la orquestación dinámica de los servicios Web en función de la lógica de coordinación de las actividades contenidas en el proceso y su aplicabilidad en diversas patologías crónicas.

La efectividad del framework se evalúa a través de un caso de estudio de enfermedades crónicas respiratorias en un hospital público pediátrico de Santiago. Se utilizan signos vitales históricos de diferentes pacientes crónicos, y se expone a los profesionales del hospital a los servicios Web desarrollados. Este estudio intenta clasificar el nivel de riesgo del paciente, con el fin de anticiparse a un estado potencial de crisis de salud, como es habitual en este tipo de pacientes en la estación invernal. Con fines ilustrativos, lo anterior se realiza mediante la aplicación de un modelo de clasificación basado en razonamiento difuso. El nivel de riesgo de crisis se entiende en este trabajo como información de contexto de alto nivel, en función de la cual se seleccionan procesos de negocios – por ejemplo un procedimiento clínico para tratar obstrucción respiratoria – y se mapean hacia los servicios Web que satisfacen los requerimientos de las actividades.

Por otro lado, se muestra un conjunto de servicios Web desarrollados para la gestión de la información del paciente una vez que éste, su cuidador y el médico han sido notificados de una posible crisis de salud. Se evalúa la usabilidad de estos servicios permitiendo que los profesionales del hospital interactúen con ellos.

La Sección 2 de este artículo introduce el trabajo relacionado. La arqui-

tectura del framework se describe en la Sección 3. La Sección 4 presenta el modelo semántico, mientras que la implementación y evaluación del framework se discute en la sección 5. El artículo se concluye en la sección 6.

2. Trabajo Relacionado

El enfoque que más se acerca a la provisión continua de atención médica es el monitoreo remoto de pacientes, que ha tenido un éxito parcial en hacer realidad tal visión. Se define como la medición continua o periódicamente frecuente de las características fisiológicas del paciente. La principal limitación de los casos existentes es que están diseñados y construidos desde una perspectiva técnica y no centrada en el paciente, constituyendo aplicaciones aisladas, especializadas en una o pocas enfermedades y para un grupo específico de pacientes.

Por otro lado, los frameworks que proveen servicios u-health en base a datos de biosensores u otros dispositivos, presentan el problema que se desarrollan de manera ad hoc. Por lo tanto, de ser necesaria la implementación de un nuevo biosensor en una aplicación, se debe desarrollar desde cero la capacidad de soportarlo. Esto significa que, dada la falta de flexibilidad de los frameworks existentes, las aplicaciones no se pueden obtener en tiempo de ejecución.

Recientemente, han aparecido servicios de salud autónomos que incorporan inteligencia artificial, como machine learning, que asisten a profesionales de salud en la interpretación de datos médicos y en la toma de decisiones [3]. Tales servicios implementan algoritmos de minería de datos sobre bioseñales para agrupar pacientes, predecir sus estados de salud o realizar pre-diagnósticos médicos mediante reglas lógicas de primer orden [13].

Por otro lado, la aparición de servicios de salud que resultan de la composición estática de servicios Web y su despliegue en dispositivos móviles, ha permitido un fácil acceso y una provisión proactiva de información médica en cualquier lugar y momento [5]. Tal avance ha requerido la integración de diferentes tecnologías, sistemas e infraestructuras de comunicación [14]. En este sentido, la literatura muestra que estos sistemas tienden a evolucionar hacia arquitecturas de software orientadas a servicio (SOA por su nombre en Inglés), que ofrece flexibilidad para la integración e interoperabilidad [11].

En esta línea, el Health Level Seven Group y el Object Management Group han juntado esfuerzos para promover la interoperabilidad dentro de organizaciones de salud que estén en búsqueda de implantar el enfoque SOA, creando el Proyecto de Salud de Especificación de Servicios (Healthcare Services Speci-

fication Project en Inglés). El objetivo de este proyecto es la generación para el sector salud de estándares SOA que definan el comportamiento de los distintos servicios e interfaces [8].

El trabajo relacionado enfocado al uso de la CU en este tipo de sistemas es escaso. En efecto, si bien el término servicio u-health ha sido acuñado para identificar los servicios de salud computacionales provistos en ambientes de CU [9], se ha hecho poco esfuerzo en crear plataformas lo suficientemente flexibles para desarrollar, compartir y reusar de manera efectiva diversos servicios. Un caso cercano a lo que se propone en este artículo pero aún insuficiente es la arquitectura que proponen Han et al., la cual permite que se puedan registrar y utilizar características comunes durante el desarrollo de servicios u-health [5], tales como la obtención, gestión y análisis de diversos tipos de datos fisiológicos del paciente, como la extracción de conocimiento a partir de ellos y el apoyo a la toma de decisiones médicas.

Entre los pocos modelos basados en la combinación de CU y SOA se encuentra el trabajo de Giuli Paganelli et al. [12], quienes además proponen una ontología de la información de contexto. Los recursos principales de esta ontología son: localización del paciente, tipos y valores de datos fisiológicos, actividad en que se encuentra involucrada la persona, síntomas, patología, red de cuidado del paciente, entre otros. El sistema propuesto por estos autores incorpora en la ontología aspectos claves del proceso, como el razonamiento acerca del contexto del paciente, y una apropiada política de alertas preventivas.

Siguiendo un enfoque similar, otros autores han propuesto ontologías para sistemas de manejo remoto de salud y detección de alertas, los cuales de manera explícita integran en las ontologías minería de datos, en conjunto con el juicio experto de los médicos, pero sólo estableciendo lineamientos generales sin alcanzar una lógica acabada [13].

En síntesis, el mayor problema de las soluciones existentes para la provisión contínua de atención médica es que utilizan una asociación estática entre las aplicaciones de salud y los servicios u-health. La principal limitación es que la provisión de estos servicios sólo puede ser realizada de acuerdo a la forma en que fueron configurados en el momento de diseño.

3. Arquitectura del Framework

Los principales elementos del framework consisten en el *context manager*, el *process manager* y el *repository system*. En el *context manager* existe el módulo *raw data collector*, que es un listener que captura señales generadas por diferentes sensores, las cuales deben estar previamente suscritas al *context manager*. Los sensores pueden generar datos fisiológicos, ambientales, geográficos, etc. Además, el *raw data collector* puede clasificar los distintos tipos de datos, o hacer una simple fusión de éstos. La información generada es procesada por el *high-level context generator*, que de acuerdo a la naturaleza de los datos recibidos, genera un contexto de alto nivel a través de la activación de diferentes servicios contextuales. La Sección 5.2 ejemplifica este proceso para dos tipos de contexto de alto nivel: el nivel de riesgo de un paciente para caer en una crisis, y la imprecisión de la predicción de un modelo para evaluar el nivel de riesgo. En este framework encapsulamos tales servicios de contexto en servicios REST, los cuales son invocados por el *high-level context generator*.

El *process manager* es el encargado de seleccionar los procesos de negocios en función del contexto de alto nivel generado en el *context manager*. Ésto se realiza a través de los siguientes módulos: *context listener*, *properties matchmaker*, *semantic measurer* y el módulo *process selector*. El primero es un listener basado en eventos que suscribe tipos específicos de contexto de alto nivel, y activa al módulo *properties matchmaker* para que éste haga un match entre el contexto y los valores de las propiedades de la ontología de procesos residente en el *repository system* que se explica en la Sección 4.1.

El *process manager* obtiene la definición del proceso seleccionado desde el *repository system*. Esta definición consiste en un archivo BPEL (Business Process Execution Language) que describe la lógica de coordinación de cada proceso haciendo uso de patrones secuenciales, concurrentes o iterativos, y de diferentes tipos de compuertas lógicas, tales como OR-split, OR-join, X-OR-split, AND-join, AND-split, etc.

El archivo BPEL es procesado por el *interpreter* del *service engine*. Este módulo recorre el archivo para extraer las actividades que componen el proceso, además de su información: variables de entrada y salida, condiciones previas y variables de efecto, definiciones funcionales, URIs (Uniform Resource Identifier) y parámetros de calidad de servicios. La Sección 4.1.1 muestra un ejemplo simplificado de una definición en BPEL.

La información anterior extraída por el *interpreter* es procesada por el

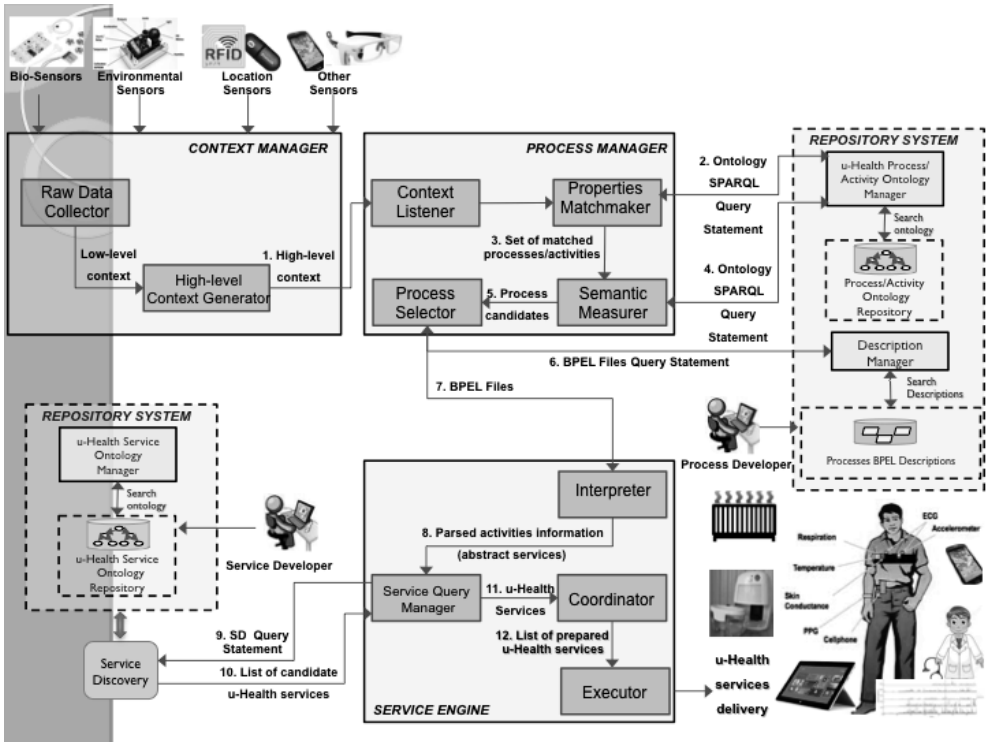


Figura 1: Visión General de la Arquitectura del Framework basado en Web Semántica para la Provisión Continua de Servicios u-health

service query manager, el cual define una serie de servicios abstractos para satisfacer los requerimientos funcionales y calidad de servicio de cada actividad. Luego, este módulo genera una sentencia de consulta hacia el *service discovery*, encargado de buscar servicios apropiados y disponibles en la ontología de servicios, de una forma similar a la realizada por el *properties matchmaker*.

Una vez que se descubren estos servicios, el *coordinator* los asocia con las actividades del proceso. La asociación servicio-actividad no necesariamente es uno a uno, es decir, una actividad podría requerir un conjunto de servicios compuestos entre sí. Posteriormente, el *coordinator* envía una lista de operaciones de servicios dispuestos en conformidad con la lógica de coordinación establecida por la definición del proceso. Finalmente, los servicios son activados y ejecutados por el módulo *executor*.

El *repository system* tiene un *ontology manager* y un *description manager*, los cuales contemplan funciones de almacenamiento, modificación, eliminación y actualización de ontologías y definiciones. En particular, los desarrolladores de procesos interactúan con el *description manager* para crear y almacenar procesos de negocio recientemente generados, de la misma forma que los desa-

rrolladores de servicios interactúan con el *ontology manager* para hacer lo mismo con los servicios.

4. Modelo Semántico

En esta sección se introduce el modelo semántico que describe procesos de negocio, ilustrado en la Figura 2. Este modelo representa la información de contexto involucrada en la provisión de servicios, a través de signos vitales obtenidos desde bio-sensores, registros de salud del paciente, perfiles de los stakeholders, condiciones del entorno, entre una serie de otros recursos.

El contexto se modela en tres aspectos: médico, personal y de entorno. El contexto médico de un paciente incluye recursos que representan la información de su ficha médica, así como los signos vitales, riesgo de crisis y síntomas medidos o inferidos en un momento dado. El recurso persona se especializa en múltiples roles y perfiles tales como el paciente, médicos, cuidadores de salud, enfermeras, secretarias, etc.

El servicio de atención médica debe entregarse en función de las características contingentes del entorno, las cuales podrían afectar el estado actual y futuro del paciente. Por ejemplo, la localización geográfica de los stakeholders es crítica para identificar las actuales condiciones ambientales que pueda estar enfrentando el paciente, la distancia de un médico o un cuidador en caso de emergencia médica, o la enfermera más cercana y con disponibilidad para asistir un evento en caso que el paciente esté hospitalizado, etc. También, el nivel de temperatura actual, humedad o luminosidad del lugar pueden ser útiles para inferir la necesidad de activar actuadores para mejorar tales condiciones, en caso de restricciones del entorno que puedan ser definidas con anterioridad para un paciente. Por ejemplo, en el caso de un paciente con hospitalización domiciliaria debido a una enfermedad respiratoria crónica, podría ser necesario activar automáticamente un deshumidificador para disminuir los niveles de humedad en la habitación en que se encuentra.

En síntesis, se podría desarrollar, por ejemplo, un proceso de negocio para notificar cuidadores, médicos o enfermeras sobre el nivel de riesgo de crisis de un paciente que se encuentra con hospitalización domiciliaría, así como también las actividades requeridas para gestionar tal condición. Un proceso como este es útil siempre y cuando el *context manager* infiera, basado en algún modelo predictivo, que el paciente está en un alto nivel de riesgo de crisis dados sus actuales signos vitales. Por lo tanto, es claro que este proceso debe ser insertado como una instancia en la ontología de procesos, tomando en cuenta, por ejemplo, que las propiedades de *localización*, *enfermedad crónica*,

modelo predictivo, *nivel de riesgo*, entre otras, necesitarían ser ajustadas a valores tales como *habitación del paciente*, *enfermedad respiratoria crónica*, *modelo de razonamiento difuso* y *nivel de riesgo alto*, respectivamente.

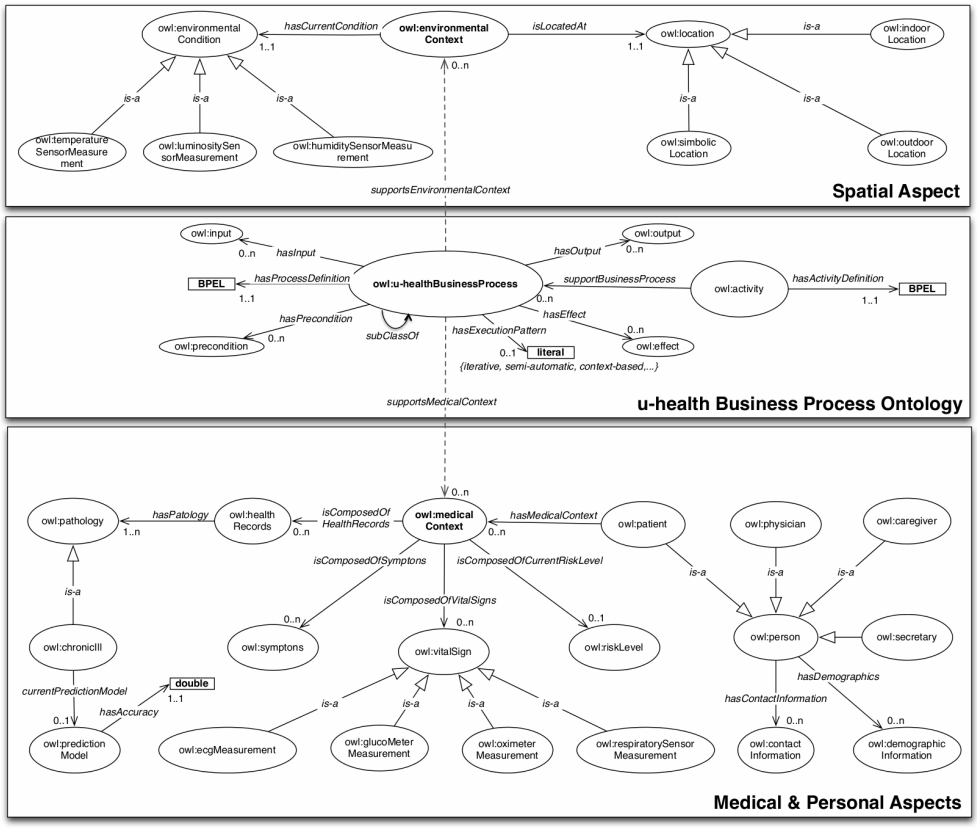


Figura 2: Extracto del Modelo Semántico que Describe Procesos de Negocio

4.1. Mecanismo de Selección de Procesos

El mecanismo de matching descrito en la Sección 3 y que ejecuta el módulo *properties matchmaker*, se implementa como una consulta en el lenguaje estándar SPARQL. La consulta consiste en identificar el número de valores de las propiedades de los nodos de la ontología de procesos que tengan un matching total o parcial con los valores de la información de contexto. Posteriormente, el módulo *semantic measurer* clasifica los procesos en términos de su valor de matching, generando un conjunto de procesos candidatos. En caso de resultar procesos con el mismo valor de matching, se determina su posición en la ontología de procesos, la cual se estructura de manera jerárquica con relaciones de subclases entre sus nodos, y se selecciona la subclase más

especializada. Este paso se explica en trabajos previos [6, 7], y esta fuera del alcance de este artículo.

4.1.1. Ejemplo de Proceso: Notificar a los Cuidadores de Salud

En esta Sección se provee un ejemplo concreto de un proceso de negocio orientado a notificar a los cuidadores sobre el nivel de riesgo de crisis del paciente. En efecto, siempre que el contexto consista en el nivel de riesgo, identificación y patología del paciente, el *process selector* extrae desde la ontología aquellos procesos cuyas propiedades *hasPatology* del recurso *healthRecords* e *isComposedOfCurrentRiskLevel* del recurso *medicalContext* calcen exactamente con el contexto. Debido a que podría ser una situación crítica, no se acepta un matching parcial en este ejemplo.

Al aplicar en el framework el contexto anterior, se seleccionan los siguientes procesos:

- Cuando el riesgo del paciente es *sin riesgo*, cualquier proceso de negocio es seleccionado.
- Cuando el riesgo del paciente es *bajo riesgo*, el *process selector* extrae los procesos etiquetados como *notificar bajo riesgo al cuidador de salud* y *proveer recomendaciones de cuidados básicos*, dentro de otros. El primero tiene una definición de procesos representada en un archivo BPeL, el cual es recorrido por el módulo *interpreter*, obteniéndose una lógica de coordinación compuesta de dos actividades secuenciales, a saber, *obtener el cuidador del paciente* y *obtener la información de contacto del cuidador del paciente*, seguido por un patrón paralelo con una compuerta OR-SPLIT de tres ramas – *enviar SMS*, *enviar e-mail*, y una tercera rama compuesta por una secuencia para *obtener el médico del paciente*, *obtener la secretaria del médico*, y *asignar la llamada telefónica a la secretaria del médico*. El archivo BPeL que define la primera actividad establece que su input consiste en un string con el identificador del paciente, y su output corresponde a un string con el identificador del cuidador de salud. Los servicios REST que implementan la actividad *obtener el cuidador del paciente*, consisten en dos consultas SPARQL, en donde la primera une de manera conjunta las propiedades *identificador* y *tieneCuidadorDeSalud* del recurso que representa al paciente (no explicitadas en la Figura 2). La segunda actividad recibe como input el *identificador* del cuidador y genera como output un arreglo de strings con tres espacios, en donde el primero representa el número telefónico móvil del cuidador de salud, el segundo su email, y el tercero su teléfono

fijo. El servicio REST que implementa la segunda actividad es una simple consulta SPARQL que une la propiedad *hasContactInformation* del cuidador de salud. La actividad *enviar SMS* se implementa mediante la utilización del servicio Google Cloud Messaging. El resto de la lógica de coordinación es trivial. El siguiente script representa una versión simplificada de su definición en BPeL:

```
1< process >
2      <sequence >
3            < receive name=‘‘receive’’ partnerLink=‘‘processSelector’’
4                        operation=‘‘request’’ variable=‘‘request’’ initiate=‘‘yes’’ >
5            < /receive >
6            < invoke name=‘‘gettingPatientCaregiver’’ partnerLink=‘‘patient’’
7                        operation=‘‘getCaregiver’’ inputVariable=‘‘patientIdentifier’’
8                        outputVariable=‘‘careGiverIdentifier’’ >
9           < /invoke >
10           < invoke name=‘‘gettingCaregiverContactInformation’’ partnerLink=‘‘caregiver’’
11                       operation=‘‘getCaregiverContactInformation’’
12                       inputVariable=‘‘caregiverIdentifier’’
13                       outputVariable=‘‘careGiverContactInformation’’ >
14           < /invoke >
15           < flow >
16                       < invoke name=‘‘sendingSMS’’ partnerLink=‘‘caregiver’’
17                                   operation=‘‘sendSMS’’ inputVariable=‘‘caregiverCellphoneNumber’’
18                       < /invoke >
19                       < invoke name=‘‘sendingEmail’’ partnerLink=‘‘caregiver’’
20                                   operation=‘‘sendEmail’’ inputVariable=‘‘caregiverEmail’’ >
21                       < /invoke >
22                       < sequence >
23                       < invoke name=‘‘gettingPatientPhysician’’ partnerLink=‘‘patient’’
24                                   operation=‘‘getPatientPhysician’’ inputVariable=‘‘patientIdentifier’’
25                                   outputVariable=‘‘physicianIdentifier’’ >
26                       < /invoke >
27                       < invoke name=‘‘gettingPhysicianSecretary’’ partnerLink=‘‘physician’’
28                                   operation=‘‘getPhysicianSecretary’’
29                                   inputVariable=‘‘physicianIdentifier’’
30                                   outputVariable=‘‘secretaryIdentifier’’ >
31                       < /invoke >
32                       < invoke name=‘‘assignPhoneCallToSecretary’’ partnerLink=‘‘secretary’’
33                                   operation=‘‘setPhoneCallToSecretary’’
34                                   inputVariable=‘‘secretaryIdentifier’’ >
```

```
35                    < /invoke >
36                    < /sequence >
37            < /flow >
38      < /sequence >
39< /process >
```

- Análogamente, cuando el riesgo del paciente es *riesgo medio* o *riesgo alto*, el *process selector* extrae el proceso etiquetado como *notificar riesgo medio/alto al cuidador de salud y al médico*, dentro de otros procesos de negocio. La única diferencia de este proceso es que el cuidador de salud y el médico son notificados del nivel de riesgo del paciente.

5. Caso de Estudio

El framework se implementa en Java con las siguientes herramientas de apoyo: (1) Protege para crear archivos OWL; (2) el Lenguaje de Ejecución de Procesos de Negocios (BPEL por su sigla en Inglés) para describir el proceso y la definición de las actividades; (3) la librería de Java Jena para implementar razonamiento semántico; (4) SPARQL como el lenguaje de consultas y como protocolo de acceso a los archivos OWL, y (5) el sistema administrador de bases de datos MySQL para almacenar los datos de los pacientes.

5.1. Descripción del Escenario

La idoneidad del framework se evalúa a través de un caso de estudio en enfermedades crónicas respiratorias, haciendo uso de datos de pacientes reales y con el apoyo de profesionales de salud del "Hospital Exequiel González Cortés" (HEGC), hospital público pediátrico de Santiago. Esta institución atiende alrededor de 300.000 pacientes cada año, la mayoría de ellos de un segmento de la población con ingresos bajos.

Las enfermedades respiratorias en Santiago tienen un aumento explosivo en otoño e invierno debido a la alta contaminación del aire. El Ministerio de Salud, con apoyo del HEGC, ha desarrollado dos programas innovadores llamados Apoyo Ventilatorio Invasivo y Apoyo Ventilatorio No-Invasivo. Ambos se han desarrollado con el objetivo de cuidar la salud de pacientes con enfermedades respiratorias crónicas en sus domicilios. Estos programas capacitan a los parientes del niño para que tomen y registren las señales fisiológicas, y utilicen una serie de dispositivos médicos. Además, los paramédicos realizan visitas domiciliarias periódicas con el fin de supervisar.

El escenario tiene relación con predecir una potencial crisis del paciente a través de una clasificación del riesgo de salud, con el objetivo de anticipar un estado general de potencial gravedad, sin necesariamente asociarlo a una crisis determinada. Esto se realiza mediante la implementación de un modelo de clasificación inteligente basado en razonamiento difuso, la generación de un contexto del riesgo del paciente a partir de los signos vitales, y el uso del framework para seleccionar procesos de negocio apropiados en función del riesgo, los cuales se mapean a servicios u-health para satisfacer las necesidades del paciente. Además, se muestra un conjunto de servicios que se desarrollaron para apoyar la gestión de alertas una vez que un paciente, su cuidador, y su médico han sido notificados de una potencial crisis de salud. Se evalúa la usabilidad de estos servicios u-health permitiendo a los profesionales de salud interactuar con la funcionalidad del servicio.

5.2. Ejemplo de Contexto de Alto Nivel: Evaluación del Nivel de Riesgo del Paciente

Esta sección describe cómo el context manager genera contexto de alto nivel a través de un servicio que encapsula un modelo de razonamiento difuso, encargado de calcular el nivel de riesgo del paciente, en un servicio REST. De hecho, los médicos del $HEGC$ han definido que la *frecuencia respiratoria*, *frecuencia cardíaca*, *temperatura* y *saturación de oxígeno en la sangre* son signos vitales necesarios para evaluar el nivel de riesgo de crisis en pacientes con enfermedad respiratoria crónica. Estos signos vitales son monitoreados de forma remota por un conjunto de bio-sensores no invasivos, tales como electrocardiogramas, oxímetros, termómetros, entre otros. Se configura un conjunto de micro controladores para integrar y transmitir señales digitales en estos dispositivos hacia el servidor del hospital, pero sus detalles están fuera del alcance de este trabajo.

El *context manager* periódicamente obtiene los signos vitales capturados por los biosensores, y alimenta el *high-level context generator*. Cada vez que este módulo recibe signos vitales, gatilla una consulta SPARQL hacia la ontología, con el objetivo de obtener la instancia del modelo de predicción predefinida para la patología del paciente. El modelo obtenido se ejecuta con los signos vitales medidos para evaluar el nivel de riesgo contingente del paciente: sin riesgo, bajo riesgo, riesgo medio o moderado, riesgo alto. En este escenario, con fines ilustrativos, la evaluación se realiza utilizando un modelo de razonamiento difuso que integra el juicio experto de los médicos del $HEGC$. Sin embargo, como se verá, es perfectamente posible hacerlo con otro modelo de clasificación.

La generación del modelo se lleva a cabo por médicos y analistas, quienes son apoyados por un proceso de negocio semi-automatizado por el *process manager* cada vez que el modelo no alcance un cierto nivel de especificidad y sensibilidad. Es decir, el modelo se ha vuelto obsoleto o sus parámetros deben ser ajustados. Este proceso se describe en la Sección 5.3.1. De esta forma, el nivel de riesgo del paciente es un contexto de alto nivel generado por el *context manager*, el cual es transmitido hacia el *process manager*. Posteriormente, se utiliza este contexto de alto nivel recién generado para identificar un conjunto de procesos de negocio candidatos.

5.2.1. Evaluación del Servicio Contextual Basado en Razonamiento Difuso

Se prueba el servicio contextual que implementa el modelo con razonamiento difuso mediante la utilización de signos vitales provenientes de 31 pacientes con enfermedad respiratoria crónica, bajo la autorización del comité ético del $HEGC$.

Un registro de la base de datos contiene campos para caracterizar anónimamente al paciente – edad, género, diagnóstico, etc. –, almacenar los signos vitales – temperatura, pulso, frecuencia respiratoria y saturación oxígeno en la sangre –, e indicar el tipo de ventilación al que ha estado sometido últimamente. Los signos vitales se registran cada dos o cuatro horas, dependiendo si el paciente está en la unidad de cuidado intensivo, o la unidad pediátrica respectivamente.

El modelo se prueba sin individualizar los registros, utilizando el 80 % de los datos para entrenar y el 20 % para evaluar. El enfoque sin individualización es similar a lo que el personal de salud realiza la mayor parte del tiempo en el triage, en que se asume que las personas tienen un comportamiento parecido. Evidentemente es necesario explorar el caso en que se individualiza al paciente, para lo cual hay que utilizar otros métodos de clasificación distintos a lógica difusa.

El modelo obtiene una especificidad de 78 %, una sensibilidad de 56,6 % y una exactitud de 79,3 % . En este escenario el médico está más interesado en la especificidad y la sensibilidad que en la exactitud. En efecto, una alta sensibilidad significa que los casos falsos negativos son una fracción pequeña, es decir, si el modelo clasifica a un paciente con alto riesgo, la probabilidad de que sea verdad es alta. Sin embargo, no se obtuvo un valor aceptable de sensibilidad, pero sí de especificidad, que es lo que más le importa al médico. De todos modos, es necesario más trabajo probando otros modelos de clasificación para este escenario.

Tabla 1: Tabla de Desviaciones de Puntajes para los Indicadores de Salud.

Esquema de Puntaje	Estado de Bioseñal
1	Temp. Alta o Temp. Baja
2	Frecuencia Cardíaca Alta
2	Frecuencia Respiratoria Alta
2	SAT $90-93$ (CR)
2	SAT $94-95$ (NCR)
3	Frecuencia Cardíaca Baja
3	Frecuencia Respiratoria Baja
4	SAT $90-93$ (NCR)
5	SAT $<$ 90

5.3. Ejemplo de Procesos de Negocio y Servicios U-Health

5.3.1. Ejemplo de Proceso: Desarrollar un Modelo Predictivo

Como se ha explicado, el proceso de negocio conocido como *desarrollar un modelo predictivo* intercala las acciones llevadas a cabo por el analista, como el ajuste de parámetros, y los servicios REST requeridos para simular variables y validar el servicio contextual implementado por el modelo de razonamiento difuso. Se desarrolla este modelo de razonamiento (que fue evaluado en la Sección 5.2.1) siguiendo el mencionado proceso. La primera actividad consiste en definir, por parte de los médicos, un conjunto de rangos de referencia específicos para diferentes segmentos de edad del paciente, ya sea para frecuencia cardíaca (FC), temperatura (T), o frecuencia respiratoria (FR). Estos rangos representan los límites inferiores y superiores aceptables para estos signos vitales. Sin importar la edad, el rango normal de temperatura se establece entre 36 y 37,5 grados Celsius, mientras que la saturación de oxígeno en la sangre (SAT) entre 96 y 100. En segundo lugar, los médicos necesitan desarrollar un esquema de puntajes para caracterizar diferentes estados de estos signos vitales (ver Tabla 1). Por ejemplo, si el paciente sólo muestra una alta o baja temperatura, se le asigna un puntaje de 1. Sin embargo, si el paciente muestra una baja frecuencia respiratoria, entonces se le asigna un puntaje de 3.

Una vez que se establece el esquema de puntuación, es necesario definir la correspondencia entre la combinación de las puntuaciones de los signos vitales y el riesgo de una crisis respiratoria (ver Tabla 2). Por lo tanto, la ejecución de este proceso de negocio semi-automatizado puede proporcionar una evaluación del riesgo basada en esta información.

Finalmente, este conocimiento se formaliza en forma de funciones, como

Tabla 2: Mapeo Niveles de Riesgo con Indicadores de Salud.

Nivel de Riesgo	Enfermedades Crónicas Respiratorias (CR)	Enfermedades No-Crónicas Respiratorias (NCR)
Bajo $(2-3$ puntos)	Frecuencia Cardíaca Alterada	Frecuencia Cardíaca Alterada
Bajo $(2-3$ puntos)	Frecuencia Respiratoria Alterada	Frecuencia Respiratoria Alterada
Bajo $(2-3$ puntos)	SAT $90-93$	SAT $94-95$
Medio $(4-5$ puntos)	SAT $90-93$ y ((FC) o (FR) alterada)	SAT $94-95$ y ((FC) o (FR) alterada)
Medio $(4-5$ puntos)	Alta FC y Alterada FR	Alta FC y Alterada FR
Medio $(4-5$ puntos)	Alta FR y Alterada FC	Alta FR y Alterada FC
Medio $(4-5$ puntos)	Alterada Temperatura y Baja FC	Alterada Temperatura y Baja FC
Medio $(4-5$ puntos)	Alterada Temperatura y Baja FR	Alterada Temperatura y Baja FR
Medio $(4-5$ puntos)		SAT $90-93$
Alto (≥ 6 puntos)	Baja (FC) y Baja (FR)	SAT $90-93$ y ((FC) or (FR) alterada)
Alto (≥ 6 puntos)	SAT < 90	Baja FC y Baja FR
Alto (≥ 6 puntos)	Alta FC y Baja FR y Alterada Temp.	Alta FC y Baja FR y Alterada Temp.
Alto (≥ 6 puntos)	Alta FR y Baja FC y Alterada Temp.	Alta FR y Baja FC y Alterada Temp.
Alto (≥ 6 puntos)		SAT < 90

se muestra en las ecuaciones (1) – (3). Estas ecuaciones definen tres funciones de pertenencia para denotar un rango alto (RA), rango normal (RN) y rango bajo (RB) de una variable.

$$RA_i(x) = \begin{cases} 0 & \text{si } x < c_i \\ \frac{(x-c_i)}{(d_i-c_i)} & \text{si } c_i \leq x \leq d_i \\ 1 & \text{si } x > d_i \end{cases} \quad (1)$$

$$RN_i(x) = \begin{cases} 0 & \text{si } x < a_i \text{ or } x > d_i \\ \frac{(x-a_i)}{(b_i-a_i)} & \text{si } a_i \leq x \leq b_i \\ 1 & \text{si } b_i \leq x \leq c_i \\ \frac{(d_i-x)}{(d_i-c_i)} & \text{si } c_i \leq x \leq d_i \end{cases} \tag{2}$$

$$RB_i(x) = \begin{cases} 0 & \text{si } x > b_i \\ \frac{(b_i-x)}{(b_i-a_i)} & \text{si } a_i \leq x \leq b_i \\ 1 & \text{si } x < a_i \end{cases} \tag{3}$$

donde para cada signo vital i su limite inferior se denota como *Mini* y su límite superior como *Maxi*. Adicionalmente, se introducen dos variables por límite para introducir flexibilidad a estas medidas: *FlexMini* y *FlexMaxi* respectivamente. De esta forma, los parámetros para las funciones de pertenencia se escriben de la siguiente manera:

$$a_i = Mini * (1\text{-}FlexMini) \quad i \in \{T, FR, FC\} \tag{4}$$

$$b_i = Mini * (1\text{+}FlexMini) \quad i \in \{T, FR, FC\} \tag{5}$$

$$c_i = Maxi * (1\text{-}FlexMaxi) \quad i \in \{T, FR, FC\} \tag{6}$$

$$d_i = Maxi * (1\text{+}FlexMaxi) \quad i \in \{T, FR, FC\} \tag{7}$$

Estas ecuaciones se utilizan como un modelo de razonamiento difuso, que se añade a la ontología de procesos como una instancia del recurso *modelo predictivo*.

La calibración del modelo se realiza a través de la similitud con que éste clasifica el riesgo de crisis con respecto al razonamiento de varios médicos expertos. Los parámetros que flexibilizan los límites de los intervalos se ajustan hasta que el razonamiento del modelo sea muy similar al de los médicos evaluadores.

5.3.2. Ejemplo de Servicios U-Health: Gestionando Alertas e Información del Paciente

Se desarrolla un conjunto de servicios para apoyar las alertas y gestión de la información de este caso de estudio. Estos servicios pueden reutilizarse para soportar múltiples actividades y procesos de negocio. Los siguientes servicios pueden ser activados cada vez que una notificación es generada hacia los médicos:

- Un servicio para visualizar la información general de los pacientes, las alertas emitidas relacionadas, el diagnóstico de la enfermdad crónica y los datos relacionados con la última descompensación. Esta lista está ordenada por el nivel de riesgo de los pacientes. Además de esto, los profesionales de salud pueden ver una lista de las alertas generadas, así como su estado correspondiente.

- Un servicio para mostrar una visión general de un paciente específico seleccionado por un profesional de salud. Este servicio muestra los datos personales, indicadores de salud de referencia para cada uno de los signos vitales, el diagnóstico inicial y la crisis anterior, entre otros aspectos.

- Un servicio para visualizar varios comportamientos de signos vitales por paciente. Este servicio permite a los profesionales de salud identificar la evolución de los diferentes indicadores relacionados con la salud del paciente. Presenta gráficos para visualizar los datos transmitidos en el último período, por ejemplo, las últimas 24 horas. Lo relevante aquí es que el profesional de salud tiene toda la información necesaria para detectar cualquier anomalía con respecto al estado de salud del paciente.

- En caso de necesitar información adicional se puede realizar un análisis temporal en cualquier indicador de salud mediante la elección de un intervalo de fechas.

Además, se desarrollaron múltiples servicios móviles que se ejecutan en la plataforma Android, con el objetivo de permitir a los médicos comprobar la información en tiempo real. La Figura 3 muestra algunos de estos servicios, que permiten gestionar las alertas y los registros de salud.

Evaluación del Servicio U-health La utilidad y facilidad de uso de estos servicios son validados mediante un estudio de usuario por los profesionales de salud del *HEGC*. Se analizan tres casos de estudio de pacientes con las siguientes edades: dos meses, un año con dos meses y ocho años. La simulación se realiza mediante el uso de un formulario web, donde los signos vitales se

Figura 3: Servicios Móviles para Gestionar Alertas y Registros de Salud.

introducen a través de un dispositivo tablet. Las interfaces gráficas de usuario de los servicios móviles se despliegan en un Smartphone.

Esta simulación se valida a través de una encuesta que utiliza una escala Likert de cinco niveles. Se encuestó un total de 16 profesionales. La composición de la muestra contiene un 6 % de enfermeras, 19 % de médicos y 75 % de kinesiólogos. La Tabla 3 muestra que cada ítem evaluado tiene un promedio mayor o igual a 4.0, lo que indica que los usuarios consideran a los servicios como útiles (100 %), fáciles de usar (81 %) y que contribuyen a desencadenar acciones preventivas en un momento oportuno (81 %). En relación a las notificaciones preventivas, el 63 % está de acuerdo en que los servicios generan notificaciones apropiadas en relación al riesgo del paciente, y el 19 % está muy de acuerdo con esto.

Se utiliza el alfa de Cronbach para evaluar la confiabilidad de la muestra. Este indicador se basa en los siguientes parámetros: número de ítems, covarianza promedio dentro de los ítems, y la varianza de la puntuación total. El alfa de Cronbach de la muestra tiene un valor de 0.84, considerando un total de 9 ítems y 16 personas encuestadas. Este indicador muestra una buena consistencia interna de la muestra, lo que indica que los resultados son bastante confiables.

Adicionalmente, se lleva a cabo un análisis estadístico de los resultados para determinar el intervalo de confianza para cada elemento con un 95 % de confianza. De la tabla 3 se puede observar que los elementos mejor valorados corresponden a la utilidad (ítem 2), y el atractivo de los servicios (ítem 1). Los ítems con peor evaluación corresponden a la capacidad de uso de los servicios (ítem 3), y las notificaciones basadas en el riesgo del paciente (ítem 7). Estos resultados indican que las principales áreas de mejora se relacionan con la interfaz gráfica de usuario de los servicios, para facilitar la navegación dentro de ellos, y cómo se visualiza la información asociada al estado del paciente.

Tabla 3: Resultados de Usabilidad de Servicios Móviles.

ID	Ítem	Puntaje Promedio	Intervalo de Confianza (95 %)
1	Encuentro los servicios atractivos	4.8	[4.61 - 5.00]
2	Los servicios son útiles	4.8	[4.53 - 4.97]
3	Los servicios son fáciles de usar	4.1	[3.73 - 4.40]
4	Los servicios permiten visualizar data relevante del paciente	4.6	[4.25 - 4.87]
5	Los servicios permiten generar acciones de manera oportuna	4.4	[3.98 - 4.77]
6	Los servicios permiten visualizar de manera correcta la información transmitida	4.4	[4.02 - 4.73]
7	Los servicios generan notificaciones adecuadas en base al riesgo del paciente	4.0	[3.69 - 4.31]
8	Estoy satisfecho con las características generales	4.1	[3.73 - 4.40]
9	Los servicios muestran información valiosa para la toma de decisiones preventivas	4.3	[3.92 - 4.70]
10	Evaluación general de 1 a 7	6.2	[5.92 – 6.42]

Se realiza un análisis de correlación con el fin de proporcionar información acerca de las características que los usuarios más valoran. Los resultados indican que los ítems con mayor correlación con la evaluación general de los servicios son los siguientes: el ítem relacionado a la generación de notificaciones apropiadas basadas en el riesgo del paciente (ítem 7, correlación 0.84) y el ítem relativo a la capacidad para ver datos relevantes del paciente (ítem 4, correlación 0.73).

6. Conclusiones

Este trabajo propone un framework basado en Web semántica para la provisión continua de servicios u-health para pacientes con enfermedades crónicas. Sus necesidades de salud están representadas en procesos de negocio, los cuales se mapean a servicios u-health disponibles. Además, se define un modelo de descripción semántico de los procesos de negocio, actividades, servicios u-health, contexto médico, y el contexto del entorno.

Como parte del trabajo futuro, se planea extender el modelo semántico para incluir ontologías médicas existentes en la literatura, así como también aplicar el framework en otras enfermedades crónicas. Además, se tiene la intención de explorar la posibilidad de aplicar un enfoque evolutivo para la ontología para permitir una expansión autónoma del modelo de descripción semántica. En relación a los modelos de clasificación de riesgo de salud, se trabajará en su mejora, abordando el caso de los falsos positivos.

Agradecimientos: *Este trabajo fue financiado por el programa CONICYT - IDEA FONDEF, código de proyecto: CA13i-10300, y por el programa CONICYT - FONDECYT, código de proyecto: 11130252. Los autores desean agradecer el apoyo continuo del "Instituto Sistemas Complejos de Ingeniería"(ICM: P-05-004-F, CONICYT: FBO16). Además, se agradece a las distintas personas del hospital que apoyaron el desarrollo del proyecto, como la directora del HEGC Dra. Begoña Yarza, la enfermera gestora de camas Francisca Molina, la directora de los programas AVI Dra. Rebeca Paiva, y la interna de medicina Katrina Lolas. Se agradece al Ingeniero Fabián García por probar el modelo de razonamiento difuso con una base de datos más confiable.*

Referencias

[1] B. Arnrich, O. Mayora, J. Bardram, y G. Troster. Pervasive healthcare: paving the way for a pervasive, user-centered and preventive healthcare model. *Methods Inf Med*, 49(1):67 – 73, 2009.

[2] P. Fortier y B. Viall. Development of a mobile cardiac wellness application and integrated wearable sensor suite. In *The Fifth International Conference on Sensor Technologies and Applications*, pages 301–306, 2011.

[3] D. Fotiadis, A. Likas, y V. Protopappas. *Intelligent Patient Monitoring*. John Wiley and Sons, Inc., 1st edition, 2006.

[4] R. Gao. A phone-based e-health system for osas and its energy issue. In *2012 International Symposium on Information Technology in Medicine and Education*, pages 682–686, 2012.

[5] D. Han, M. Lee, y S. Park. The-muss: Mobile u-health service system. *Comput. Methods Prog. Biomed.*, 97(2):178–188, 2010.

[6] A. Jimenez-Molina, J. Kim, H. Koo, B. Kang, y I. Ko. A semantically-based task model and selection mechanism in ubiquitous computing environments. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 829–837, 2009.

[7] A. Jimenez-Molina y I. Ko. Spontaneous task composition in urban computing environments based on social, spatial, and temporal aspects. *Engineering Applications of Artificial Intelligence*, 24(8):1446 – 1460, 2011.

[8] K. Kawamoto, A. Honey, y K. Rubin. The hl7-omg healthcare services specification project: Motivation, methodology, and deliverables for enabling a semantically interoperable service-oriented architecture for healthcare. *Journal of the American Medical Informatics Association:JAMIA*, 16(6):1874 – 881, 2009.

[9] J. Kim, S. Ahn, J. Soh, y K. Chung. U-health platform for health management service based on home health gateway. In *IT Convergence and Security*, pages 351–356, 2012.

[10] F. Miao, X. Miao, W. Shangguan, y Y. Li. Mobihealthcare system: Body sensor network based m-health system for healthcare application. *E-Health Telecommunication Systems and Networks*, 1(1):12 – 18, 2012.

[11] W. Omar, B. Ahmad, A. Taleb-Bendiab, y Y. Karam. A software framework for open standard self-managing sensor overlay for web services. In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, pages 72–81, 2005.

[12] F. Paganelli y D. Giuli. An ontology-based system for context-aware and configurable services to support home-based continuous care. *IEEE Transactions on Information Technology in Biomedicine*, 15(2):324–333, 2011.

[13] D. Patil, V. Wadhai, M. Gund, R. Biyani, S. Andhalkar, y B. Agrawal. An adaptive parameter free data mining approach for healthcare application. *International Journal of Advanced Computer Science and Applications*, 3(1), 2012.

[14] M. Serhani, A. Benharref, y E. Badidi. Towards dynamic non-obtrusive health monitoring based on soa and cloud. In *Proceedings of the Second International Conference on Health Information Science*, pages 125–136, 2013.

Una aplicación del problema del cartero rural a la recolección de residuos reciclables en Argentina

Gustavo Braier[*]
Guillermo Durán[**]
Javier Marenco[***]
Francisco Wesner[***]

Resumen

En este trabajo reportamos la aplicación de técnicas de programación matemática a la optimización de las rutas de vehículos de recolección de residuos reciclables en Morón, una municipalidad en el Gran Buenos Aires, Argentina. Este problema es un caso particular del problema del cartero rural abierto en grafos mixtos, y se resuelve por medio de un modelo de programación lineal entera. Las rutas generadas por este procedimiento son significativamente mejores que las rutas designadas a mano y que estaban en uso antes de la implementación del presente proyecto. El beneficio más importante consiste en que estas rutas cubren el 100 % del sector a ser recolectado, mientras que con las rutas diseñadas a mano se omitía hasta el 16 % de las cuadras. Las rutas generadas por el modelo fueron implementadas por la municipalidad en 2014.

Palabras Clave: Problema del cartero rural, Recolección de residuos, Ruteo de vehículos.

[*]Braier & Asociados Consultores (papyro.com), Argentina.
[**]CONICET, Argentina. Instituto de Cálculo, FCEyN, Universidad de Buenos Aires, Argentina. Departamento de Matemática, FCEyN, Universidad de Buenos Aires, Argentina. Departamento de Ingeniería Industrial, FCFM, Universidad de Chile, Chile.
[***]Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina.

1. Introducción

Este trabajo presenta un enfoque basado en programación lineal entera para mejorar el ruteo de vehículos de recolección de residuos reciclables en la municipalidad de Morón, al oeste de la Ciudad de Buenos Aires, Argentina. Morón tiene una población de 320,000 habitantes (de acuerdo con el Censo 2010) y cubre un área de 55.6 km^2. La municipalidad está dividida en cinco distritos administrativos (Castelar, Morón centro, Haedo, Palomar y Villa Sarmiento), que a su vez están divididos en hasta siete sectores, de modo tal que cada sector es recorrido por un único camión de recolección. Los residuos reciclables en cada distrito se recolectan un día de la semana (de lunes a viernes). Como el sistema de recolección no cuenta con contenedores o basureros centralizados, los vehículos deben recorrer todos los frentes domiciliarios para realizar la recolección.

Las rutas de los camiones de la municipalidad para la recolección de los residuos reciclables habían sido diseñadas, previamente a la concreción de este proyecto, de manera manual. Como consecuencia de ello los recolectores terminaban no recorriendo algunas de las cuadras del municipio en las horas estipuladas. Esta situación generaba numerosas protestas de los vecinos, que motivaron la realización de este trabajo.

El problema de encontrar una ruta óptima que asegura que todas las cuadras de un sector son visitadas es un caso particular del problema del *cartero rural abierto* en grafos mixtos [15], que es NP-hard (en este trabajo, "cuadra" se refiere a un segmento de una calle a lo largo de una manzana entre dos esquinas consecutivas). Dado un grafo y un conjunto de arcos/aristas seleccionados, el problema del cartero rural solicita hallar un ciclo de costo mínimo que visite cada arco/arista seleccionado al menos una vez, y es una generalización del problema del cartero Chino. El problema del *cartero rural abierto* solicita un camino con las mismas propiedades. El problema del cartero rural ha sido objeto de mucho interés por parte de la comunidad de optimización combinatorial, tanto para grafos dirigidos [9, 13, 14, 4] como para grafos mixtos [6, 7, 5, 8].

Las reglas de tránsito complican el diseño de los recorridos de los camiones de recolección, y las reglas más relevantes en este contexto son la prohibición de girar a la izquierda en esquinas con semáforo y la prohibición de realizar "giros en U". El enfoque propuesto en este trabajo utiliza un algoritmo de planos de corte sobre un modelo natural de programación lineal entera para el

problema. La principal contribución teórica de este trabajo es la introducción de un procedimiento para combinar *subtours*, que permite reducir la cantidad de rondas de planos de corte, reduciendo así los tiempos de ejecución necesarios para obtener soluciones óptimas. Este procedimiento no se puede aplicar en instancias generales del problema del cartero rural, aunque demostró ser útil en el caso particular considerado en este trabajo.

Un survey interesante de problemas de ruteo de vehículos para la recolección de residuos está dado en [11]. Se pueden hallar en la literatura aplicaciones de técnicas de programación matemática a la recolección de residuos en distintas ciudades, incluyendo Chicago (EEUU) [10], Kaoshiung (Taiwan) [3], Lisboa (Portugal) [12], Hamilton (Canadá) [16], Santiago (Chile) [1] y la Ciudad de Buenos Aires [2]. La Ciudad de Buenos Aires no incluye la municipalidad de Morón, y el problema de recolección de residuos estudiado en [2] tiene una estructura diferente a la considerada en este trabajo.

Este trabajo está organizado de la siguiente forma. La Sección 2 describe el problema a resolver y presenta un modelo de programación lineal entera. En la Sección 3 presentamos el procedimiento que utilizamos para resolver este modelo, obteniendo así las rutas de recolección en cada sector. La Sección 4 reporta los resultados computacionales sobre el municipio de Morón, y la Sección 5 cierra el trabajo con nuestras conclusiones.

2. El problema

Cada sector –que será recolectado por un único camión– está representado por un grafo mixto cuyos vértices corresponden a las esquinas del sector. Se tienen las coordenadas geográficas de cada vértice de este grafo, y se tiene también una indicación especificando si hay un semáforo en esa esquina. Los arcos del grafo representan calles de un sentido de circulación (calles de "mano única"). Las aristas del grafo representan calles con dos sentidos de circulación (calles "doble mano") pero que son angostas y alcanza con recorrerlas en cualquiera de los dos sentidos para realizar la recolección de residuos. Hay también avenidas anchas con dos sentidos de circulación, pero en ese caso se representan las cuadras como dos arcos paralelos, uno en cada dirección. Cada arco y cada arista tiene asociada una distancia, que se calcula como la distancia en línea recta entre las dos esquinas.

Los vehículos de recolección no pueden realizar giros en U. Para incorporar esta restricción, expandimos el grafo dividiendo cada vértice en varios vértices que representan las formas posibles de arribar a la esquina. Se agregan arcos

auxiliares entre estos nuevos vértices, que representan las transiciones permitidas entre los vértices de la esquina. La Figura 1 muestra un ejemplo de esta expansión en una esquina de una calle de un sentido con una calle de doble mano. En la figura, se muestran con líneas punteadas los arcos auxiliares. De acuerdo con los funcionarios de la municipalidad, los camiones de recolección no tienen restricciones en cuanto a la posibilidad de girar en las esquinas, incluso en calles angostas. Esto implica que un giro en una esquina no tiene un costo adicional demasiado alto, y entonces se asigna una distancia pequeña y positiva a los arcos auxiliares. Esta distancia es la misma para todos los arcos auxiliares.

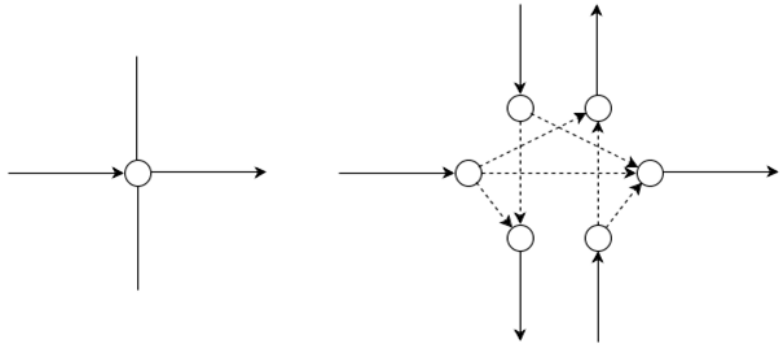


Figura 1: Expansión de un vértice del grafo para representar los giros permitidos.

La prohibición de realizar giros en U no aplica para las calles sin salida. Para permitir esta situación, se modifica levemente el grafo agregando arcos auxiliares entre los vértices finales de una calle sin salida, como en la Figura 2.

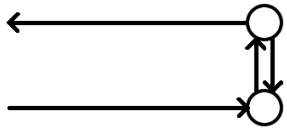
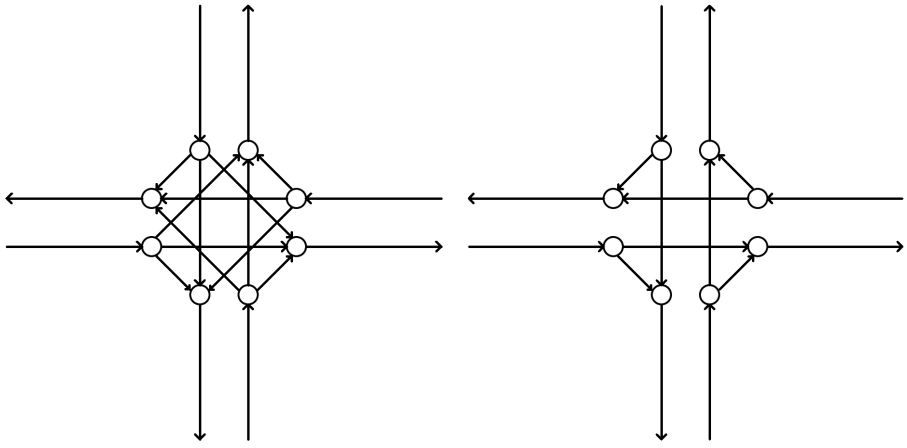


Figura 2: Arcos auxiliares para permitir giros en U en calles sin salida.

Además de los giros en U, las reglas de tránsito en Argentina prohiben girar a la izquierda en esquinas con semáforos de calles de doble mano, excepto cuando el semáforo permite explícitamente este giro con una luz especial. La expansión del grafo descrita arriba permite incorporar estas restricciones, eliminando los arcos auxiliares que representan giros ilegales. Un ejemplo de esta situación está dado en la Figura 3, para la intersección de dos calles de

doble mano.



(a) (b)

Figura 3: Giros permitidos en la intersección (a) sin semáforos y (b) con semáforos.

El punto de inicio de la ruta de recolección de un sector debe estar ubicado en el borde del sector más cercano al depósito de origen, desde donde parten los camiones. Cualquier vértice en este borde puede ser seleccionado como el punto de inicio del recorrido. La selección del vértice final del recorrido no tiene un impacto significativo en la performance total del mismo, dado que una vez que la recolección ha sido completada el camión vuelve directamente al centro de reciclado. Como los sectores no son muy extensos y el centro de reciclado se encuentra bastante alejado de los diferentes sectores, se puede despreciar este costo adicional.

En algunas instancias, en especial en sectores con muchas calles de mano única, los camiones de reciclado pueden tener que realizar pequeñas desviaciones para ingresar a algunas calles. Para reducir este efecto y evitar desvíos innecesarios, es posible que los conductores deban ingresar brevemente a un sector vecino. Para contemplar estas situaciones, se agregan al grafo las manzanas inmediatamente contiguas al sector en cuestión, formando una banda perimetral de manzanas en sectores adyacentes que no se deben recolectar pero que pueden ser utilizadas para optimizar el recorrido. Estos nuevos arcos y aristas se tratan como arcos auxiliares, con el costo adecuadamente calculado.

La Figura 4 muestra una ejemplo de esta situación para un sector en el distrito de Villa Sarmiento. Las cuadras que conforman la banda perimetral están mostradas con líneas oscuras. En esta figura no se agregaron manzanas

de cuadros adyacentes hacia el norte (arriba) ni el sur (abajo), dado que el norte está cerrado por una autopista que no se puede atravesar fácilmente y el sur es adyacente a un municipio vecino, para el cual no se disponía de información actualizada.

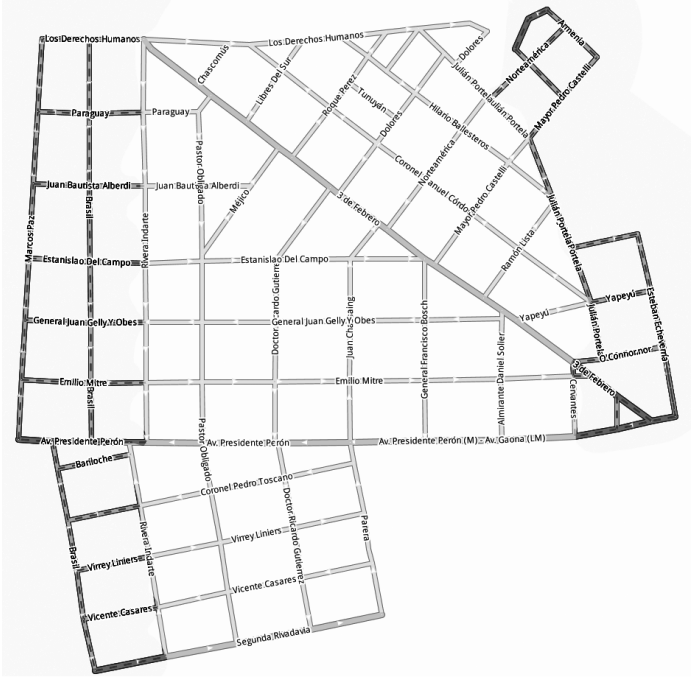


Figura 4: Sector en Villa Sarmiento con manzanas de sectores adyacentes que se pueden recorrer en forma auxiliar.

Los párrafos previos describen la construcción del grafo mixto que representa el mapa del sector a recorrer. Sea $G = (V, E, A)$ este grafo, donde los vértices corresponden a todas las formas posibles de llegar a cada esquina, el conjunto de aristas E representa las calles de doble mano que alcanza con visitar en cualquiera de las dos direcciones, y el conjunto de arcos A representa las cuadras que se deben recorrer en la dirección especificada. Como parte de la entrada se tiene una partición $A = A_M \cup A_{\text{AUX}}$, donde A_M contiene los arcos obligatorios (que deben ser visitados) y A_{AUX} es el conjunto de arcos auxiliares (que pueden ser visitados sólo si es conveniente). Tenemos además una función de peso $w : E \cup A \to \mathbb{R}$ que asocia un peso a cada arista y cada arco, y suponemos que los arcos auxiliares entre vértices asociados a una misma esquina tienen el mismo peso, positivo y pequeño. Finalmente, tenemos un conjunto $I \subseteq V$ de vértices iniciales, que especifica los puntos permitidos de inicio del recorrido.

Con estas definiciones, podemos dar un modelo de programación lineal

entera para este problema. Para cada arista $ij \in E$ introducimos las variables enteras x_{ij} y x_{ji}, que representan la cantidad total de veces que se recorren ij y ji, respectivamente. Para cada arco $ij \in A$ introducimos la variable entera y_{ij}, que representa la cantidad total de veces que se recorre el arco ij. Finalmente, para cada vértice $i \in I$ definimos la variable binaria s_i que especifica si i es el primer vértice de la ruta, y para cada vértice $j \in V$ definimos la variable binaria t_j que especifica si j es el último vértice de la ruta. Para cualquier conjunto $S \subseteq V$, definimos $E(S, \bar{S}) = \{ij \in E : i \in S, j \notin S\}$ y $A(S, \bar{S}) = \{ij \in A : i \in S, j \notin S\}$. Podemos ahora formular el modelo como sigue.

$$\text{mín} \sum_{ij \in E} w_{ij}(x_{ij} + x_{ji}) + \sum_{ij \in A} w_{ij} y_{ij} \tag{1}$$

$$y_{ij} \geq 1 \qquad \forall ij \in A_M \tag{2}$$

$$x_{ij} + x_{ji} \geq 1 \qquad \forall ij \in E \tag{3}$$

$$s_i + \sum_{j:ji \in E} x_{ji} + \sum_{j:ji \in A} y_{ji} = \sum_{j:ij \in E} x_{ij} + \sum_{j:ij \in A} y_{ij} + t_i \qquad \forall i \in I \tag{4}$$

$$\sum_{j:ji \in E} x_{ji} + \sum_{j:ji \in A} y_{ji} = \sum_{j:ij \in E} x_{ij} + \sum_{j:ij \in A} y_{ij} + t_i \qquad \forall i \in V \backslash I \tag{5}$$

$$\sum_{i \in I} s_i = 1 \tag{6}$$

$$\sum_{i \in V} t_i = 1 \tag{7}$$

$$\sum_{ij \in E(S,\bar{S})} x_{ij} + \sum_{ij \in A(S,\bar{S})} y_{ij} \geq 1 \qquad \forall S \subset V, S \neq \emptyset \tag{8}$$

$$x_{ij} \in \mathbb{Z}_+ \qquad \forall ij \in E \tag{9}$$

$$y_{ij} \in \mathbb{Z}_+ \qquad \forall ij \in A \tag{10}$$

$$s_i \in \{0, 1\} \qquad \forall i \in I \tag{11}$$

$$t_i \in \{0, 1\} \qquad \forall i \in V \tag{12}$$

La función objetivo intenta minimizar el costo total de la ruta de recolección, es decir, la suma de los costos individuales de las cuadras recorridas por el vehículo. Las restricciones (2) solicitan que cada arco sea visitado al menos una vez, mientras que las restricciones (3) requieren que cada arista sea visitada al menos una vez en alguna de las dos direcciones. Las restricciones (4)-(5) aseguran que la solución es un camino, imponiendo la conservación de flujo en cada vértice con excepción de los vértices de inicio y finalización del recorrido. Las restricciones (6) y (7) garantizan que estos dos vértices son únicos.

A pesar de que la incorporación de arcos auxiliares permite modelar las restricciones de tránsito, también agrega una complicación en el modelado. Dado que no es necesario cubrir todos los arcos en el grafo, entonces requerir que el grado de entrada sea igual al grado de salida no es suficiente para garantizar que las soluciones factibles del modelo sean rutas conexas. De hecho, las restricciones (2)-(7) permiten la formación de subtours, que deben ser evitados.

Por ejemplo, si se resuelve el modelo dado por (1)-(7) y (9)-(12) sobre la instancia dada por la Figura 5, se puede obtener una solución como la especificada en la Figura 6. Esta solución tiene un camino principal (la línea punteada) formado por los vértices $C_1 = (4, 1, 2, 3, 6, 9, 8, 7, 4, 5, 2)$ y un subtour (la línea rayada) formado por los vértices $C_2 = (5, 6, 9, 8, 5)$. Recordemos que como el grafo está expandido, cada vértice en la figura en realidad representa varios vértices en el grafo. Una vista detallada del vértice 5 revela cómo se encuentra separado el subtour del resto de la ruta, y se tiene una situación similar en los vértices 6, 9 y 8. Para evitar esta situación, las *restricciones de rompimiento de subtours* (8) impiden la formación de subtours. Estas restricciones imponen la existencia de al menos un arco o arista conectando el subtour al resto de los vértices. Finalmente, las restricciones (9)-(12) especifican los valores posibles para las variables.

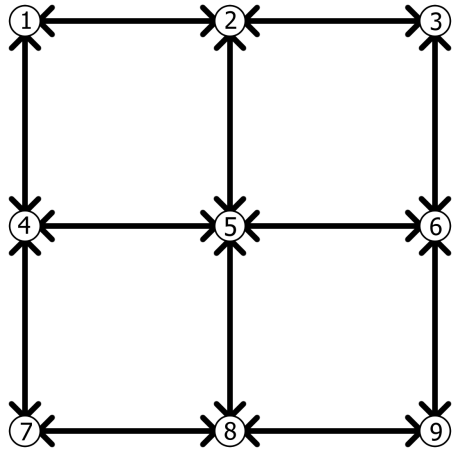


Figura 5: Grafo de un sector compuesto por cuatro manzanas y calles de doble mano.

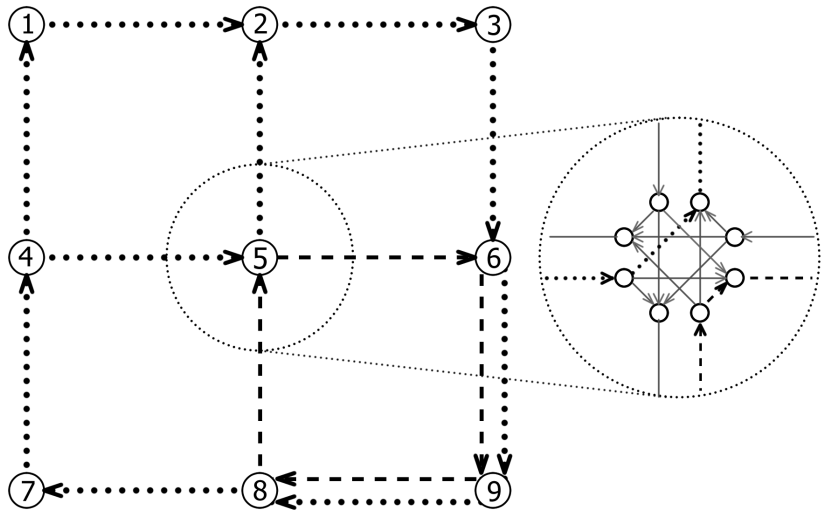


Figura 6: Formación de un subtour en una solución para el modelo sin las restricciones de eliminación de subtours, para la instancia de la Figura 5.

3. Procedimiento propuesto

Proponemos en esta sección un procedimiento sencillo basado en planos de corte para resolver el modelo de programación entera presentado en la sección anterior. La Figura 7 muestra el pseudocódigo de este algoritmo. Se resuelve un modelo relajado, compuesto por las restricciones (2)-(7) y (9)-(12) (Pasos 1 y 2). Si la solución obtenida no contiene subtours (Paso 4), entonces es óptima y el algoritmo termina. En caso contrario, la solución contiene un subtour que debe ser evitado.

Para excluir esta solución, una técnica estándar consiste en agregar la restricción de rompimiento de subtours (8) asociada con el subtour, cortando así esta solución. Sin embargo, esta estrategia tiene el problema de que puede requerir muchas iteraciones hasta hallar una solución sin subtours. Para manejar esta situación, primero intentamos combinar el subtour con el camino principal por medio de intercambio de arcos auxiliares. Por ejemplo, la solución que se muestra en la Figura 6 se puede modificar intercambiando los arcos auxiliares en el vértice 5, como en la Figura 8, obteniendo así una ruta conexa. Como el costo de los arcos auxiliares entre vértices de una misma esquina es el mismo, la nueva ruta obtenida por este procedimiento también es óptima.

1. Inicializar el modelo relajado $\mathcal{M} :=$ (1)-(7) y (9)-(12).
2. Resolver el modelo $\mathcal{M}$.
3. Si $\mathcal{M}$ es no factible, retornar "no factible" y terminar.
4. Si la solución óptima de $\mathcal{M}$ no tiene subtours, retornar esta solución y terminar.
5. Si los subtours se pueden combinar con el camino principal entre el nodo de inicio y el nodo de finalización, combinarlos, retornar la solución obtenida y terminar.
6. En caso contrario, agregar a $\mathcal{M}$ una restricción de rompimiento de subtours (8) para cada subtour en la solución y volver al Paso 2.

Figura 7: Algoritmo de planos de corte para resolver en forma exacta el caso particular del problema del cartero rural planteado en este trabajo.

A pesar de su gran eficiencia computacional, esta técnica de *combinación de subtours* no funciona en todos los casos. Por ejemplo, consideremos el mapa de la Figura 9 y la solución (con un subtour) de la Figura 10. En este caso, el subtour (línea rayada) no se puede combinar con el camino principal (línea punteada) por el método descrito arriba.

La estrategia definitiva consiste entonces en utilizar el procedimiento de combinación de subtours en conjunto con el agregado dinámico de restricciones de rompimiento de subtours. Si una solución no tiene subtours, se retorna la solución como resultado (Paso 4). Si la solución tiene uno o más subtours que se pueden combinar con el camino principal, se realiza esta combinación y se retorna la solución obtenida (Paso 5), que es óptima. Finalmente, si la solución tiene subtours que no se pueden combinar con el camino principal, se agrega al modelo una restricción de rompimiento de subtours (8) por cada subtour en la solución y se repite el procedimiento (Paso 6).

La idea no es eliminar completamente el agregado dinámico de restricciones de rompimiento de subtours, sino reducir la cantidad de rondas de planos de corte necesarias para obtener una solución factible (es decir, conexa). El algoritmo propuesto garantiza que se encuentra una solución óptima en un número finito de pasos, dado que en el peor caso se agregan todas las restricciones de rompimiento de subtours y el número de estas restricciones es

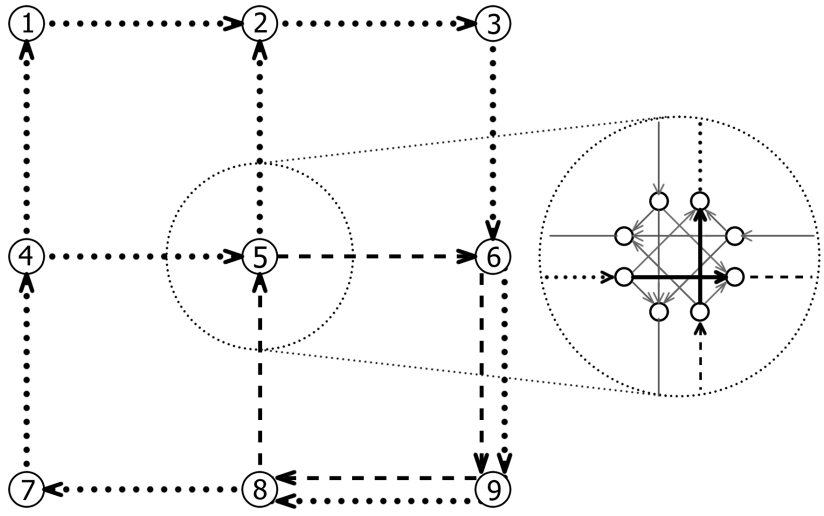
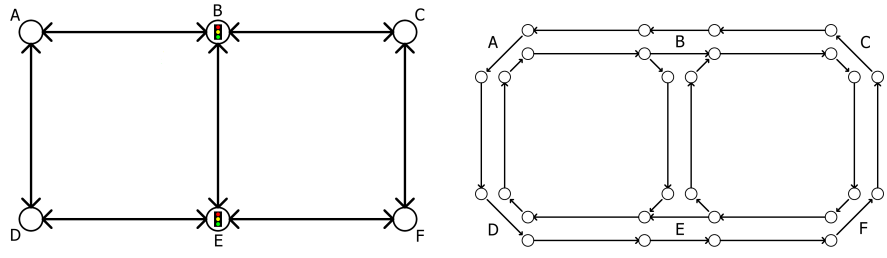


Figura 8: Combinación de un subtour con el camino principal para la ruta de la Figura 5.

exponencial pero finito. La intención es que esto no sea necesario, buscando una solución factible y óptima por medio del procedimiento de combinación de subtours.



(a) (b)

Figura 9: Un grafo simple y el grafo expandido correspondiente.

El agregado dinámico de restricciones de rompimiento de subtours se realiza una vez que el procedimiento de resolución del modelo $\mathcal{M}$ ha terminado. Esto permite identificar los subtours rápidamente, dado que las variables en la solución obtenida toman valores enteros. Un enfoque alternativo consiste en separar las restricciones (8) durante el procedimiento branch-and-bound (convirtiendo así este procedimiento en un algoritmo de tipo branch-and-cut), pero la detección de restricciones de rompimiento de subtours violadas pasa a ser

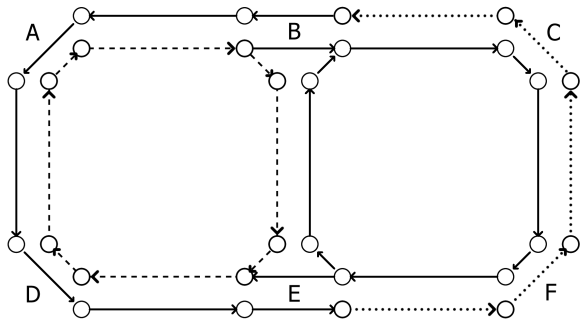


Figura 10: Una ruta con subtours que no se pueden combinar utilizando la técnica propuesta, para el grafo de la Figura 9.

más complicada. Por este motivo, decidimos en este trabajo buscar subtours una vez que el modelo se resuelve en forma óptima y se obtiene una solución entera.

4. Resultados computacionales

En esta sección presentamos y analizamos los resultados de los experimentos realizados para evaluar la efectividad del algoritmo propuesto en la sección anterior. Se fijó un máximo de 10 minutos para la resolución del modelo de programación entera. Se determinó este límite luego de comprobar que la mayoría de los sectores considerados en este trabajo se resuelven en forma óptima en menos de 10 minutos, y para el resto de los sectores el gap de optimalidad es muy bajo.

Los experimentos fueron realizados en una PC con un procesador AMD Phenom II x4 945 corriendo a 3GHz, y 4GB de memoria RAM. El sistema operativo es Linux Ubuntu 12.04 (32 bits) y el solver para los modelos de programación entera es SCIP. A pesar de que existen solvers comerciales con mejor performance que SCIP sobre bancos de prueba estándar, el uso de SCIP no trajo aparejados problemas importantes, y su performance fue adecuada para este trabajo.

La Tabla 1 muestra los experimentos realizados para determinar la efectividad del procedimiento de combinación de subtours propuesto en este trabajo. Todos los datos corresponden al distrito de Castelar, que fue seleccionado para estas pruebas debido a la gran diversidad de sectores que presenta. Los sectores de este distrito varían en tamaño, forma (algunos son rectangulares y otros son más irregulares) y tipos de calles. Para cada sector, se muestra el

número de iteraciones y los tiempos de resolución sin y con el procedimiento de combinación de subtours, respectivamente.

En Castelar1 y Castelar5, los dos sectores con la menor mejora, el procedimiento de combinación de subtours redujo de 2 a 1 la cantidad de iteraciones y el tiempo de resolución en aproximadamente la mitad. Por su parte, en Castelar3 el algoritmo sin combinación de subtours requiere de 59 iteraciones y casi 10 horas de ejecución para encontrar una solución factible, mientras que en la primera iteración el procedimiento de combinación de subtours permite encontrar una solución conexa y óptima. En otros sectores se observaron resultados similares.

Sector	$\|V\|$	$\|E\|$	$\|A_M\|$	$\|A_{\text{AUX}}\|$	Sin el Paso 5		Con el Paso 5	
					Iteraciones	Tiempo	Iteraciones	Tiempo
Castelar1	1372	658	28	1570	2	5.2 seg	1	2.4 seg
Castelar2	650	100	225	559	5	19 seg	1	3 seg
Castelar3	852	200	226	782	59	9.7 hs	1	10 min
Castelar4	878	354	85	876	6	52.4 seg	1	5.6 seg
Castelar5	1090	508	37	1416	2	5.7 seg	1	3 sec
Castelar6	738	166	203	650	22	3.7 hs	1	10 min
Castelar7	1082	524	17	1341	35	5.9 hs	2	20 min
Castelar8	1276	622	46	1574	24	3.7 hs	1	17.9 seg

Tabla 1: Contribución del procedimiento de combinación de subtours para el distrito de Castelar.

La Tabla 2 presenta información provista por la municipalidad antes del proyecto descrito en este trabajo. Las rutas informadas en esa tabla corresponden a los recorridos diseñados manualmente sin más herramientas que un mapa del sector a recorrer. Como puede verse en la Tabla en todos los sectores quedan cuadras sin recorrer, llegando a 60 en Castelar4.

Para comparar esta situación con nuestros resultados, la Tabla 3 muestra las características de las rutas obtenidas por el procedimiento propuesto en este trabajo. Por construcción, las rutas propuestas por el algoritmo de la sección anterior recorren todas las cuadras, visitando así el 100 % del sector en cuestión. En algunos casos, las rutas generadas por nuestro procedimiento son más largas que las rutas manuales, debido a la restricción que solicita recorrer todas las cuadras. Una estimación sencilla nos permitió ver que las nuevas rutas, aún siendo algunas de ellas más largas que las actuales, se podían recorrer en los tiempos asignados para el recorrido (5 horas en total). Por ejemplo, para la instancia Castelar4 se recorrían 46km en 4:10 horas, y el algoritmo propuso una ruta de 50km. Es prácticamente imposible estimar con precisión el tiempo de recorrida de la nueva ruta, pero las 4:10 horas necesarias

Sector	Fecha	Distancia viajada	Cuadras salteadas	Tiempo de viaje	% del sector visitado
Castelar4	26/04/2013	46 km	60	4:10hs	88,5 %
Castelar7	19/04/2013	43,1 km	6	3:40hs	98,6 %
Haedo2	23/04/2013	17,5 km	20	2:26hs	89,7 %
Palomar1	10/04/2013	25 km	46	3:38hs	84,5 %
Palomar2	10/04/2013	24,5 km	35	4:10hs	87,5 %
Palomar3	24/04/2013	31,2 km	17	3:12hs	94,8 %
Palomar4	27/03/2013	35,2 km	35	4:46hs	91,0 %
VillaSarmiento1	06/04/2013	24,6 km	10	3:09hs	96,1 %
VillaSarmiento2	13/04/2013	23,7 km	9	2:44hs	96,3 %
VillaSarmiento3	15/04/2013	23,2 km	17	3:13hs	93,2 %

Tabla 2: Datos de GPS proporcionados por la municipalidad.

para recorrer 46 km nos permitieron asumir que la nueva ruta se podía recorrer en las 5 horas disponibles. Estas estimaciones realizadas a priori mostraron ser razonables cuando se aplicaron las nuevas rutas en la práctica.

Sector	$\|V\|$	$\|E\|$	A_M	A_{AUX}	Tiempo	Gap	Iteraciones	Costo
Castelar4	878	354	85	876	600 seg	0.008	1	50.45 km
Castelar7	1082	524	17	1341	646 seg	0.003	2	40.45 km
Haedo2	1154	396	181	1256	0.3 seg	0	1	18.31 km
Palomar1	1046	502	21	1251	30 seg	0	1	33.51 km
Palomar2	822	316	95	875	512 seg	0	1	26.90 km
Palomar3	1254	608	19	1596	1200 seg	0.003	2	30.53 km
Palomar4	1090	500	45	1353	1200 seg	0.003	2	30.53 km
VillaSarmiento1	304	12	140	258	1 seg	0	1	24.17 km
VillaSarmiento2	600	134	166	494	1 seg	0	1	24.25 km
VillaSarmiento3	384	76	116	322	1 seg	0	1	25.08 km

Tabla 3: Resultados del procedimiento para los sectores de la Tabla 2.

Los tiempos de resolución no superaron los 20 minutos para ninguno de los sectores. Estos tiempos son razonables para los requerimientos del sistema de reciclado, dado que una vez que se define una ruta, solamente se requieren cambios ocasionales en el futuro. Los gaps de optimalidad son pequeños o nulos, mostrando que el límite de 10 minutos para la resolución de SCIP es adecuado para estas instancias.

Es importante notar que las rutas obtenidas por este procedimiento fueron bien recibidas por la municipalidad. Se realizaron estudios de campo antes de implementar efectivamente las rutas para analizar su factibilidad, y las reacciones de los usuarios fueron positivas.

Cabe notar que en todos los casos las rutas óptimas obtenidas pueden ser recorridas en el tiempo disponible. Si variaran las instancias a resolver se podría dar la situación de que algunos recorridos óptimos fueran demasiado largos y no alcanzara el tiempo para realizarlos. En ese caso una posible variante sería maximizar el número de cuadras a ser visitadas incorporando al modelo una restricción que fije como cota superior el mayor número de cuadras que pueden ser recorridas en el tiempo disponible (y obviamente no obligando a pasar por todas las cuadras).

5. Conclusiones

Se propuso en este trabajo una metodología basada en programación lineal entera para optimizar las rutas de recolección de residuos reciclables en una municipalidad del Gran Buenos Aires. El problema a resolver corresponde a una variante del problema del cartero rural abierto, con restricciones que reflejan las reglas de tránsito de Argentina. El procedimiento presentado en este trabajo permite manejar prohibiciones de giros en algunas esquinas con semáforos, prohibiciones de giros en U, la selección del punto de inicio de la ruta, la inclusión de bandas perimetrales alrededor del sector a recolectar, y el sentido de circulación de las calles.

Las rutas de los vehículos de reciclado generadas por esta metodología e implementadas por la municipalidad tienen una performance significativamente mejor que las rutas diseñadas manualmente que estaban en uso antes de este trabajo. El mayor avance está dado en la cobertura provista por el servicio de recolección, que ahora llega a todas las cuadras de la municipalidad. Antes de esta implementación, hasta 60 cuadras se salteaban en algunos sectores. En algunos sectores, se logró esta mejora disminuyendo también la distancia total recorrida, con relación a las rutas manuales. En unos pocos casos la distancia total aumentó levemente.

Luego de varios meses de testeo de las rutas definidas por el procedimiento propuesto en este trabajo, la municipalidad comenzó su implementación a mediados de 2014. Esta implementación llevó a una mejor cobertura de la municipalidad, y a una notable reducción en los reclamos relacionados con la recolección de residuos reciclables. Debido al éxito de este proyecto, los autores han sido invitados por el Ministerio de Interior y Transporte de la Nación para reproducir la experiencia en otras municipalidades a lo largo del país.

Agradecimientos: Este estudio fue financiado parcialmente por los proyectos UBACyT 20020130100808BA (Argentina), ANPCyT PICT 2012-1324 (Argentina) y FONDECyT 1140787 (Chile), por el Instituto Milenio de Sistemas Complejos de Ingeniería (Chile) y por la firma papyro.com (Argentina). Los autores quisieran agradecer a la municipalidad de Morón, por su apoyo a este proyecto, y al revisor anónimo, por sus muy interesantes observaciones que permitieron mejorar la versión final de este trabajo.

Referencias

[1] C. Arribas, C. Blazquez, y A. Lamas. Urban solid waste collection systems using mathematical modelling and tools of geographic information systems. *Waste Management & Research*, 24(4):355–363, 2010.

[2] F. Bonomo, G. Durán, F. Larumbe, y J. Marenco. A method for optimizing waste collection using mathematical programming: A buenos aires case study. *Waste Management & Research*, 30(3):311–324, 2012.

[3] N. Chang, H. Lu, y L. Wei. GIS technology for vehicle routing and scheduling in solid waste collection systems. *Journal of Environmental Engineering*, 123:901–933, 1997.

[4] N. Christofides, V. Campos, A. Corberán, y R. Mota. An algorithm for the rural postman problem on a directed graph. *Mathematical Programming Study*, 26:155–166, 1986.

[5] A. Corberán, R. Martí, E. Martínez, y D. Soler. The rural postman problem on mixed graphs with turn penalties. *Computers and Operations Research*, 29(7):887–903, 2002.

[6] A. Corberán, R. Martí, y A. Romero. Heuristics for the mixed rural postman problem. *Computers and Operations Research*, 27(2):183–203, 2000.

[7] A. Corberán, E. Motta, y J. Sanchis. A comparison of two different formulations for arc routing problems on mixed graphs. *Computers and Operations Research*, 33(12):3384–3402, 2006.

[8] A. Corberán, I. Plana, y J. Sanchis. The rural postman problem on directed, mixed, and windy graphs. *In: Arc Routing: Problems, Methods, and Applications,MOS-SIAM Series on Optimization, MO20*, pages 101–127, 2014.

[9] H. Eiselt, M. Gendreau, y G. Laporte. Arc routing problems, part ii: The rural postman problem. *Operations Research*, 43(3):399–414, 1995.

[10] D. Eisenstein y A. Iyer. Garbage collection in Chicago: a dynamic scheduling model. *Management Science*, 43:922–933, 1997.

[11] B. Kim, S. Kim, y S. Sahoo. Waste collection vehicle routing problem with time windows. *Computers and Operations Research*, 33:3624–3642, 2006.

[12] M. Mourao y M. Almeida. Lower-bounding and heuristic methods for a refuse collection vehicle routing problem. *European Journal of Operational Research*, 121:420–434, 2000.

[13] W. Pearn y T. Wu. Algorithms for the rural postman problem. *Computers and Operations Research*, 22(8):819–828, 1995.

[14] A. Rodrigues y J. Ferreira. Solving the rural postman problem by memetic algorithms. *Proceedings of MIC'2001 - 4th Metaheuristics International Conference*, pages 679–683, 2001.

[15] H. Thimbleby. The directed chinese postman problem. *Software - Practice and Experience*, 33(11):1081–1096, 2003.

[16] J. Yeomans, G. Huang, y R. Yoogalingam. Combining simulation with evolutionary algorithms for optimal planning under uncertainty: An application to municipal solid waste management planning in the regional municipality of hamilton-wentworth. *Journal of Environmental Informatics*, 2(1):11–30, 2003.

Predicción de la intención de click del usuario Web, usando análisis de dilatación pupilar

Gino Slanzi R. *
Joaquín Jadue M. *
Juan D. Velásquez S. *

Resumen

Se propone un nuevo enfoque para predecir la intención de click del usuario Web, usando datos de dilatación pupilar generados por un dispositivo de *eye-tracking*. El objetivo es determinar si esta variable permite diferenciar estados de elección y no-elección de objetos Web, y de ser así, generar un modelo de clasificación para predecir la elección entendida como un click. Para esto, se realizó un experimento con 25 sujetos saludables, en que la posición ocular y la dilatación pupilar fue capturada mientras los sujetos realizaban tareas de elección entre diferentes objetos en un sitio Web simulado. Los resultados del análisis muestran que existe una diferencia significativa entre los tamaños de la pupila para los objetos escogidos frente a los no escogidos. Además, se creó un modelo de predicción de intención de click, basado en Redes Neuronales Artificiales, que obtuvo un 82 % de *accuracy*. Estos resultados sugieren que esta variable puede ser usada en la perspectiva de *Web Intelligence* como una aproximación del comportamiento del usuario Web, para generar un sistema de recomendaciones para mejorar la estructura y contenidos de sitios Web.

Palabras Clave: Comportamiento usuario web, Dilatación pupilar, Predicción de click, Eye-tracking.

*Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile.

1. Introducción

Durante los últimos 20 años, la Web ha estado creciendo en uso y penetración. Actualmente, es casi natural utilizarla diariamente para buscar cualquier tipo de información, para acceder a productos y servicios y para interactuar socialmente mediante diferentes plataformas. Dado lo anterior, las empresas y organizaciones buscan aumentar su presencia en esta red para atraer y retener clientes, ganando posición de mercado y subir niveles de venta.

Esta meta puede alcanzarse mediante la creación de sitios Web que resulten más efectivos que sus potenciales competidores en capturar la atención y preferencias de sus usuarios.

El conocimiento acerca de las necesidades de los potenciales clientes y la habilidad para establecer servicios perzonalizados que satisfagan esas necesidades, es un aspecto clave para las empresas y organizaciones para destacar frente al resto [24]. En estos tiempos, para un vendedor es posible perzonalizar su producto para clientes individuales en una escala masiva, fenómeno denominado como customización masiva [27].

La personalización es importante por cuanto permite preparar de mejor manera la oferta en base a la extrapolación de las conductas de navegación y preferencias a partir de la caracterización del usuario.

Web usage mining es el "proceso en el cual se aplican técnicas de minería de datos para el descubrimiento de patrones de uso desde los datos generados en la Web" [27]. Esta disciplina utiliza diversas metodologías para el descubrimiento de la experiencia de los usuarios mientras navegan en los sitios Web, tales como el análisis de los *Web logs*, cuestionarios y encuestas [18]. Estas fuentes de datos han sido complementadas con tecnologías más recientes, como *mouse* y *eye tracking*, para obtener un modelamiento más preciso del comportamiento de los usuarios.

Los enfoques basados en neurodatos han adquirido mayor relevancia gracias al hecho que las respuestas fisiológicas a menudo sirven para explicar diferentes conductas del ser humano [19, 20, 21]. En particular, los cambios en la dilatación pupilar han sido relacionados con la carga cognitiva o la actividad mental [8], y los movimientos oculares y fijaciones han sido vinculados con enfoque y atención [30, 13].

Este estudio busca introducir un novedoso enfoque para la predicción de la elección del usuario Web, usando análisis de dilatación pupilar. Para esto, basado en el hecho que el tamaño pupilar varía en el tiempo dependiendo de los estímulos visuales a los que se enfrentan los ojos, se plantea la siguiente

hipótesis: "*Los cambios en los tamaños pupilares en el tiempo corresponden a una variable predictiva de la intención de click de los usuarios Web*. Para validarla, se condujo un experimento en el que 25 sujetos saludables navegaron por un sitio Web simulado, realizando tareas de elección entre diferentes objetos, mientras se registraba su posición ocular y dilatación pupilar, para posteriormente aplicar técnicas de minería de datos.

El paper está estructurado como sigue: La Sección 2 describe estudios relacionados con nuestra propuesta de investigación, explicando como el proceso de elección del usuario Web ha sido modelado, usando variados puntos de vista y cómo las técnicas de *eye tracking* han ganado presencia en este campo. Posteriormente, la Sección 3 muestra detalladamente nuestra propuesta, estableciendo qué modelos y métricas fueron usados junto a las preguntas de investigación que fueron planteadas para validar la hipótesis. En la Sección 4 se entrega una descripción de los experimentos realizados para la adquisición de los datos, mientras que su tratamiento y los resultados son ilustrados en la Sección 5. Finalmente, la Sección 6 presenta las conclusiones finales del trabajo en conjunto con posibles líneas de trabajo futuro.

2. Trabajo relacionado

Esta Sección revisará las principales líneas de investigación relacionadas con el modelamiento y predicción de la intención de click, utilizando variables biológicas como dilatación pupilar y posicionamiento ocular.

Adicionalmente, Buscher et al. introdujeron el concepto de *fixation impact* que permite la identificación de un conjunto de elementos que están bajo la mirada en un cierto instante de tiempo. La definición de este concepto se basa en que la visión humana está definida por una estrecha ventana de alta agudeza junto con el area de visión. Entonces, cuando se está mirando a un objeto, el sujeto está también mirando elementos que están rodeando al objeto principal. De esa manera, dada una fijación, una zona DOM es seleccionada para identificar cuáles elementos están demarcados por ésta. Luego, un puntaje de distancia es asignado a cada elemento según su cobertura, asumiendo una distribución Gaussiana. El *fixation impact* es calculado usando esta distancia e incorporando una dimensión de tiempo relativa al tiempo de la fijación.

Paralelamente, Loyola propone una caracterización de la posición ocular del usuario Web basada en Teoría de Grafos, concluyendo que sus resultados sugieren que un enfoque basado en grafos puede capturar, en una manera confiable, la dinámica del comportamiento del usuario y la identificación de

objetos salientes dentro de un sitio Web [13, 12].

Román y Velásquez en [22] crean un modelo de uso Web inspirado por una descripción neurofisiológica estocástica de la toma de decisión y la utilidad del contenido de las páginas Web. En su estudio, implementan un proceso estocástico de alta dimensión basado en el modelo neuronal *leaky competing accumulator* (LCA), logrando una efectividad del 73 %.

Por otro lado, en [23], los autores revisaron variados métodos para modelar Juicio y Toma de Decisión (*JDM, Judment and Decision Making* en inglés). En particular, exploraron las principales contribuciones del uso de los movimientos oculares en ese campo. Describieron diversos estudios, destacando que el uso de las fijaciones como la principal característica a analizar.

Otro tipo de estudios intentan relacionar la elección o clicks del usuario Web con diferentes variables. Por ejemplo, en 2007, Chandon et al. realizaron un experimento de *eye-tracking* para analizar situaciones de elección de objetos asociados a distintas marcas. Concluyeron que la atención visual es relevante en el proceso de elección de los usuarios, sugiriendo que los objetos con baja probabilidad de ser escogidos pueden ser resaltados poniendolos cerca a los objetos de mayor probabilidad [5].

Reutskaja et al. estudiaron el comportamiento de los usuarios mientras escogían entre objetos bajo condiciones de tiempo y carga usando tecnologías de *eye-tracking*. De su trabajo concluyeron que los objetos ubicados en el centro de la pantalla tienen mayor probabilidad de ser elegidos que los objetos de similares características ubicados en otras zonas de la pantalla. Esto puede permitir inluenciar decisiones, centrando aquellos objetos que se pretende sean escogidos. Adicionalmente, concluyeron que el 70 % de los objetos que fueron elegidos tuvieron más largas fijaciones [16].

Otro estudio fue realizado por Krajbich et al. [9] donde se intentó relacionar el proceso de elección con la posición ocular de los usuarios. Particularmente, desarrollaron un modelo de predicción de elección basado en tres observaciones principales: *a)* la primera y la última fijación son más cortas que las centrales, aunque esto no afecta la probabilidad de elección de cada objeto; *b)* el último objeto visto tiene una mayor probabilidad de ser elegido que el resto; *c)* los objetos con fijaciones más largas tienen mayores probabilidades de ser seleccionados.

Las tecnologías de *eye-tracking* pueden ser usadas para recolectar otro tipo de datos como la dilatación pupilar a través del tiempo. Esta variable está directamente relacionada con diferentes procesos cognitivos, ya que está enlazada al sistema simpático y parasimpático. En la literatura existen variados trabajos que utilizan esta variable para describir diferentes tipos de fenómenos. Por ejemplo, Beatty [1] utilizó el tamaño de la pupila para medir el esfuerzo mental

relativo a tareas cognitivas; Steinhauer et al. encontraron que a mayor complejidad en diversas tareas, mayor era la dilatación pupilar [28]; además, Bradley et al. estudiaron la relación del tamaño pupilar con la excitación emocional producida por estímulos visuales [2].

Dada la evidencia empírica mostrada en esta sección, se encontró un campo de acción específico donde se puede centrar la presente investigación para relacionar la intención de click con la dilatación pupilar a través del tiempo. En particular, se propone un enfoque novedoso para predecir la intención de click usando un análisis de curvas de tamaño pupilar en la navegación.

Este enfoque es interesante debido a que puede ser visto como un complemento a otro tipo de análisis de comportamiento de usuarios en la Web, como el análisis de *clickstream data*, es decir, qué caminos siguen los usuarios por medio de los clicks en los sitios Web y qué conclusiones se pueden obtener con este tipo de estudios.

En este ámbito, Bucklin y Sismeiro en [3] desarrollaron un modelo de comportamiento de navegación de usuarios en un sitio Web tomando en cuenta la decisión de seguir navegando (no salir del sitio) y el tiempo empleado en cada página. Como resultado notaron que la propensión a seguir explorando cambia dinámicamente con la profundidad y la repetición en las visitas, además si aumenta la cantidad de visitas, la cantidad de páginas visitadas se reduce, no así el tiempo de navegación.

En [15] los autores concluyen que los caminos seguidos por los usuarios pueden reflejar de buena forma sus objetivos en los sitios Web, lo que puede servir en la predicción de la conversión de compra. Por otro lado, Moe et al. implementaron un modelo para la evolución del comportamiento de visita a sitios Web, basado en *clickstream data*. Obtuvieron resultados lógicos, como la existencia de un cambio en la conducta de navegación relacionado con la experiencia del usuario, o como el aumento en la propensión de compra dado un aumento en la frecuencia de visita al sitio. Concluyen que la evolución (los cambios) en el comportamiento de visita a los sitios otorga información valiosa de qué tipo de clientes van a comprar en mayor cantidad [14].

Si bien el estudio de los datos de navegación por medio de clicks es un enfoque importante en la investigación del comportamiento del usuario Web, no considera aspectos que son interesantes de evaluar como las respuestas fisiológicas de los usuarios frente a los estímulos presentados en los sitios Web. En particular, el análisis de la secuencia de clicks puede entregar resultados valiosos desde un cierto punto de vista, es decir, los usuarios tienen un comportamiento en la navegación dependiendo a intereses y objetivos y en función de las opciones mostradas en el sitio. Sin embargo, este enfoque puede ser mejorado utilizando una retroalimentación desde otro punto de vista: analizando

si efectivamente las opciones mostradas son las que mejor se acomodan a cada tipo de usuario o si realmente están capturando la atención que se quiere captar. Este tipo de conocimiento puede ser encontrado con análisis de variables fisiológicas como la dilatación pupilar o la actividad cerebral.

Lo que busca este estudio es generar un modelo de clasificación de la intención de click basado en el tamaño de la pupila de los usuarios, con un modelo de este tipo, el análisis de *clickstream* puede ser complementado para desarrollar sistemas de recomendación más precisos y personalizados.

3. Propuesta de Investigación

El principal objetivo de este estudio es clasificar y predecir el proceso de elección entendido como un click en un sitio Web, de acuerdo a variables fisiológicas. Para alcanzar este objetivo, se propone un novedoso enfoque basado en *eye-tracking*, en el que se utiliza la dilatación pupilar como variable predictiva de la intención de click en sitio Web simulado.

3.1. Análisis de Dilatación Pupilar

La utilización de un dispositivo de *eye-tracking* permite capturar en tiempo real la visión de los usuarios y además los cambios en los tamaños de las pupilas dentro de intervalos de tiempo. Los datos de dilatación pupilar puden ser representados como un flujo de datos con una componente de tiempo y ser mostrados como una curva de contracciones y dilataciones dependiendo del estímulo mostrado y el proceso cognitivo implementado como respuesta.

Basado en lo anterior, se declara la siguiente hipótesis de investigación: **Los cambios en los tamaños pupilares en el tiempo corresponden a una variable predictiva de la intención de click de los usuarios Web**. En este sentido, se utiliza la curva de la dilatación pupilar para caracterizar los estados de elección y no-elección, entendiendo una elección como: **el acto visible en el que un usuario realiza un click en uno de los objetos presentados como opción en un instante de tiempo determinado, dada como instrucción la navegación y elección libre**.

La hipótesis fue propuesta de manera que permita responder dos preguntas de investigación principales:

- **Pregunta 1: ¿Es la *Dilatación Pupilar* una variable que sirve para caracterizar estados de elección y no-elección para usuarios Web?**

- **Pregunta 2: ¿Es posible generar un modelo de predicción de intención del click basado en esta variable?**

Para responder estas preguntas de investigación, se diseñó e implementó un experimento en el que se recolectaron los datos del posicionamiento ocular y dilatación pupilar de los usuarios mientras realizaban tareas de elección en un sitio Web simulado.

3.2. Modelo de Clasificación

Se propone el uso de modelos de clasificación binaria para predecir la intención de click. Luego, se utilizan algoritmos de aprendizaje supervisado, donde el input es la curva de dilatación pupilar y el output es el parámetro de decisión con valor 1 para *elección* y 0 para *no-elección*. En ese sentido, el modelo considera tres algoritmos típicamente usados en *Web mining*, como Redes Neuronales Artificiales, Regresión Logística y Support Vector Machine.

Para los tres modelos se usa el 70 % de los datos como set de entrenamiento y el 30 % restante, como set de prueba. A continuación se describen rápidamente los algoritmos usados.

Para evaluar la performance de cada modelo de clasificación, se utilizan dos ratios derivados de la Matriz de Confusión: *Accuracy* y *Recall*. El primero sirve para medir cuán asertivo fue el modelo en todas sus predicciones, mientras que el segundo permite medir qué tan asertivo fue entre el total de casos positivos.

4. Recolección de Datos

Para obtener los datos necesarios, se realizó un experimento que considera los diferentes aspectos que permiten reproducir el proceso de elección del usuario Web, mientras se monitorea y recolecta los datos de dilatación pupilar y posicionamiento ocular. Esta etapa experimental fue llevada a cabo en el Laboratorio de Neurosistemas de la Facultad de Medicina de la Universidad de Chile.

4.1. Diseño Experimental

La serie de experimentos fue diseñada considerando los siguientes aspectos:

4.1.1. Grupo Experimental

El grupo consistió en 25 sujetos sanos (11 mujeres y 14 mujeres), estudiantes y profesionales de diversas áreas y disciplinas. La edad promedio del grupo

fue 26,1 años con una varianza de 2,2 años. Todos los sujetos declararon tener visión correcta y no presentaban enfermedades psiquiátricas o neurológicas que pudiesen interferir con los propósitos del experimento. Todos los sujetos firmaron un consentimiento informado aprobado por el *Comité de Ética de la Facultad de Medicina de la Universidad de Chile*.

4.1.2. Instrumentos

Se requiere un dispositivo de *eye tracking* para grabar el posicionamiento ocular y la dilatación pupilar durante las tareas del experimento. Para esto se utilizó el *SR Research Eye Link 1000* [26] a una tasa de sampleo de 1000 Hz. Por otro lado, los estímulos visuales, es decir, el sitio Web simulado, fue presentado en una pantalla *LG* de $32''$ situada a 60 centímetros del sujeto en una habitación experimental. Cada sujeto debía apoyar el mentón a un soporte para mantener la cabeza lo más quieta posible. La habitación experimental tenía la luz apagada durante el experimento. La Figura 1 muestra el *set up* donde los usuarios tomaron las pruebas, donde el dispositivo de *eye tracking*, el soporte y la pantalla pueden ser vistos claramente.

Figura 1: Set up del Experimento.

4.1.3. Protocolo

El sitio Web simulado está compuesto de tres partes principales: página de inicio, páginas de elección y página de salida, tal como se ve en la Figura 2. Cada parte está descrita a continuación:

1. Página de Inicio: Primero que todo, una se muestra una página de inicio con las instrucciones del experimento.

2. Páginas de Elección: En esta etapa, una página de forma de grilla que contiene 9 objetos de una misma categoría es mostrada para que el usuario eliga uno de los nueve objetos. Todos los objetos son tomados de la base de datos de la *International Affective Picture System (IAPS)* [10] y corresponden al tipo de valencia neutral. Los sujetos tienen que hacer click sobre uno de los nueve objetos para elegirlo. Una vez que el usuario selecciona un objeto, una imagen de ruido rosa es presentada por 2 segundos, para presentar una nueva grilla de 9 objetos nuevamente. Este proceso es repetido 90 veces, para el que la instrucción explícita es: **Usando el mouse, elija uno de los objetos mostrados en la página, haciendo un click sobre él**.

3. Página de Salida: Finalmente, una página de despedida y agradecimiento es mostrada en pantalla.

Además, la serie de experimentos fue conducida según el siguiente protocolo experimental, que fue presentado con el software *SR Research Experiment Builder* [25]:

1. Bienvenida: El sujeto es agradecido por su participación en el experimento, se le explica el procedimiento y se le pide que lea y firme el consentimiento informado.

2. Los instrumentos deben estar encendidos y funcionando correctamente.

3. El sujeto es sentado de manera cómoda, apoyando el mentón en el soporte frente a la pantalla a unos $60 - 80$ cm, asegurándose que los ojos coincidan con el centro de ésta.

4. El dispositivo de *eye-tracking* debe ser calibrado para el funcionamiento perfecto

5. Una vez realizada la calibración, empieza el experimento con la página de inicio, donde el sujeto lee las instrucciones.

6. Luego de las 90 categorías, se muestra la página de salida y el experimento es finalizado.

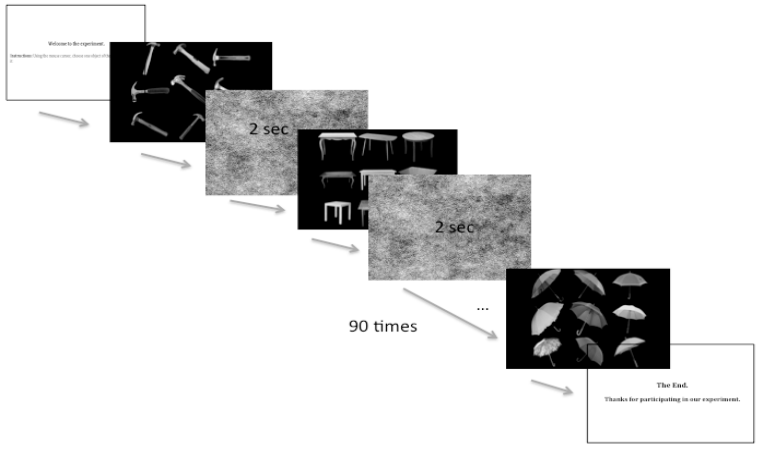


Figura 2: Secuencia del Experimento.

5. Resultados y Discusión

Posterior a la adquisición de los datos, se efectuó un análisis para poder responder a las preguntas de investigación propuestas en 3.1. En esta sección se describe dicho proceso y se entregan respuestas a esas interrogantes, discutiendo resultados y rendimientos.

5.1. Preprocesamiento y Transformación de los Datos

Es importante eliminar ruidos que afecten el resultado del estudio, por lo tanto, primero que todo, cada curva de dilatación pupilar fue preprocesada interpolando linealmente los pestañeos, ya que cada vez que un sujeto cierra los ojos, esos datos se pierden. Posteriormente, se arreglaron los errores producidos por las sacadas y se utilizó un filtro pasa-bajo de 2 Hz, que permite suavizar la curva de dilatación pupilar, descartando las frecuencias mayores al umbral impuesto, eliminando ruidos y artefactos debidos a errores de medición del hardware.

Luego de tener los datos limpios, se definió una *observación* considerando los siguientes aspectos:

- Una observación empieza cuando el sujeto realiza una fijación sobre un objeto y termina cuando empieza otra fijación en otro objeto.
- Se estableció un umbral mínimo de tiempo de fijación de 300 milisegundos para ser considerada observación. De otra manera, sería considerada

como que el sujeto no prestó atención en ese objeto y sólo paso por encima durante una sacada.

- Para el análisis se consideraron los primeros 600 milisegundos de cada observación.

Las observaciones fueron transformadas usando el *Z-score* como forma de estandarizarlas y hacer sencilla la comparación entre los sujetos. Finalmente, se centraron las observaciones, removiendo una línea base de los 200 milisegundos previos a la observación.

5.2. Predicción de Click

La predicción del click fue analizada en dos pasos. El primero intentó responder la pregunta de investigación 1, es decir, intentar determinar si es que existen diferencias entre los tamaños pupilares para elección y no-elección. Luego, se aplica el modelo de clasificación para responder la pregunta 2, esto es, poder predecir la intención de click, utilizando la dilatación pupilar como input.

5.2.1. Pregunta 1 - Diferencia entre elección y no-elección

Usando los datos de *eye-tracking* fue posible de etiquetar las observaciones en dos grupos, *elección* y *no-elección*. Luego, un promedio fue calculado para todas las observaciones de todos los sujetos para cada grupo. La Figura 3 muestra las dos curvas con sus respectivos intervalos de confianza, donde se pueden ver gráficamente las diferencias entre los tamaños pupilares para elección y no-elección; la curva de elección (en azul), es mayor que la curva de no elección (en rojo).

Para validar estadísticamente esta diferencia, se llevó a cabo un Test *Lilliefors* de normalidad, para ver si ambas curvas presentaban una distribución normal. Con un 95 % de confianza, el test entregó un *p-valor* de 0,0970, mayor que el valor crítico de 0,0518, lo que implica que estas curvas no están normalmente distribuidas. Por lo tanto, para determinar si son estadísticamente diferentes, se realizó un Test *Wilcoxon*, cuyo resultado, con un 95 % de confianza, permitió rechazar la hipótesis nula que plantea que las curvas poseen la misma media.

En otras palabras, se pudo validar la hipótesis, debido a que para los objetos que fueron visualmente explorados y *elegidos* (clickeados), la dilatación pupilar due estadísticamente mayor que para aquellos que no fueron seleccionados.

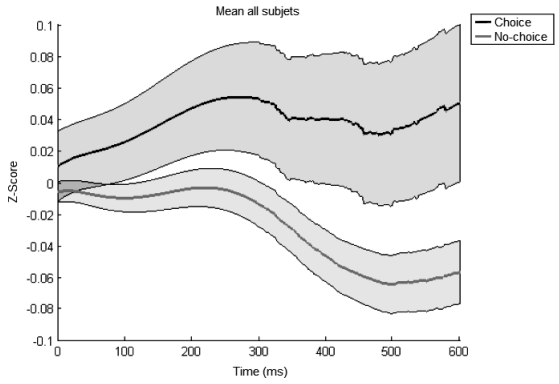


Figura 3: Diferencia entre *elección* y *no-elección*.

5.2.2. Pregunta 2 - Modelo Predictivo

Finalmente, se implementó el modelo de predicción declarado en 3.2. Cada algoritmo fue utilizado por separado, usando una distribución de $70-30\,\%$ de los datos para entrenamiento y prueba respectivamente. La idea era generar más de una opción para la predicción y así analizar distintos resultados.

El primer algoritmo empleado fue la Regresión Logística, para la cual se utilizó la función logística estándar para minimizar la función de costos y así obtener los valores beta de la regresión. Luego, probando los resultados, se obtuvo un nivel de *accuray* del $75\,\%$ y un *recall* del $12\,\%$.

Enseguida, se implementó el algoritmo SVM junto con una función de kernel *Gaussian Radial Basis* con un factor de escalamiento σ 1. De forma similar al caso anterior, este algoritmo presentó un *accuracy* aceptable del $72\,\%$ y un bajo *recall* del $15\,\%$.

Para terminar, se aplicó el algoritmo RNA, donde se utilizó la función *Log-Sigmoid Transfer* como función de activación para cada capa, definida como $A = logsig(n)$, donde n corresponde a datos de input. Este modelo fue el que entregó los mejores resultados con un *accuracy* del $82\,\%$ y un *recall* del $19\,\%$.

En definitiva, el modelo que se propone para estudiar y predecir la intención de click en un sitio Web es usar datos de dilatación pupilar generados por un dispositivo de *eye-tracking* para la aplicación de un algortimo de RNA. Este algoritmo es preferido sobre los demás, porque entregó el mejor resultado en cuanto a *accuracy* y *recall*, como se puede apreciar en la Tabla 1.

Es importante destacar que aun cuando los tres modelos entregaron altos niveles de *accuracy*, los valores de *recall* son bajos. Esto se debe a que los casos en que la variable objetivo toma el valor de 1 (elección), son eventos raros dentro de la totalidad del conjunto de observaciones. De esa manera, el clasificador intentó predecir de mejor manera cuando la variable objetivo toma

Tabla 1: Resultados de los modelos de predicción.

Modelo	**Accuracy**	**Recall**
Regresión Logística	75 %	12 %
Support Vector Machine	72 %	15 %
Redes Neuronales Artificiales	82 %	19 %

el valor de 0 (no-elección), entonces debido a que ese tipo de observaciones equivalen en cantidad a diez veces al tipo de elección, el modelo entrega como resultado un elevado *accuracy* y un disminuido *recall*.

5.3. Discusión General

Como es posible ver en 5.2.1, la dilatación pupilar muestra patrones de comportamiento determinados que diferencian la elección de la no elección. Sin embargo, dados los resultados del modelo de predicción, se puede interpretar que el fenómeno en estudio no fue clasificado correctamente. Esto es porque si la elección, estudiada como en este trabajo, puede ser clasificada como un evento extraño, los modelos típicos de clasificación no actúan de manera perfecta con este tipo de datos.

No obstante, existen diferentes maneras para solucionar este problema, como usar modelos de clasificación de eventos raros o aplicar técnicas de transformación como sobre-muestreo de los datos de casos extraños; tomar un subconjunto de casos normales que calce con la cantidad de casos extraños; o cambiar los costos en la matriz de confusión, otorgando mayores costos a la función objetivo cuando falle en la clasificación de eventos raros.

Además, la variabilidad de la dilatación pupilar entre sujetos es alta como para validar la hipótesis individualmente. De todas maneras, el promedio entre todos los sujetos muestra que la tendencia de las curvas permite separar la decisión de elegir y no elegir en dos grupos de manera estadísticamente significativa. Basado en esto, diversos tipos de análisis de curvas pueden ser realizados para extender y corroborar estos resultados. Determinar cuáles son las características más importantes de la curva de dilatación pupilar podría ser útil para generar modelos más empíricos para predecir la intención de click del usuario Web, por ejemplo, obtener variables tales como la máxima dilatación y mínima contracción, o velocidad y aceleración de la curva.

6. Conclusiones y Trabajo Futuro

En este trabajo se ha explorado la relación entre la dilatación pupilar y la intención de click del usuario Web, entendida como "el acto visible en el que un usuario realiza un click en uno de los objetos presentados como opción en un instante de tiempo determinado, dada como instrucción la navegación y elección libre". Para la recolección de los datos necesarios, se condujo un experimento que consistió en recopilar los datos de posición ocular y dilatación pupilar con un dispositivo de *eye-tracking*, mientras los sujetos realizaban tareas de elección de objetos en un sitio Web simulado.

Considerando diferentes aspectos, se definió una observación de un objeto y se compararon las observaciones correspondientes a objetos seleccionados y no seleccionados por los usuarios. Se encontró una diferencia significativa entre los tamaños pupilares de esos objetos. Más precisamente, los objetos elegidos presentaron mayores tamaños de pupila que los elementos que no fueron escogidos; hecho que permitió validar la hipótesis de investigación planteada para este estudio.

Además, se propuso un modelo de predicción basado en el uso de Redes Neuronales Artificiales. Cada observación fue etiquetada como 1 o 0 para elección y no-elección respectivamente. Aunque el modelo se comportó bien en términos de *accuracy*, los valores de *recall* fueron bajos. Esto significa que no se pudo generar un modelo de calidad, ya que el fenómeno en estudio no fue clasificado correctamente.

Como trabajo futuro se proponen tres líneas de desarrollo. Primero, mejorar los resultados de la predicción mediante la aplicación de técnicas como sobre muestreo de casos raros o cambiando los costos de la matriz de confusión. Segundo, tratar de caracterizar la curva de dilatación pupilar con diversas *features* como dilatación máxima y mínima contracción o velocidad y aceleración de la curva. Finalmente, se propone complementar el análisis de *eye-tracking* con el uso de un dispositivo de electroencelografía (EEG), para analizar las ondas del cerebro y poder determinar qué es lo que está realmente pasando al momento de tomar decisiones de elección a un nivel más cognitivo.

Agradecimientos: Este trabajo fue financiado por el proyecto FONDEF-CONICYT IT13I20049 y por el Instituto Sistemas Complejos de Ingeniería (ICM: P-05-004-F, CONICYT: FBO16).

Referencias

[1] J. Beatty. Tas-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, 91(2):276, 1982.

[2] M. Bradley, L. Miccoli, M.A. Escrig, y P.J. Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607, 2008.

[3] R. Bucklin y C. Sismeiro. A model of web site browsing behavior estimated on clickstream data. *Journal of marketing research*, 40(3):249–267, 2003.

[4] G. Buscher, E. Cutrell, y M. Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30. ACM, 2009.

[5] P. Chandon, W.J. Hutchinson, y S.H. Young. *Measuring the value of point-of-purchase marketing with commercial eye-tracking data*. INSEAD, 2001.

[6] L.E. Dujovne y J.D. Velásquez. Design and implementation of a methodology for identifying website keyobjects. In *Proceedings of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 301–308. KES, 2009.

[7] L. González y J.D. Velásquez. Una aplicación de herramientas de eye-tracking para analizar las preferencias de contenido de los usuarios de sitios web. *Revista de Ingeniera de Sistemas*, 26(1):95–118, September 2012.

[8] E. Hess y J. Polt. Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611):1190–1192, 1964.

[9] I. Krajbich, C. Armel, y A. Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10):1292–1298, 2010.

[10] P. Lang, M. Bradley, y B. Cuthbert. International affective picture system (iaps): Technical manual and affective ratings, 1999.

[11] P. Loyola, G. Martínez, y J.D Velásquez. Caracterizando los patrones de la mirada del usuario web: Una aproximacin basada en teora de grafos. *Revista de Ingeniera de Sistemas*, pages 87–107, 2014.

[12] P. Loyola, P.E. Román, y J.D. Velásquez. Predicting web user behavior using learning-based ant colony optimization. *Engineering Applications of Artificial Intelligence*, 25(5):889–897, 8 2012.

[13] P. Loyola y J.D. Velásquez. Characterizing web user visual gaze patterns: A graph theory inspired approach. In Dominik Slezak, Ah-Hwee Tan, JamesF. Peters, y Lars Schwabe, editors, *Brain Informatics and Health*, volume 8609 of *Lecture Notes in Computer Science*, pages 586–594. Springer International Publishing, 2014.

[14] W. Moe y P. Fader. Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing*, 18(1):5–19, 2004.

[15] A. Montgomery, S. Li, K. Srinivasan, y J. Liechty. Modeling online browsing and path analysis using clickstream data. *Marketing Science*, 23(4):579–595, 2004.

[16] E. Reutskaja, R. Nagel, C.F. Camerer, y A. Rangel. Search dynamics in consumer choice under time pressure: An eye-tracking study. *The American Economic Review*, pages 900–926, 2011.

[17] S. A Ríos, J.D. Velásquez, E. Vera, H. Yasuda, y T. Aoki. Using sofm to improve web site text content. In *Advances in Natural Computation*, pages 622–626. Springer, 2005.

[18] P. Román, G. L'Huillier, y J.D. Velásquez. Web usage mining. In *Advanced Techniques in Web Intelligence-I*, pages 143–165. Springer, 2010.

[19] P. Román y J.D. Velásquez. Cognitive science for web usage analysis. *Advanced Techniques in Web Intelligence-2: Web User Browsing Behaviour and Preference Analysis*, 452:35, 2012.

[20] P. Román y J.D. Velásquez. Cognitive science for web usage analysis. In *Advanced Techniques in Web Intelligence-2*, pages 35–73. Springer Berlin Heidelberg, 2013.

[21] P. Román, J.D. Velásquez, V. Palade, y L. Jain. New trends in web user behaviour analysis. In *Advanced Techniques in Web Intelligence-2*, volume 452 of *Studies in Computational Intelligence*, pages 1–10. Springer Berlin Heidelberg, 2013.

[22] P.E. Román y J.D. Velásquez. A neurology-inspired model of web usage. *Neurocomputing*, 131:300–311, May 2014.

[23] M. Schulte-Mecklenbeck, A. Ḱuhberger, y R. Ranyard. The role of process data in the development and testing of process models of judgment and decision making. *Judgment and Decision Making*, 6(8):733–739, 2011.

[24] M. Spiliopoulou. Web usage mining for web site evaluation. *Commun. ACM*, 43(8):127–134, August 2000.

[25] SR Research Ltd. *SR Research Experiment Builder User Manual*, 1.6.121 edition, 2004-2010.

[26] SR Research Ltd. *EyeLink User Manual*, 1.4.0 edition, 2005-2008.

[27] J. Srivastava, R. Cooley, M. Deshpande, y P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl.*, 1(2):12–23, January 2000.

[28] S.R. Steinhauer, G.J. Siegle, R. Condray, y M. Pless. Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International Journal of Psychophysiology*, 52(1):77 – 86, 2004. Pupillometric Measures of Cognitive and Emotional Processes.

[29] J.D. Velásquez. Web site keywords: A methodology for improving gradually the web site text content. *Intelligent Data Analysis*, 16(2):327–348, 2012.

[30] J.D. Velásquez. Combining eye-tracking technologies with web usage mining for identifying website keyobjects. *Eng. Appl. Artif. Intell.*, 26(5-6):1469–1478, May 2013.

[31] J.D. Velásquez, L.E. Dujovne, y G. L'Huillier. Extracting significant website key objects: A semantic web mining approach. *Engineering Applications of Artificial Intelligence*, 24(8):1532–1541, 2011.

Selección de atributos y support vector machines adaptado al problema de fuga de clientes

Álvaro Flores [*]
Sebastián Maldonado[**]
Richard Weber[*]

Resumen

Uno de los grandes desafíos de la Minería de Datos aplicada al Análisis de Negocios es la selección de atributos para un modelo de clasificación. La mayoría de las técnicas de selección de atributos se basan en criterios de validación estadística, perdiendo en muchos casos el objetivo del negocio en sí mismo, lo que no necesariamente lleva a modelos que optimicen las metas definidas. Para generar el modelo y la selección de atributos se utiliza un enfoque basado en utilidades utilizando el modelo de *Support Vector Machines*, donde las métricas basadas en utilidades simulan la realización de una campaña de retención de clientes considerando beneficios y costos (Maximum Profit Criterion (MPC) y Expected Maximum Profit Criterion (EMPC)) o bien sólo costos, como es el caso de *H-measure*. El enfoque presentado en este trabajo consiste en un método de selección de atributos empotrado en la construcción del modelo clasificador, que apunta a la eliminación secuencial de atributos removiendo los que tienen menor relevancia de acuerdo a estas métricas. Utilizando un caso del área de Telecomunicaciones, los resultados indican que estos métodos de selección de atributos y evaluación de modelos son más estables y obtienen mejores resultados tanto en términos de métricas usuales de evaluación de modelos predictivos, como en métricas de desempeño basadas en utilidades orientadas al negocio.

Palabras Clave: Fuga de Clientes, Selección de atributos, *Support Vector Machines*.

[*]Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile.

[**]Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Santiago, Chile.

1. Introducción

La clasificación es una tarea relevante en muchas aplicaciones orientadas a mejorar las utilidades de una empresa, tales como *Credit Scoring* o Predicción de Fuga de Clientes [2]. Además, se ha demostrado que el desempeño de un clasificador puede ser mejorado enfocándose en los atributos más relevantes usados para la construcción de éste. La selección de atributos tiene importantes ventajas:

1. Una representación usando menos atributos realza el poder predictivo de los modelos de clasificación disminuyendo su complejidad, reduciendo de esta manera el riesgo de *Overfitting (sobreajuste)*, causado por la *Maldición de la Dimensionalidad* [20].

2. Dicha selección permite una mejor interpretación del clasificador, lo que es particularmente importante en *Business Analytics*, puesto que muchos profesionales consideran que las técnicas de *Machine Learning* son *cajas negras* y se rehúsan a emplear estos métodos debido a su complejidad [2].

En este trabajo se propone abordar la problemática de fuga de clientes en telecomunicaciones, desde un punto de vista proactivo, generando un listado de clientes candidatos a ser contactados para una campaña de retención, maximizando las utilidades que ésta generaría. Para esto se compara un estado basal equivalente a no realizar acción comercial alguna, versus efectuar una campaña de retención focalizada en el conjunto de clientes que el clasificador determina como potenciales fugas. Se propone un método de selección de atributos que se incluye en la construcción del modelo de clasificación, usando la herramienta *Support Vector Machine* y proponiendo métricas de desempeño asociadas a utilidades que permiten discriminar las variables que aportan más a diferenciar los clientes prospectos a fugarse de la compañía.

Tratar de predecir qué clientes van a dejar una compañía, fugarse o simplemente *churn*[1], como se conoce en la literatura, es una de las tareas más importantes de las empresas de servicio, principalmente la banca y telecomunicaciones. La importancia de la predicción de fuga de cliente se incrementa debido a la creciente cantidad de clientes dispuestos a cambiar sus proveedores, junto a la fuerte competencia por captar a clientes nuevos. Es por esto que surge la necesidad de crear y desarrollar modelos capaces de identificar

[1]En este trabajo se hace referencia indistintamente a fuga o *churn*.

clientes actuales con tendencia a dejar la compañía en un periodo dado de tiempo.

El *Churn* puede ser observado de dos maneras diferentes: **voluntario**, en donde el cliente decide terminar el contrato, o bien **involuntario**, donde la compañía en cuestión decide terminar el contrato con el cliente [3]. En este trabajo se estudia el *churn* como un fenómeno voluntario.

Uno de los objetivos relevantes al realizar modelos de predicción de fuga, es establecer estrategias orientadas a la retención del cliente. Si la compañía es capaz de identificar los posibles *churners*, el siguiente paso es desarrollar campañas comerciales y estrategias de retención enfocadas en este grupo en particular, potenciando de esta manera la lealtad del cliente y obteniendo otros beneficios, como por ejemplo:

- Un incremento en la proporción de clientes fieles, los cuales generan 1,7 veces más ingreso que los otros clientes [12].
- Un impacto directo en la rentabilidad: un 5 % de incremento en la tasa de retención de clientes, puede llevar a un 18 % de reducción en costos operacionales [12].
- Una disminución del gasto en retención innecesario, enfocando los recursos en clientes en riesgo de fuga y no en la base de clientes completa; reduciendo así los costos operacionales y de marketing [25].

Atendiendo a estos hechos, la tasa de fuga es puesta explícitamente en la fórmula para el *Customer Lifetime Value* (en adelante **CLV**), que toma la siguiente forma considerando periodos anuales [3]:

$$CLV = \sum_{t=1}^{\infty} \frac{m(1-c)^{t-1}}{(1+r)^{t-1}} = m\frac{(1+r)}{(r+c)} \tag{1}$$

en donde c es la tasa anual de fuga y m es el retorno esperado medio por cliente. El parámetro r es la tasa de descuento anual. Existen dos maneras clásicas de determinar este último valor. La primera es obtener directamente el *Weighted Average Cost of Capital (WACC)* de la compañía. La segunda, es usar la tasa de descuento del sector industrial en particular. Se puede entender el **CLV** como el valor presente neto del beneficio por cliente, luego un descenso en la tasa de fuga de clientes impacta de manera directa en las utilidades de la empresa.

El fenómeno de la fuga de clientes puede ser modelado con técnicas que dependen del tiempo [3], o bien como predicciones sobre el siguiente periodo.

En el primer caso este tipo de modelos no asume que la fuga ocurrirá explícitamente en un período de tiempo, proponiendo probabilidades de fuga hasta un número fijo de períodos desde el origen de los datos, pudiendo incluso variar con el tiempo [3]. En el segundo caso, encontramos formas de enfocarnos a predecir si un cliente decide o no fugarse en el período siguiente, donde los enfoques más tradicionales son regresión logística [5, 17, 22], modelos estadísticos no paramétricos como *K-nearest neighbors (KNN)* [8], árboles de decisión [31], y redes neuronales [16]. Una revisión sobre la modelación de fuga de clientes, puede ser encontrada en Verbeke et al. [29]. En este trabajo se utilizan clasificadores basados en SVM, prediciendo la fuga de los clientes en el siguiente período de tiempo.

Para poder evaluar el desempeño de un clasificador, es necesario usar alguna medida de rendimiento que permita conocer el grado de asertividad que tiene dicho clasificador, esto a su vez permite comparar entre clasificadores. En la literatura se proponen muchas métricas de rendimiento para evaluar el desempeño de un clasificador. Una revisión más exhaustiva del tema puede ser encontrada en [7].

Cuando se usan modelos de predicción de fuga mensual, la tasa de fuga usualmente se mantiene bajo el 5 % [28], lo que nos lleva naturalmente al problema de desbalance de clases que también será tratado en los modelos propuestos en este artículo.

La estructura de este trabajo es la siguiente: La Sección 2 presenta la derivación del método de clasificación *Support Vector Machines*. Técnicas recientes de selección de atributos para *Support Vector Machines* se presentan en la Sección 3. La Sección 4 describe la metodología para selección de atributos propuesta en este trabajo. La Sección 5 presenta los principales resultados. Finalmente, la Sección 6 muestra las conclusiones del trabajo.

2. *Support Vector Machines*

Support Vector Machine (SVM) [27] es un modelo de clasificación basado en minimizar el error cuadrático de la clasificación, construyendo un hiperplano que separa los datos de la forma más precisa posible. SVM es considerado uno de los modelos más precisos y robustos posibles dentro de los algoritmos de clasificación binaria [32], siendo ampliamente utilizado por su versatilidad y efectividad a la hora de clasificar. Esto último se logra principalmente porque incluye la minimización del error estructural al clasificar nuevos objetos, que está directamente relacionado con uno de los principales objetivos de un

clasificador, que es tener la habilidad de *generalizar* de forma correcta.

El método SVM define un hiperplano en $\mathbb{R}^M$, donde M es la cantidad de atributos, que separa lo mejor posible una clase de la otra. El objetivo es maximizar la distancia entre el hiperplano óptimo y los hiperplanos canónicos (que representan los *bordes* de cada clase). Para lograr esto se minimiza la norma Euclidiana de $\mathbf{w}$, que coresponde a los coeficientes que definen el hiperplano, dando origen al siguiente problema de minimización como formulación primal:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.a.} \quad & y_i \cdot (\mathbf{w}^\top \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \qquad i = 1,\ldots,N, \\ & \xi_i \geq 0, \qquad i = 1,\ldots,N, \end{aligned} \tag{2}$$

donde $\mathbf{x_i} \in \mathbb{R}^{\mathbf{M}}$, $y_i \in \{-1, 1\}$, y ξ_i $(i = 1,\ldots,N)$ son variables de holgura que tienen por objetivo relajar las restricciones, permitiendo que ocurran errores, pero penalizándolos en la función objetivo. Esta penalización se controla con un parámetro C.

La formulación previa puede ser extendida a clasificadores no lineales, usando el *kernel trick*: Los datos de entrenamiento son transformados en un espacio de mayor dimensionalidad $\mathscr{H}$, a través de una función $\phi : \mathbf{x} \to \phi(\mathbf{x}) \in \mathscr{H}$ [23]. Una función de Kernel $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^\top \cdot \phi(\mathbf{y})$ define un producto interno en el espacio $\mathscr{H}$, lo que nos lleva a la siguiente formulación (luego de calcular el dual):

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,s=1}^{N}\alpha_i\alpha_s y_i y_s K(\mathbf{x}_i, \mathbf{x}_s) \\ \text{s.a.} \quad & \sum_{i=1}^{N}\alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, \qquad i = 1,\ldots,N. \end{aligned} \tag{3}$$

3. Selección de Atributos para SVM

En esta sección se hará una revisión del estado del arte en lo que a selección de variables se refiere, y se explican los tres enfoques principales [13], así como los métodos que serán usados en este trabajo.

3.1. Métodos de Filtro (*Filter Methods*)

La aplicación de este enfoque, ocurre antes de aplicar cualquier algoritmo de clasificación, y usa propiedades estadísticas de los atributos, con el objetivo de dejar fuera los que aportan menos *información* (de acuerdo a alguna métrica) al modelo. Ejemplos clásicos de la literatura, son **el estadístico de** χ^2, que mide la dependencia entre la distribución de cada atributo y las etiquetas de las observaciones [26], la **Ganancia de Información** que usa la entropía para medir la relevancia de un atributo [26], y finalmente el **Fisher Score** (F), que estima la relevancia de cada atributo calculando la diferencia (absoluta) entre las medias del valor de la variable en ambas clases, normalizando según la suma de las varianzas intra-clases:

$$F(j) = \left| \frac{\mu_j^+ - \mu_j^-}{(\sigma_j^+)^2 + (\sigma_j^-)^2} \right| \tag{4}$$

en donde μ_j^+ (μ_j^-) es la media del j-ésimo atributo en la clase positiva (negativa) y σ_j^+ (σ_j^-) es la correspondiente desviación estándar.

Usando este indicador, es posible observar qué atributos difieren *más* entre clases. El **Método de Fisher** para selección de atributos consiste en obtener el indicador señalado en la ecuación (4) para todos los atributos y elegir la cantidad de atributos deseada que tengan mejor *Fisher Score*.

3.2. Métodos de Envoltura (*Wrapper Methods*)

Estos métodos buscan entre los posibles subconjuntos de atributos, evaluando su potencial predictivo. Esto es altamente demandante en términos computacionales, puesto que la cantidad de subconjuntos a revisar tiene un tamaño exponencial en la cantidad de atributos. Las estrategias más populares para llevar a cabo esta tarea son *Sequential Forward Selection (SFS)* y *Sequential Backward Elimination (SBE)*. SFS empieza con un conjunto vacío de atributos, y luego intenta agregar variables de manera secuencial, donde en cada paso se agrega la más relevante (de acuerdo a algún método de clasificación en particular) del conjunto de atributos pendientes por agregar. SBE, por su lado, empieza con el conjunto completo de variables, y calcula la significancia estadística de cada una, eliminando en cada iteración la menos relevante.

3.3. Métodos Empotrados (*Embedded Methods*)

Estos métodos realizan la selección de atributos de manera simultánea a la construcción del clasificador y son específicos para cada técnica de clasificación.

Por ende incluyen la interacción entre los atributos y el clasificador en el proceso de modelación. Los métodos *embedded* son computacionalmente menos intensivos que las estrategias *wrapper* [13].

Una técnica muy popular y relevante para el desarrollo de este trabajo es *Recursive Feature Elimination* (RFE-SVM [14]). El objetivo de este método es encontrar un subconjunto de tamaño r entre n variables (con $r < n$), eliminando aquellos atributos cuya extracción contribuye a alcanzar el mayor margen de separación entre clases. Esto puede ser logrado utilizando una estrategia SBE, eliminando de manera secuencial atributos basándose en las componentes del vector de pesos en SVM $\mathbf{w}$. El caso lineal toma la siguiente forma:

Algoritmo 1 *Recursive Feature Elimination, SVM* - Caso Lineal

1. **repetir**
2. $\mathbf{w} \leftarrow$ Entrenamiento SVM (formulación primal).
3. Eliminar el atributo p con el valor más pequeño de $|w_p|$.
4. **hasta** reducir la cantidad de atributos a r.

Cabe notar que el caso RFE-SVM lineal puede extenderse al caso no-lineal (es decir, usando funciones de Kernel), notando que la distancia de los hiperplanos canónicos que separan ambas clases (el **margen**) es inversamente proporcional a la norma del vector de pesos $\mathbf{w}$ y este último valor puede ser escrito en términos de las variables duales del modelo generado por SVM, tomando la siguiente forma funcional:

$$W^2(\alpha) = \sum_{i,s=1}^{N} \alpha_i \alpha_s y_i y_s K(\mathbf{x}_i, \mathbf{x}_s) \tag{5}$$

El atributo removido en cada iteración, es aquel cuya eliminación minimice la variación de $W^2(\boldsymbol{\alpha})$. Este procedimiento queda descrito en el Algoritmo 2.

Algoritmo 2 Recursive Feature Elimination, SVM - Caso no Lineal

1. **repetir**
2. $\mathbf{w} \leftarrow$ Entrenamiento SVM (formulación dual).
3. Eliminar el atributo p con menor valor de $|W^2(\boldsymbol{\alpha}) - W^2_{(-p)}(\boldsymbol{\alpha})|$.
4. **hasta** reducir la cantidad de atributos a r.

4. Metodología propuesta

A continuación se detalla la metodología propuesta para la selección de atributos maximizando el beneficio asociado. En primera instancia, se describen las métricas utilizadas para evaluar el desempeño de una campaña de retención de clientes a partir de la solución entregada por un modelo de clasificación, para luego detallar el algoritmo propuesto.

4.1. Fuga de Clientes y Medidas de desempeño

La fuga de clientes puede ser modelada como una clasificación binaria donde un cliente pertenece a una de dos clases: clientes leales (*non-churners*) o clientes fugados (*churners*). Dado un objeto $\mathbf{x}$ (observación, evento, cliente, etc.) un clasificador $\mathcal{C}$ producirá una puntuación s, donde por convención una puntuación más alta implica que tiene mayor tendencia a ser etiquetado con $(+1)$, es decir, *churn* en nuestro caso. Se fija un valor de **umbral** t para proveer una clasificación binaria de la base completa basada en sus puntajes. De esta manera todas las instancias que tengan una puntuación s menor que t son clasificados como *non-churners* (-1) y los clientes con s mayor o igual a t son clasificados como *churners* $(+1)$.

Se consideran las siguientes definiciones (notación presentada en [28]):

- **Probabilidades a Priori**: π_{-1} y π_1 son las probabilidades a priori de que una observación posea la etiqueta -1 o 1, respectivamente. Notemos que $\pi_{-1} + \pi_1 = 1$, es decir, son las únicas posibilidades para una observación.
- **Distribuciones de Probabilidades**: Dado una puntuación s, las funciones de densidad de probabilidad para los *non-churners* y los *churners*

son respectivamente $f_{-1}(s)$ y $f_1(s)$, y las funciones de densidad acumulativa son denotadas por $F_{-1}(s)$ y $F_1(s)$.

- **Términos de Costo-Beneficio**: Se define b_{-1} (b_1) como el beneficio obtenido de clasificar correctamente a un *non-churner (churner)*, y c_{-1} (c_1) al costo de clasificar incorrectamente un *non-churner (churner)*. Así mismo se define $\theta = (b_1 + c_1)/(b_{-1} + c_{-1})$ como el **Ratio Costo Beneficio** para simplificar notación. Tanto el beneficio esperado de realizar la campaña, como el umbral óptimo dependerán de este ratio de costos y beneficios.

Con la ayuda de la notación recién presentada, es posible construir la siguiente matriz, conocida como la **Matriz de Confusión**:

		Clasificado como	
		Clase -1	Clase 1
Pertenece a	Clase -1	**True Negative (TN)** $[c(-1\|-1) = b_{-1}]$	**False Positive (FP)** $[c(1\|-1) = c_{-1}]$
	Clase 1	**False Negative (FN)** $[c(-1\|1) = c_1]$	**True Positive (TP)** $[c(1\|1) = b_1]$

Tabla 1: Matriz de confusión para un problema de clasificación binaria

Una medida usada frecuentemente en la literatura para evaluar el desempeño de un clasificador es el AUC, que corresponde al área bajo la curva de ROC. La curva de ROC (*Receiver Operating Characteristic*, es una representación del desempeño del clasificador en la medida que cambia el valor umbral t, que marca el límite para determinar si una observación pertenece a la clase positiva o negativa dado el score obtenido por el clasificador para esa observación. Dicho de otro modo, la curva de ROC corresponde a un gráfico que muestra la **Sensibilidad** vs $1 -$ **Especificidad**, es decir $F_{-1}(t)$ como función de $F_1(t)$, en donde:

$$\text{Sensibilidad} = F_{-1}(t),$$
$$\text{Especificidad} = 1 - F_1(t),$$
$$\text{AUC} = \int F_{-1}(s) f_1(s) \mathrm{d}s.$$

La sensibilidad indica la tasa de todos los positivos que el modelo es capaz de reconocer como positivos, mientras la especificidad es la tasa de todos los negativos que el modelo es capaz de reconocer como negativos.

En términos simples, el AUC de un método de clasificación es la probabilidad de que una observación positiva elegida de forma aleatoria sea puntuada más alta que una clasificación negativa escogida al azar [10]. Lo que implica que a mayor AUC, mejor capacidad predictiva.

Al realizar una campaña de retención, una fracción η de los clientes más propensos a la fuga es contactada, incurriendo en un costo de f por persona, para ofrecerles un incentivo con un costo monetario d para la empresa. Entre este conjunto de clientes, habrá una fracción que no tiene intenciones de fugarse. Se asume entonces que éstos aceptan el incentivo y no se fugan. Para aquellos que sí tienen intenciones de fugarse, se estima que una fracción γ acepta la oferta, lo que resulta en una ganancia directa en CLV (Ecuación (1)), mientras que una fracción $1-\gamma$ efectivamente deja la compañía a pesar del incentivo. Los clientes no contactados por la campaña de retención mantendrán su decisión, ya sea fugarse o permanecer en la empresa según corresponda. Este proceso se puede ver gráficamente en la Figura 1.

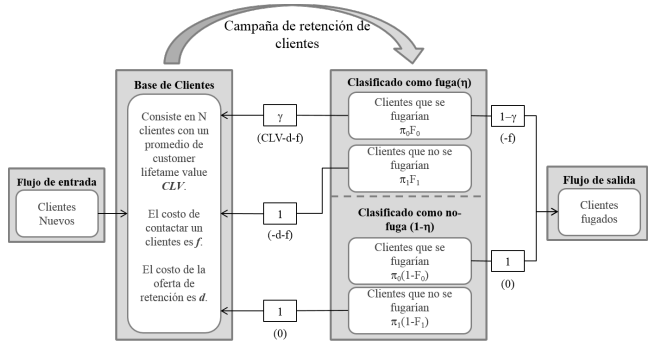


Figura 1: Esquema de evaluación de desempeño de una campaña de retención de clientes. (Fuente: [30])

El proceso descrito en la Figura 1, se sintetiza en la siguiente ecuación para el *Profit* [22]:

$$\text{Profit } = N\eta\left[(\gamma CLV + d(1-\gamma))\,\pi_{-1}\lambda - d - f\right] - A, \tag{6}$$

donde η es la fracción de clientes contactados, CLV es el *Customer Lifetime Value* (ver ecuación (1)), d es el costo del incentivo, f es el costo de contactar al cliente, y A son los costos administrativos fijos para la campaña. El coeficiente de *lift* λ es la fracción de clientes que se fugarían en la fracción η de clientes dividido por la tasa base de fuga para todos los clientes (a saber, π_{-1} [28]). Finalmente, γ se interpreta como la probabilidad de que un cliente que tenía

pensado fugarse acepte el incentivo y no se vaya de la empresa. Para todo efecto, CLV, A, f y d son positivos, y $CLV > d$ para que esto tenga sentido en primera instancia. Es importante notar que claramente η depende de la selección del valor umbral t, lo que permite a la empresa tener el control sobre qué fracción de clientes contactar para ofrecer el incentivo de retención. Al ajustar la matriz de confusión vista en la Tabla 1, se obtiene lo siguiente:

		Clasificado como	
		Clase -1	Clase 1
Pertenece a	Clase -1	$\pi_{-1}F_{-1}(t)\,N$ $[c(-1\|-1)=b_{-1}]$	$\pi_{-1}(1-F_{-1}(t))\,N$ $[c(1\|-1)=c_{-1}]$
	Clase 1	$\pi_1 F_1(t)\,N$ $[c(-1\|1)=c_1]$	$\pi_1(1-F_1(t))\,N$ $[c(1\|1)=b_1]$

Tabla 2: Matriz de confusión a nivel agregado.

Se estudiará el *Profit* promedio en vez del *Profit* total. Además se puede descartar A puesto que es un costo fijo que no depende del clasificador. Dicho esto, se define el **Profit medio generado por un clasificador para fuga de clientes** como la siguiente expresión:

$$P_{\mathcal{C}}(t;\gamma,CLV,\delta,\phi) = CLV\,(\gamma(1-\delta)-\phi)\,\pi_{-1}F_{-1}(t) - CLV(\delta+\phi)\cdot\pi_1F_1(t). \tag{7}$$

donde $\delta = \frac{d}{CLV}$ y $\phi = fCLV$. Es posible notar que $b_{-1} = CLV\,(\gamma(1-\delta)-\phi)$, y $c_1 = CLV(\delta+\phi)$. También es posible deducir las siguientes relaciones, identificando términos y viendo la estructura de la ecuación (7):

- $\eta(t) = \pi_{-1}F_{-1}(t) + \pi_1F_1(t)$
- $\lambda(t) = \frac{F_{-1}(t)}{\pi_{-1}F_{-1}(t)+\pi_1F_1(t)}$.

De lo anterior y teniendo en consideración un conjunto de entrenamiento $\mathcal{T}$ con un conjunto de atributos $\mathcal{F}$, se estudian tres métricas diferentes, descritas a continuación:

H- Measure

En Hand [15] se propone esta métrica como una alternativa al AUC. La diferencia principal entre este indicador y las MP que serán descritas posteriormente, es que la medida H sólo se enfoca en costos. El foco de esta medida

no es maximizar beneficios, si no minimizar los costos esperados. La pérdida media de clasificación Q se define como:

$$Q_{\mathcal{C}}(t; c, b) = b \cdot \left[c\pi_{-1}\left(1 - F_{-1}(t)\right) + (1 - c)\pi_1 F_1(t)\right], \tag{8}$$

donde $c = c_0/(c_{-1} + c_1)$ y $b = c_{-1} + c_1$. Calcular el valor de la pérdida mínima esperada requiere hacer supuestos sobre la densidad de probabilidad tanto de b como de c. Si se asume independencia de estos valores; y definiendo $w(b, c)$ como la distribución conjunta de b y c, y $u(c)$ y $v(b)$ como las densidades de probabilidad marginal de c y b respectivamente; se obtiene la siguiente relación: $w(b, c) = u(c)v(b)$. Usando esto último, la pérdida mínima esperada L toma la siguiente forma:

$$\mathrm{L} = \mathrm{E}[b] \int_0^1 Q_{\mathcal{C}}(T(c); b, c) \cdot u(c)\mathrm{d}c. \tag{9}$$

Se asume que c sigue una distribución Beta con parámetros α y β, cuya forma funcional es:

$$u_{\alpha,\beta}(x) = \begin{cases} \frac{x^{\alpha-1} \cdot (1-x)^{\beta-1}}{B(1,\alpha,\beta)} & \text{si } x \in [0, 1], \\ 0 & \text{en caso contrario,} \end{cases} \tag{10}$$

donde α, $\beta \in \mathbb{R}$ y α, $\beta > 1$, y además:

$$B(x, \alpha, \beta) = \int_0^x t^{\alpha-1} \cdot (1 - t)^{\beta-1}\mathrm{d}t. \tag{11}$$

Una vez definido esto, para llegar a una métrica final (*H-measure*), se normaliza lo obtenido (para obtener una métrica acotada) entre cero y uno:

$$\mathrm{H} = 1 - \frac{\int_0^1 Q_{\mathcal{C}}(T(c); b, c) \cdot u(c)\mathrm{d}c}{\pi_0 \int_0^{\pi_1} c \cdot u(c)\mathrm{d}c + \pi_1 \int_{\pi_1}^1 (1 - c) \cdot u(c)\mathrm{d}c}. \tag{12}$$

Acá $u(c)$ es una abreviación de notación de $u_{\alpha,\beta}(c)$. El denominador corresponde a la pérdida ocasionada por el peor clasificador posible, que corresponde a una función de predicción aleatoria.

Maximum Profit

Si se asume que todos los parámetros de la ecuación (7) son conocidos, para un clasificador $\mathcal{C}$ es posible generar una medida determinista. Considerando el valor máximo de $P_{\mathcal{C}}$ sobre todos los umbrales t, se obtiene la siguiente medida de rendimiento [28]:

$$\text{MPC} = \max_t P_{\mathcal{C}}(t; \gamma, CLV, \delta, \phi). \tag{13}$$

De esta manera es posible obtener la fracción de clientes $\bar{\eta}_{\text{mpc}}$ que debe ser contactada para maximizar el beneficio generado por la campaña de retención:

$$\bar{\eta}_{\text{mpc}} = \pi_{-1} F_{-1}(T) + \pi_1 F_1(T), \tag{14}$$

en donde

$$T(\gamma) = \arg\max_t P_{\mathcal{C}}(t; \gamma, CLV, \delta, \phi). \tag{15}$$

Expected Maximum Profit

Para este caso particular, se modela γ, que corresponde a la probabilidad de que un cliente que se iba de la empresa acepte el incentivo y se quede, como una variable aleatoria distribuida como una función Beta, lo que conduce a la siguiente expresión:

$$\text{EMPC} = \int_\gamma P_{\mathcal{C}}(T(\gamma); \gamma, CLV, \delta, \phi) \cdot h(\gamma) \mathrm{d}\gamma, \tag{16}$$

donde $T(\gamma)$ es el corte óptimo según la ecuación (15) y $h(\gamma)$ la densidad de probabilidad para γ. Los parámetros α y β, relacionados a la distribución Beta de γ fueron obtenidos de un trabajo previo de fuga de clientes, que puede ser revisado en detalle en [30]. De manera análoga al **MPC**, el porcentaje o fracción de clientes contactados en la campaña de retención sugerida por esta métrica es:

$$\bar{\eta}_{\text{empc}} = \int_\gamma \left[\pi_{-1} F_{-1}\left(T(\gamma)\right) + \pi_1 F_1(T\left(\gamma\right))\right] \cdot h(\gamma) \mathrm{d}\gamma. \tag{17}$$

4.2. Algoritmo basal

Se construyen nuevos métodos de selección de atributos en base al algoritmo HOSVM [19]. El concepto fundamental que está detrás de éstos es eliminar aquellos atributos cuya extracción tenga menor impacto en métrica considerada, tomando en cuenta los costos y beneficios que un clasificador puede ocasionar (dependiendo de la medida de desempeño). Para lograr esto, se busca encontrar el mejor desempeño predictivo en un conjunto desconocido del conjunto de entrenamiento, utilizando las métricas MPC, EMPC, y

H-Measure. Los métodos de selección de atributos propuestos toman los siguientes nombres: SVM_{MPC}, SVM_{EMPC}, y SVM_{H}; de acuerdo a qué métrica de desempeño es utilizada.

Dado que el problema de fuga de clientes presenta usualmente un alto desbalance de clases [2], primero se redefine el algoritmo de *Holdout SVM* para incorporar un paso donde se realizan técnicas de *resampling* para mitigar el efecto del desbalance. En este trabajo se proponen dos estrategias: *Undersampling* aleatorio, y una combinación de *Undersampling* aleatorio y *SMOTE Oversampling* para aumentar el tamaño de la clase minoritaria.

Undersampling consiste en eliminar de manera aleatoria observaciones de la clase mayoritaria, para equilibrar la distribución de clases. Análogamente al *Undersampling*, el *Oversampling* tiene como objetivo inducir un balance entre las clases del conjunto de entrenamiento generando ejemplos artificiales a partir de la clase minoritaria. Una manera práctica de realizar esta técnica consiste en crear observaciones sintéticas interpolando los ejemplos de un subconjunto de observaciones de dicha clase, lo que se conoce como ***SMOTE Oversampling*** [7].

El propósito final del algoritmo es encontrar un subconjunto $\mathcal{K}$ ($\mathcal{K} \subseteq \mathcal{F}$) de atributos, de tal manera que el desempeño predictivo del clasificador sea maximizado. Se considera un conjunto de entrenamiento $\mathcal{T}$, el cual es particionado en un subconjunto de entrenamiento $\mathcal{TR}$ y otro de validación $\mathcal{V}$. En $\mathcal{TR}$ se aplican las técnicas de *resampling* mencionadas, dando origen a un nuevo conjunto $\mathcal{TR}'$, en donde se construye el clasificador. La función de contribución de cada atributo se construye en $\mathcal{V}$ en base a las métricas MPC, EMPC, y H-Measure. Esquemáticamente, lo recién expuesto se presenta en el Algoritmo 3.

El clasificador entrenado sobre el conjunto $\mathcal{TR}'$ en el paso 5 del Algoritmo 3 corresponde al par ordenado $\Lambda = (\boldsymbol{\alpha}, b)$, y esta información se usa para calcular una función de pérdida en el conjunto de validación, a saber $\text{LOSS}^{(-j)}(\boldsymbol{\Lambda}, \mathcal{TR}', \mathcal{V})$. Se propone calcular las medidas MPC, EMPC, y H-Measure utilizando el subconjunto $\mathcal{V}$ cuando el atributo j es eliminado. El atributo cuya eliminación lleve a obtener el beneficio mayor (o menor costo en el caso de H-Measure) debe ser eliminado de la base. Para adaptar estas métricas, la versión propuesta del algoritmo solo difiere de las versiones originales de MPC, EMPC y H-Measure en el cálculo de los *scores* para cada observación, mientras que ni los costos y beneficios de una solución dada ni la definición de γ se ven afectados. Se define $s_k^{(-j)}$ como el *score* de la observa-

[2]Incluso cuando en muchos sectores de la industria de servicios se encuentran tasas de fuga anuales en torno al 20 % y 50 % [22], cuando se usan modelos de predicción de fuga mensual, las tasa de fuga usualmente se mantiene bajo el 5 % [28]

Algoritmo 3 Algoritmo de *Holdout* para *Backward Feature Elimination* usando SMOTE

Input: El conjunto original de atributos $\mathcal{F}$

Output: Un vector ordenado de atributos $\mathcal{F}^\dagger$

1. $\mathcal{F}^\dagger \leftarrow \emptyset$
2. **repetir**
3. $(\mathcal{TR}, \mathcal{V}) \leftarrow$ Holdout usando $\mathcal{T}$
4. $\mathcal{TR}' \leftarrow$ Resampling($\mathcal{TR}$)
5. $\mathbf{\Lambda} \leftarrow$ Entrenamiento SVM usando $\mathcal{TR}'$
6. $\mathcal{I} \leftarrow \text{argmin}_{\mathcal{I}} \sum_{j \in \mathcal{I}} \text{LOSS}^{(-j)}(\mathbf{\Lambda}, \mathcal{TR}', \mathcal{V}),\ \mathcal{I} \subset \mathcal{F}$
7. $\mathcal{F} \leftarrow \mathcal{F} \setminus \mathcal{I}$
8. $\mathcal{F}^\dagger \leftarrow (\mathcal{F}^\dagger, \mathcal{I})$
9. **hasta** $\mathcal{F} = \emptyset$

ción $k \in \mathcal{V}$ cuando el atributo j es eliminado, y toma la forma descrita en la ecuación (18).

$$s_k^{(-j)}(\mathbf{\Lambda}, \mathcal{TR}', \mathcal{V}) = \sum_{i \in \mathcal{TR}'} \alpha_i y_i K(\mathbf{x}_i^{(-j)}, \mathbf{x}_k^{v(-j)}) + b, \tag{18}$$

donde $\mathbf{x}_i^{(-j)}$ corresponde a una observación del conjunto de entrenamiento originado por el *resampling* cuando el atributo j es eliminado y $\mathbf{x}_k^{v(-j)}$ a un objeto de validación k con la variable j eliminada. Análogo a RFE-SVM, el vector α se asume que es igual a la solución obtenida en el paso 5 del Algoritmo 3 para reducir la complejidad computacional.

Luego de lo anterior, se proponen las siguientes métricas de desempeño basadas en utilidades para el paso 6 del algoritmo 3, donde la única diferencia respecto a la definición original de las métricas es la inclusión de la fórmula para $s^{(-j)}$:

- **H measure**:

$$\text{LOSS}^{(-j)}(\mathbf{\Lambda}, \mathcal{TR}', \mathcal{V}) = \text{H}(s^{(-j)}) \tag{19}$$

- **Maximum Profit (MPC)**:

$$\text{LOSS}^{(-j)}(\mathbf{\Lambda}, \mathcal{TR}', \mathcal{V}) = \text{MPC}(s^{(-j)}) \tag{20}$$

- **Expected Maximum Profit (EMPC)**:

$$\text{LOSS}^{(-j)}(\mathbf{\Lambda}, \mathcal{TR}', \mathcal{V}) = \text{EMPC}(s^{(-j)}) \tag{21}$$

Utilizando estas funciones en el paso 6 del Algoritmo 3, se generan tres variantes del enfoque propuesto: HOSVM$_H$, HOSVM$_{MPC}$, y HOSVM$_{EMPC}$; cuyos desempeños serán evaluados junto a las otras metodologías de selección de atributos en la siguiente sección.

Finalmente, en el paso 6 el Algoritmo 3 determina un conjunto de atributos a ser eliminado ($\mathcal{I}$). Si bien es posible eliminar un sólo elemento en cada iteración, esto es ineficiente dado que en general existe un número considerable de atributos irrelevantes. Por otro lado, remover muchos atributos a la vez aumenta el riesgo de eliminar atributos relevantes [14].

5. Resultados

En esta sección se analizan los resultados de tres problemas de predicción de fuga de clientes mediante diversas técnicas de selección de atributos. Primero se describen las bases de datos utilizadas y el diseño experimental, y luego se presentan los resultados obtenidos.

5.1. Descripción de las Bases de Datos y Diseño Experimental

Las tres bases de datos utilizadas de fuga de clientes se describen a continuación:

- **UCI-Telecom**: Esta base de datos de clientes fue extraída desde el repositorio UCI [1] y contiene la información de 5,000 clientes de una compañía de telecomunicaciones, descritos por 20 atributos.
- **Operator 1**: Esta base de clientes de telecomunicaciones fue originalmente estudiada por [21], y contiene 47,761 clientes descritos por 47 variables. Esta base de datos fue además utilizada en Verbeke et al. [28] bajo el nombre de Operator 1 (O1).
- **Cell2Cell**: Esta base de datos fue propuesta en [9] como caso de estudio, contiene información de 20,406 clientes descritos por 73 variables. Esta base de datos fue además utilizada en Verbeke et al. [28] bajo el nombre de D2.

Para los diferentes enfoques de SVM se usó LIBSVM [6] para Matlab. La tabla 3 resume la información relevante de cada base de datos:

Base de Datos	#variables	#obs(min.,may.)	tasa de churn
UCI-Telecom	20	(707;4,293)	16.5 %
Operator 1	47	(1,761;46,000)	3.8 %
Cell2Cell	73	(406;20,000)	2.0 %

Tabla 3: Número de variables, número de observaciones de cada clase y tasa de *churn* para cada una de las 3 bases.

El diseño experimental consiste en la metodología KDD [11], que es una técnica con éxito comprobado en *business analytics*, tanto en predicción de fuga como en *credit scoring* [4].

- **Recopilación y consolidación de datos:** El primer paso consiste en identificar las fuentes relevantes de datos y consolidarlas en un repositorio único creado para la construcción de los modelos de clasificación. Este paso fue directo, puesto que todas las bases de datos ya estaban consolidadas.
- **Pre-procesamiento de datos:** El siguiente paso, es la eliminación de observaciones con valores faltantes, y transformación de datos. Estos pasos también fueron directos, puesto que las bases ya estaban pre-procesadas para el análisis.
- **Minado de datos:** Se siguió el siguiente procedimiento para realizar el ranking de atributos y la selección de hiper-parámetros del modelo: Para cada una de las bases, los conjuntos de entrenamiento y testeo fueron generados mediante Validación Cruzada en 10 particiones, que es una técnica común en la predicción de fuga de clientes [28], luego se realiza tanto el ranking de atributos como la clasificación de las observaciones en el conjunto de entrenamiento. El desempeño de la clasificación final es calculado promediando los resultados de la clasificación de los diferentes conjuntos de testeo, utilizando tanto métricas usuales como indicadores de rendimiento basados en utilidades. La selección del modelo fue realizada a través de una búsqueda en grilla y evaluando los siguientes valores para los parámetros C y σ (solo para Kernel RBF): $C \in \{2^{-7}, ..., 2^{7}\}$ y $\sigma \in \{2^{-7}, ..., 2^{7}\}$.
- **Evaluación de los resultados:** El desempeño de todos los métodos fue estudiado para diferentes valores de los hiper-parámetros y comparados con diferentes indicadores, para ver cuál es el impacto en el clasificador final.

5.2. Presentación de Resultados

En esta sección se presenta un resumen de los resultados para facilitar la evaluación de la mejor instancia de cada uno de los distintos enfoques. En las tablas 4, 5, y 6 se resume el desempeño promedio entre diferentes conjuntos de atributos para cada método en las bases de datos Telecom1, Cell2Cell, y Operator1 respectivamente. Para este efecto, consideramos las siguientes métricas: AUC, EMPC, MPC y H-measure. Las métricas EMPC y MPC están medidas en Euros por cliente. El mejor desempeño

entre todos los métodos se destaca en negrita. Adicionalmente, se destaca con un asterisco cuando el desempeño es significativamente más bajo que el mejor método a un 10 % de nivel de significancia estadística, con doble asterisco a un 5 %, y con 3 asteriscos a un 1 %. Un test t es usado para hacer comparaciones entre las medias de pares de enfoques y el mejor método para cada base de datos. Los resultados son mostrados para la mejor estrategia de *resampling*, que corresponde a *undersampling* aleatorio en cada base de datos.

	Fisher	RFE	HOSVM$_{EMPC}$	HOSVM$_{H}$	HOSVM$_{MPC}$
AUC	62.4**	63.5*	64.5	64.4	**64.6**
EMPC	2.21**	2.45	2.58	2.55**	**2.61**
MPC	2.06**	2.36	2.51	2.49**	**2.55**
H	0.064**	0.089	0.092	0.085	**0.094**

Tabla 4: Desempeño medio para todos los métodos e indicadores, para la base Telecom1

	Fisher	RFE	HOSVM$_{EMPC}$	HOSVM$_{H}$	HOSVM$_{MPC}$
AUC	64.81***	94.09	94.09	94.13	**94.13**
EMPC	0.224***	**0.879**	0.860	0.860	0.859
MPC	0.223***	**0.876**	0.859	0.860	0.859
H	0.097***	**0.462**	0.429	0.381	0.385

Tabla 5: Desempeño medio para todos los métodos e indicadores, para la base Cell2Cell

	Fisher	RFE	HOSVM$_{EMPC}$	HOSVM$_{H}$	HOSVM$_{MPC}$
AUC	49.65**	54.35	54.49	**55.18**	54.47
EMPC	0.006	0.006	**0.008**	0.007	0.007
MPC	0.005	0.006	0.007	**0.007**	0.006
H	0.001***	0.001	**0.002**	0.001	0.002

Tabla 6: Desempeño medio para todos los métodos e indicadores, para la base Operator1

De la Tabla 4, para la base Telecom1, observamos que el método propuesto usando la métrica MPC para eliminación de atributos y construcción del clasificador tiene mejor desempeño para todos los indicadores considerados. El método superó a la selección vía *Fisher Score* con un nivel de significancia del 5 % en todos los indicadores y al método RFE-SVM a un nivel de 10 % de significancia en AUC. Mientras que el

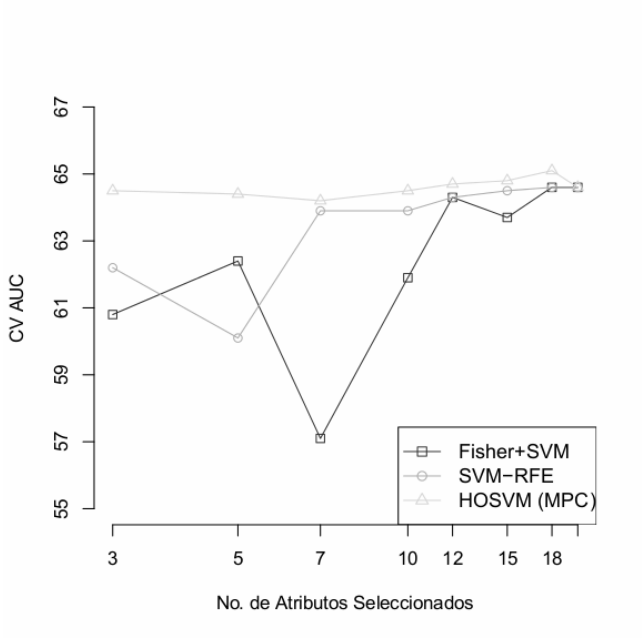


Figura 2: AUC versus el número de variables para diferentes enfoques de selección de atributos. Base Telecom1.

método HOSVM con EMPC nunca es significativamente peor que el mejor de los métodos para todas las métricas consideradas.

Para la Tabla 5, que corresponde a la base Cell2Cell, el método propuesto HOSVM con MPC tiene mejor AUC en general, mientras que RFE usando SVM tiene mejor desempeño para el resto de las métricas. Nuevamente *Fisher Score* fue superado por todos los métodos con un nivel de significancia del 1 % en todas las métricas, mientras que los otros métodos nunca se comportaron significativamente peor que el mejor método.

Adicionalmente, para la Tabla 6, que corresponde a la base Operator1, el método propuesto de HOSVM basado en EMPC tiene mejor desempeño para las métricas EMPC y H-measure, mientras que HOSVM basado en H-measure alcanza mejores resultados tanto en AUC como en MPC. Una vez más, *Fisher Score* es superado en AUC (significancia del 5 %) y en el caso de H-measure alcanza un 1 % de significancia, mientras que los otros métodos nunca están significativamente por debajo del mejor método para todas las métricas.

A modo ilustrativo, es posible analizar gráficamente el comportamiento de la selección de atributos para diferentes subconjuntos de variables. La Figura 2 presenta el desempeño predictivo para un número creciente de atributos elegidos para la base Telecom1. Para cada subconjunto de atributos, la media de AUC para los métodos *Fisher Score*, RFE-SVM y el mejor método propuesto (HOSVM basado en MPC para Telecom).

En la Figura 2 se puede observar que el método de selección de atributos propuesto

(HOSVM basado en MPC), alcanza el mejor desempeño AUC 0.648 con 18 atributos, y luego suavemente disminuye su desempeño. A diferencia de *Fisher Score* o bien RFE-SVM, que disminuyen abruptamente su desempeño en la medida que se van removiendo atributos.

6. Conclusiones

En este trabajo se propone un enfoque de eliminación de atributos recursivo, y empotrado en la construcción del modelo de clasificación usando SVM. El método propuesto estudia tres medidas de desempeño diferentes para la fuga de clientes: el *H-measure*, el *Maximum Profit Measure for Customer Churn (MPC)*, y el *Expected Maximum Profit Measure for Customer Churn (EMPC)*. Mientras que *H-measure* provee una estructura capaz de considerar de manera explícita los costos de clasificar de forma incorrecta como medida de la capacidad predictiva de un modelo, como se puede ver en [15], la medida MPC [28] y el EMPC [30] van un paso adicional e incorporan los potenciales beneficios de una campaña de retención realizada para evitar una eventual fuga de clientes, lo que genera una medida muy robusta y con una visión de negocios más recabada para evaluar el desempeño de un modelo de clasificación. La principal diferencia entre MPC y EMPC, es que este último considera la decisión de un cliente candidato a fuga como una variable aleatoria, y luego calcula el valor esperado del beneficio de una campaña de retención realizada orientada a los clientes que fueron señalados por el clasificador.

A diferencia de la literatura disponible en esta área, que se enfoca en seleccionar el mejor modelo entre varios métodos de clasificación, el objetivo de este trabajo es proporcionar un sustento teórico que permita la correcta selección de parámetros y atributos en la construcción del clasificador, basándonos en la herramienta *Support Vector Machines*. El enfoque presenta las siguientes ventajas:

- El método permite la incorporación explícita de los costos y beneficios obtenidos al clasificar en problemas de predicción de fuga de clientes, llevando a un proceso de selección de atributos especialmente diseñado para este problema en particular.
- El enfoque alcanza mejores resultados que otras técnicas de selección de atributos en problemas de predicción de fuga, considerando tanto métricas usuales de desempeño (como por ejemplo AUC), e indicadores basados en ganancias.
- La estrategia en sí misma es muy flexible y permite elegir diferentes funciones de Kernel para selección de atributos en entornos no lineales, y clasificación usando SVM. Incluso el enfoque permite ser extendido a otras herramientas de clasificación, no necesariamente SVM.

Existen muchas oportunidades de trabajo futuro, a modo de ejemplo se puede considerar las siguientes:

- El proceso de selección de atributos puede ser extendido a otras aplicaciones de *business analytics*, como por ejemplo *credit scoring* [4, 24]. El EMPC puede

adaptarse para incorporar los costos y los beneficios de aceptar o rechazar a los solicitantes de créditos, la regresión logística se puede establecer como el clasificador de referencia, puesto que es el método de clasificación más común para esta tarea debido a razones regulatorias [24].

- También es posible incorporar el costo de la adquisición de las variables en el modelo, enriqueciendo de esta manera el proceso de selección de atributos. Es posible ver este enfoque en [18], en donde el costo de los atributos se considera explícitamente en el modelo a través de variables binarias y una restricción presupuestaria explícita.

Agradecimientos: Este trabajo fue financiado por el Instituto Sistemas Complejos de Ingeniería (ICM: P-05-004-F, CONICYT: FB016) y Fondecyt (1140831).

Referencias

[1] A. Asuncion y D.J. Newman. UCI machine learning repository, 2007.

[2] B. Baesens. *Analytics in a Big Data World*. John Wiley and Sons, 2014.

[3] R.C. Blattberg, B.D. Kim, y S.A. Neslin. Database marketing: Analyzing and managing customers. 2008.

[4] C. Bravo, S. Maldonado, y R. Weber. Methodologies for granting and managing loans for micro-entrepreneurs: New developments and practical experiences. *European Journal of Operational Research*, 227(2):358–366, 2013.

[5] J. Burez y D. Van den Poel. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636, 2009.

[6] C.C. Chang y C.J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[7] N. Chawla. *Data mining for imbalanced datasets: An overview*. Springer, Berlin, 2010.

[8] P. Datta, B. Masand, D.R. Mani, y B. Li. Automated cellular modeling and prediction on a large scale. *Artificial Intelligence Review*, 14:485–502, 2000.

[9] Center for Customer Relationship Management Duke University, February 2014.

[10] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[11] U. Fayyad. Data mining and knowledge discovery- making sense out of data. *IEEE Expert-Intelligent Systems and Their Aplications*, 11:20–25, 1996.

[12] J.H. Fleming y J. Asplund. *Human Sigma: Managing The Employee-Customer Encounter*. Gallup Press, New York, 2007.

[13] I. Guyon, S. Gunn, M. Nikravesh, y L. A. Zadeh. *Feature extraction, foundations and applications*. Springer, Berlin, 2006.

[14] I. Guyon, J. Weston, S. Barnhill, y V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[15] D.J. Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123, 2009.

[16] S.Y. Hung, D.C. Yen, y H.Y. Wang. Applying data mining to telecom churn management. *Expert Systems with Applications*, 31:515–524, 2006.

[17] A. Lemmens y C. Croux. Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, 43(2):276–286, 2006.

[18] S. Maldonado, J. Pérez, M. Labbé, y R. Weber. Feature selection for support vector machines via mixed integer linear programming. *Information Sciences*, 279:163–175, 2014.

[19] S. Maldonado y R. Weber. A wrapper method for feature selection using support vector machines. *Information Sciences*, 179:2208–2217, 2009.

[20] S. Maldonado, R. Weber, y J. Basak. Kernel-penalized SVM for feature selection. *Information Sciences*, 181(1):115–128, 2011.

[21] M. Mozer, R. Wolniewicz, D. Grimes, E. Johnson, y H. Kaushansky. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, 11(3):690–696, 2000.

[22] S.A. Neslin, S. Gupta, W.A. Kamakura, J. Lu, y C.H. Mason. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, 43(2):204–211, 2006.

[23] B. Schölkopf y A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, USA., 2002.

[24] L.C. Thomas, J.N. Crook, y D.B. Edelman. *Credit Scoring and its Applications*. SIAM, 2002.

[25] D. Van den Poel y B. Lariviére. Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, 157(1):196–217, 2004.

[26] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano, y R. Wald. Feature selection with high-dimensional imbalanced data. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 507–514, 2009.

[27] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

[28] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, y B. Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211–229, 2012.

[29] W. Verbeke, D. Martens, C. Mues, y B. Baesens. Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38:2354–2364, 2011.

[30] T. Verbraken, W. Verbeke, y B. Baesens. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, 25(5):961 – 973, 2012.

[31] C.P. Wei y I.T. Chiu. Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications*, 23:103–112, 2002.

[32] X. Wu, V. Kumar, J. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. Hand, y D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14:1–37, 2008.

Modelo de simulación aplicado a procesos de atención presencial de contribuyentes en la Dirección Regional Metropolitana Santiago Oriente del Servicio de Impuestos Internos de Chile

Raúl Carpio Carrasco [*]
Juan Carlos Vilchez Pardo [**]
Patricio Duhalde Albornoz [**]

Resumen

Con la finalidad de disponer de una herramienta que ayude en la toma de decisiones, en el ámbito táctico- estratégico, que permita mejorar la eficiencia y calidad de los procesos de atención presencial de contribuyentes que el Servicio de Impuestos Internos de Chile proporciona en la Dirección Regional Metropolitana Santiago Oriente, se estudiaron los procesos de atención de público en la Plataforma de Atención y Asistencia de Contribuyentes de esta unidad, con el objetivo de desarrollar un modelo de simulación que permita, en función de las variables que determinan los objetivos estratégicos del SII, evaluar, para horizontes de mediano a largo plazo, los impactos de distintas configuraciones de atención.

Palabras Clave: Servicio de Impuestos Internos, Simulación, Diseño de Procesos.

[*]Subdirección de Contraloría Interna del Servicio de Impuestos Internos de Chile
[**]Subdirección de Fiscalización del Servicio de Impuestos Internos de Chile

1. Introdución

Durante el 2013, el Servicio de Impuestos Internos de Chile[1], en adelante indistintamente SII o el Servicio, recibió en su Plataforma de Atención y Asistencia de Contribuyentes a más de 3 Millones de usuarios que requirieron de una atención presencial a lo largo de Chile.

En la actualidad, la administración del proceso de atención se basa en un esquema en que las decisiones son tomadas conforme a los resultados obtenidos, a la cantidad de contribuyentes que pertenecen a la jurisdicción y a la operación observada y percibida in situ por los funcionarios y jefaturas del Departamento Plataforma de Atención y Asistencia.

Con la finalidad de disponer de una herramienta que ayude en la toma de decisiones en el ámbito táctico- estratégico, y que permita mejorar la eficiencia[2] y calidad[3] de los procesos de atención presencial de contribuyentes, se desarrolló un modelo de simulación aplicado específicamente al proceso de atención de la Dirección Regional Metropolitana Santiago Oriente (DRMSO), por ser ésta una de las oficinas de mayor demanda a nivel nacional.

Dentro de las aplicaciones del modelo desarrollado están las siguientes:

- Medición del impacto en el sistema de atención frente a cambios en la demanda.
- Evaluación de proyectos de mejoramiento de la atención de contribuyentes que tengan por objeto invertir en infraestructura y/o tecnología.
- Determinación de tasas de ocupación máximas que permitan que los sistemas de atención operen de manera estable, minimizando la posibilidad de colapsos ante imprevistos.
- Determinar la ubicación y cantidad de módulos de atención.
- Evaluación de nuevas políticas de atención.

[1]Servicio público que tiene a su cargo la aplicación y fiscalización de todos los impuestos internos de Chile.

[2]Mejoramiento en la calidad de la atención sin necesidad de emplear recursos adicionales.

[3]Atributos del proceso de atención que permiten que pueda ser comparado con procesos de atención similares.

2. Modelo operacional general de la plataforma de atención y asistencia de contribuyentes

La Plataforma de Atención y Asistencia de Contribuyentes tiene la función de actuar como contacto permanente en la atención presencial de los requerimientos del contribuyente [4]. Su estructura está compuesta por:

1. Un área de información y asistencia: Es donde se atienden las consultas que presentan los contribuyentes, entregando información, orientación y promoviendo la autoatención mediante el uso del portal Web del SII o bien derivando al área de atención correspondiente.

2. Un área de atención para trámites generales: Corresponde a la primera línea de atención o *front office* y es la encargada de recibir, admitir y procesar los requerimientos de los contribuyentes. Los principales trámites que se realizan en la primera línea son:

   - Obtención de RUT[5], Registro de Inicio de Actividades y Modificaciones (RIAC).
   - Timbraje de Documentos[6] y Timbraje Express [7].
   - Peticiones Administrativas[8].
   - Término de Giro[9].

3. Un área de atención para trámites específicos: Corresponde a la segunda línea de atención o *back office* de la Plataforma. Está conformada por un equipo de fiscalizadores habilitados para recibir los requerimientos que deriven las áreas de información y asistencia, atención de trámites generales o atender directamente a contribuyentes con situaciones pendientes con el SII. Estos funcionarios son los encargados de resolver los trámites

[4] Persona natural o jurídica, o administradores o tenedores de bienes ajenos afectados por impuestos.

[5] Corresponde al trámite de obtención del Rol Único Tributario.

[6] Es la autorización de los documentos tributarios para que un contribuyente pueda operar.

[7] Consiste en el timbraje de boletas o libros, los cuales son de menor complejidad y requieren de un menor tiempo de proceso.

[8] Son solicitudes y avisos de tipo general que los contribuyentes realizan ante la administración tributaria.

[9] Trámite obligatorio que los contribuyentes deben realizar cuando ponen término a su actividad.

de mayor complejidad y que requieren de mayor tiempo y competencia para su resolución.

2.1. Flujo de atención de contribuyentes

Los contribuyentes que acuden a la Plataforma de Atención y Asistencia, pueden concurrir como primera instancia al Área de Información y Asistencia para requerir información del trámite que requieren realizar[10].

Aquellos contribuyentes que requieran realizar algún trámite de RIAC, Peticiones Administrativas, Término de Giro o Timbraje de Documentos, pueden acudir directamente al Área de Atención para Trámites Generales a efectuar su diligencia.

Finalmente, son atendidos en el Área de Atención para Trámites Específicos sólo aquellos contribuyentes que son derivados desde el Área de Atención para Trámites Generales o desde el Área de Información y Asistencia, previa validación de la jefatura correspondiente.

El diagrama presentado en la Figura 1 resume el flujo de atención descrito.

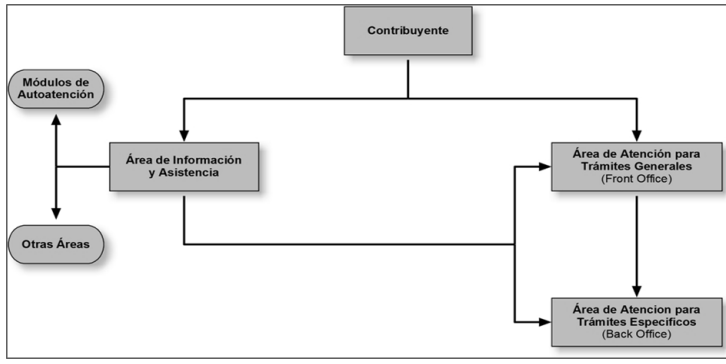


Figura 1: Flujo de atención de contribuyentes.

2.2. Sistema de control y gestión de atención de público (SCGA)

El SII, con el objetivo de ofrecer servicios que faciliten el cumplimiento tributario de los contribuyentes, implementó a mediados del año 2011, en la Plataforma de Atención y Asistencia de Contribuyentes, un "Sistema de Control y Gestión de Atención de Público (SCGA)" cuya finalidad es mejorar la administración y gestión de la atención masiva de público.

[10]Si el trámite puede ser realizado por Internet, los funcionarios orientarán y asistirán al contribuyente para concretar su trámite en un Módulo de Autoatención. Cuando el trámite no pueda ser realizado en Internet, se derivará -al contribuyente- al Área de Atención para Trámites Generales o bien al área que corresponda.

Este sistema, mediante el registro de la llegada y el tipo de trámite a realizar por el usuario, contribuyente o mandatario[11], permite generar un número de atención, el cual es asignado de forma automática a un puesto de trabajo que se encuentre activo y desocupado, para así iniciar el proceso de atención.

Por medio del SCGA es posible monitorear, registrar y almacenar la información del proceso de atención, obteniendo indicadores de gestión, tales como: el tiempo de espera y atención de los usuarios, tipos de trámites realizados y la cantidad de números emitidos y atendidos, posibilitando además el control y administración remota del sistema.

2.3. Delimitación del modelo

La DRMSO se encuentra ubicada en la comuna de Providencia en Santiago-Chile, consta de un edificio del cual los pisos 1° al 12° pertenecen al SII, con un total de 6.000 m^2 aproximadamente.

La Plataforma de Atención y Asistencia de Contribuyentes de la DRMSO se encuentra circunscrita al piso N°2, en donde se realizan trámites de RIAC, Timbraje y Timbraje Express (Trámites Generales o *Front Office*), junto con los trámites realizados por fiscalizadores (Trámites Específicos o *Back Office*); y al piso N°7 en donde se realizan solo trámites de Peticiones Administrativas. Sin embargo, para efectos de este trabajo, sólo se considerará el funcionamiento del servicio de atención proveído en el 2° piso.

3. Herramienta de simulación

Si bien la actual forma de operar del SII ha permitido mejorar los parámetros de atención, no se ha considerado en el proceso de toma de decisiones otros factores, tales como la aleatoriedad propia de los procesos de atención y la medición del nivel de impacto que podría generar una decisión en el proceso de atención, por lo que se hace necesario revisar otras formas de planificar y operar en las oficinas. Esta situación condujo a evaluar la opción de utilizar la simulación como herramienta para modelar el proceso de atención de una oficina, la que permitiría analizar y evaluar distintas configuraciones de atención y de esta manera facilitar la toma de decisiones. Con este objetivo, se desarrolló un modelo de simulación, para lo cual se utilizó el software ProModel [5].

En sistemas complejos en los cuales existen líneas de espera, y cuando por la cantidad de variables no es práctico utilizar métodos analíticos, el desarrollo

[11] Persona que recibe la confianza de otra para actuar en su nombre.

de modelos de simulación [7, 3, 1, 2] se utiliza ampliamente como herramienta en la toma de decisiones.

Para desarrollar modelos de simulación se utiliza software especializado. Se distinguen software de simulación basada en eventos discretos (Discrete-Event) y software de simulación basada en agentes (Agent-Based). Los primeros están focalizados en los procesos y los segundos en las entidades [6].

ProModel utiliza simulación basada en eventos discretos, la cual es la más comúnmente utilizada en líneas de espera.

Otro tipo de técnica utilizada en la simulación por eventos discretos, es la simulación dirigida por datos (Data - Driven), en la cual los modelos de simulación se generan automáticamente en base a datos operacionales. Las ventajas son la flexibilidad y que no se requiere demasiada expertise durante el proceso de modelamiento [4].

De la literatura consultada se extrae que en sectores industriales en los cuales la gestión de líneas de espera es un factor que determina la competitividad, la simulación es ampliamente utilizada. Por ejemplo bancos, call centers, estaciones de servicio para automóviles, servicios de urgencia médica o el movimiento de camiones en obras de construcción. En estos casos los modelos de simulación dependen de los modelos de negocios específicos de cada sector. En el caso del SII la simulación se orienta al cumplimiento de sus objetivos estratégicos asociados al cumplimiento tributario.

4. Metodología

La metodología utilizada consistió en la caracterización del proceso de llegada y atención de contribuyentes para luego crear un modelo de simulación del sistema de referencia.

4.1. Flujo de atención

El flujo de atención de un contribuyente o mandatario que ingresa al sistema es el siguiente:

- Etapa 1: Entrada al sistema del contribuyente o mandatario.
- Etapa 2: Emisión de número de atención en dispensador. En esta instancia se determina el orden de atención del contribuyente o mandatario y si la atención del trámite se llevará a cabo en un módulo de atención *front office* o *back office*.

Una atención será derivada a *back office* cuando el contribuyente mantenga situaciones pendientes con el SII.

- Etapa 3: Ingreso de contribuyente o mandatario a sala de espera *front office* si el dispensador derivó la atención a *front office* o a sala de espera *back office*, si el dispensador derivó la atención a *back office*.
- Etapa 4: Salida del contribuyente o mandatario desde sala de espera *front office* e ingreso a un módulo de atención *front office* o salida del contribuyente o mandatario desde sala de espera *back office* e ingreso a un módulo de atención *back office*.
- Etapa 5: Atención de contribuyente o mandatario en un módulo *front office* o en un módulo *back office*, según corresponda.

  Si la atención es en *front office* y se determina que el trámite es complejo, el contribuyente o mandatario es derivado a *back office*, con lo cual el flujo de atención continúa de la siguiente manera:

  - Ingreso de contribuyente o mandatario a sala de espera *back office*.
  - Salida desde sala de espera *back office*.
  - Atención de contribuyente o mandatario en un módulo *back office*.
- Etapa 6: Salida del sistema del contribuyente o mandatario y fin de atención.

Las etapas del flujo de atención determinan dos estados para el contribuyente o mandatario:

1. Estado en espera: Cuando se encuentra esperando para emitir un número de atención o está en alguna de las salas de espera.
2. Estado en atención: Cuando está emitiendo un número de atención o está siendo atendido en alguno de los módulos de atención.

4.2. Caracterización de los procesos de llegada y atención de contribuyentes

Para construir el modelo de simulación, fue necesario primero identificar los elementos del sistema y los atributos que los caracterizan.

4.2.1. Elementos del sistema

Los elementos de la Plataforma de Atención y Asistencia de Contribuyentes son:

- Entidades: contribuyentes o mandatarios que llegan a la unidad de atención a realizar trámites.
- Locaciones: módulos o puestos de trabajo destinados a la atención y los espacios destinados a espera.

En la Tabla 1 se describen cada de los elementos del sistema.

Elemento	Tipo	Descripción
Contribuyente	Entidad	Entidad que llega a la unidad de atención a realizar un trámite.
Mandatario	Entidad	Entidad que llega a la unidad de atención a realizar un trámite en representación de uno o más contribuyentes.
Dispensador de tickets de atención	Locación	Locación cuya función es determinar en forma automática, mediante la emisión de un número de atención, el orden de atención y si la atención será derivada a un módulo *front office* o a un módulo *back office*. Se estimó en base a datos reales de la DRMSO del periodo enero-diciembre de 2013, que la probabilidad que se derive una atención a *front office* es de un 97 % y que se derive a *back office* de un 3 %.
Módulo de atención *front office* (Atención para trámites generales)	Locación	Locación de primera línea en la cual un funcionario realiza la atención a contribuyentes o mandatarios.
Módulo de atención *back office* (Atención para trámites específicos)	Locación	Locación de segunda línea en la cual un funcionario fiscalizador realiza la atención a contribuyentes o mandatarios.
Sala de espera	Locación	Locación en la cual el contribuyente o mandatario espera cuando no existen módulos de atención desocupados.

Tabla 1: Elementos del sistema

4.2.2. Atributos

Cada uno de los elementos descritos en la Tabla 1 tiene asociado atributos que permiten determinar la dinámica del sistema.

4.2.2.1. Atributos de las entidades

Los atributos que caracterizan a las entidades, contribuyentes o mandatarios, son:

- El tipo de trámite a realizar y
- El tiempo entre llegadas.

El tiempo entre llegadas es el tiempo que transcurre desde que llega un contribuyente o mandatario al sistema, hasta que llega el siguiente contribuyente o mandatario al sistema. Este tiempo entre llegadas no es determinístico sino que es una variable aleatoria continua, la cual puede ser descrita mediante su función de densidad de probabilidad. En la Tabla 2 se caracteriza el proceso de llegada de contribuyentes o mandatarios.

Elemento	Trámite	Tiempo entre llegadas	p-value
Contribuyente o Mandatario	RIAC	LogNormal (1.09, 1.66) min	K-S test: 0.216 A-D test: 0.2
	Timbraje	Pearson6 (1.43, 9.76, 2.55) min	K-S test: 7.22x10-2
	Timbraje Express		A-D test: 0.238

Tabla 2: Proceso de llegada

Por otra parte, del 100 % de los contribuyentes o mandatarios que acuden a la plataforma a realizar el trámite de Timbraje o Timbraje Express, se estimó que existe una probabilidad de un 81 % que el trámite sea Timbraje y un 19 % de probabilidad que el trámite sea Timbraje Express. Se estimó además que la probabilidad de que llegue un mandatario a realizar el trámite es de un 41 % frente a un 59 % de que llegue un contribuyente. Estas probabilidades fueron calculadas a partir de datos reales de la DRMSO del periodo enero-diciembre 2013.

4.2.2.2. Atributos de las locaciones

Los atributos que las caracterizan son:

- Cantidad.
- Trámites que atiende.
- Prioridad asignada.
- Duración de las pausas de descanso, durante las cuales la locación no está disponible y,

- Tiempo de servicio o atención.

El tiempo de servicio o atención va a depender a su vez de los siguientes factores:

- Tipo de trámite a realizar.
- Si el trámite es realizado por un contribuyente o mandatario. Esta distinción se debe a que cuando el trámite es realizado por un mandatario, una misma atención incluye a todos los contribuyentes que representa.

Por otra parte se debe considerar que existen atenciones perdidas, que se generan cuando un contribuyente o mandatario abandona el sistema antes de iniciar su trámite. En estos casos, el tiempo de servicio del módulo asignado a la atención del contribuyente o mandatario que abandonó el sistema es el tiempo que utiliza el funcionario en esperar y/o realizar un rellamado. Este tiempo está configurado en 50 segundos en el SCGA, transcurrido este tiempo el sistema procede a anular el turno automáticamente.

En la Tabla 3 se muestra la probabilidad de una atención perdida, en función del tipo de trámite y del módulo asignado a la atención del contribuyente o mandatario. Dichas probabilidades se calcularon a partir de datos reales de la DRMSO del periodo enero-diciembre 2013.

Trámite/Área de Atención	*Front Office*	*Back Office*
RIAC	33 %	
Timbraje	23 %	22 %
Timbraje Express	25 %	

Tabla 3: Probabilidad de ocurrencia de atención perdida.

También existe una situación particular en la cual un contribuyente o mandatario es derivado desde un módulo de atención *front office* hacia un módulo de atención *back office*. En estos casos el tiempo de servicio del módulo *front office* es el tiempo que utiliza el funcionario en revisar los antecedentes del contribuyente o mandatario que justifican realizar la derivación, el cual fue estimado a partir de los registros mínimos y máximos.

Se estimó que la probabilidad que un trámite sea derivado desde *front office* a *back office* es un 5 %.Esta probabilidad fue calculada a partir de datos reales de la DRMSO del periodo enero-diciembre 2013.

Al igual que el tiempo entre llegadas, el tiempo de servicio de una locación es una variable aleatoria continua que se representó por su función de densidad

de probabilidad, salvo el tiempo que un contribuyente o mandatario utiliza en obtener el número de atención, el que conforme a lo observado y consultado a los usuarios del sistema se estimó en 16 segundos. En la Tabla 4 se detalla el proceso de atención.

Con respecto a la duración de la pausa de descanso, se estimó en 15 minutos por locación/jornada, tanto para los módulos *front office* como *back office*, tomando como precaución que se aplique en forma secuencial, de manera que por módulo no existan dos locaciones en pausa simultáneamente.

Las funciones de densidad de probabilidad de los tiempos entre llegadas y de los tiempos de servicio se estimaron utilizando el software STAT::FIT [8]. Para la estimación de los parámetros de las funciones de densidad de probabilidad el software utiliza el método Maximum likehood, y para testear la bondad de ajuste a los datos, los métodos: Kolmogorov–Smirnov test y Anderson–Darling test, cuyos p-value se indican en las Tablas 2 y 4). En estas mismas tablas la columna "Gráfico estimación función densidad de probabilidad" muestra en color rojo la curva ajustada de la función de densidad de probabilidad y en color azul los datos reales.

De la función de distribución de probabilidad de los tiempos entre llegadas de contribuyentes o mandatarios que realizan trámites de RIAC, se calculó mediante el software STAT::FIT, que existe una probabilidad de un 75 % que el tiempo entre llegadas sea menor o igual a 1,26 minutos. De igual forma se calculó una probabilidad de 75 % de que el tiempo entre llegadas de contribuyentes o mandatarios que realizan trámites de Timbraje o Timbraje Express sea menor o igual a 0,555 minutos. Esto implica que la frecuencia de llegada de contribuyentes a realizar trámites de Timbraje y Timbraje Express es mayor a la frecuencia de contribuyentes que llegan a la plataforma a realizar trámites de RIAC.

De las funciones de densidad de probabilidad de los tiempos de servicio, se puede inferir que cuando el trámite es timbraje realizado por un mandatario existe un 50 % de probabilidad que el tiempo de servicio varíe entre 3,23 y 8,55 minutos, con un rango intercuartil de 5,32 minutos (8,55 – 3,23). Por otra parte para los trámite atendidos en *back office* existe un 50 % de probabilidad que el tiempo de servicio varíe entre 5.63 y 20,2 minutos, con un rango intercuartil de 14,57 minutos (20,2 – 5,63). De lo anterior se deduce, considerando que el rango intercuartil es menor, que el trámite timbraje es más estructurado que los trámites atendidos en *back office*.

Elemento	#	**Tipo de trámite**	**Tiempo de servicio**		**p−value**
Dispensador de tickets	2	RIAC, Timbraje	16 s		No aplica
	1	Timbraje Express			
Módulo de atención *front office* (Atención trámites generales)	7	RIAC (prioridad)	Mandatario	*Weibull* *(1.12, 12.9) min*	K-S test: 0.109 A-D test : 9.95 x10-2
			Contribuyente	*Pearson 6* *(1.39, 7.95, 54.4) min*	K-S test: 0.119 A-D test: 4.84 x10-2
			Atención perdida	50 s	No aplica
			Derivación a *back office*	*Uniform* *(0.75, 3.11) min*	-
		Timbraje	Mandatario	*LogLogistic* *(2.26, 5.26) min*	K-S test: 6.93 x10-2 A-D test: 4.55 x10-2
			Contribuyente	*LogLogistic* *(2.2, 5.03) min*	K-S test: 8.88 x10-2 A-D test: 6.44 x10-2
			Atención perdida	50 s	No aplica
			Derivación a *back office* Mandatario	*Uniform* *Weibull* *(1.12, 12.9) min* *(0.75, 3.11) min*	K-S test: 0.109 A-D test: 9.95 x10-2
		RIAC	Contribuyente	*Pearson 6* *(1.39, 7.95, 54.4) min*	K-S test: 0.119 A-D test: 4.84 x10-2
			Atención perdida	50 s	No aplica
			Derivación a *back office*	*U(0.75, 3.11) min*	-

	9	Timbraje (prioridad)	Mandatario	*LogLogistic* *(2.26, 5.26) min*	K-S test: 6.93 x10-2 A-D test: 4.55 x10-2
			Contribuyente	*LogLogistic* *(2.2, 5.03) min*	K-S test: 8.88 x10-2 A-D test: 6.44 x10-2
			Atención perdida	50 s	No aplica
			Derivación a *back office*	*Uniform* *(0.75, 3.11) min*	-
	3	Timbraje Express	Contribuyente/ Mandatario	*Pearson 6* *(3.87, 5.09, 4.87) min*	K-S test: 0.765 A-D test: 0.611
			Atención perdida	50 s	No aplica
			Derivación a back office	*Uniform* *(0.75, 3.11) min*	-
Módulo de atención *back office* (Atención trámites específicos)	3	RIAC, Timbraje, Timbraje Express	Contribuyente/ Mandatario	*Weibull* *(1.23, 15.5) min*	K-S test: 8.22 x10-2 A-D test: 2.2 x10-2
			Atención perdida	50 s	No aplica
Sala de espera	2	RIAC, Timbraje, Timbraje Express	No aplica		No aplica

Tabla 4: Proceso de atención

4.2.3. Supuestos

Considerando que el desarrollo de este modelo de simulación busca soportar la toma de decisiones tácticas y estratégicas, como por ejemplo el diseño de oficinas o las necesidades de capacitación, la información para la estimación de las funciones de densidad de probabilidad, tanto de tiempos entre llegadas como de tiempos de servicios, consideró como base la información del mes de junio de 2013, por ser junio un mes sin contingencias y que representa el promedio de un año. Esto en consideración de que el modelo viene a solucionar las configuraciones base, sin considerar potenciales contingencias las cuales son parte de la administración de día a día de cada unidad del SII. Por la misma razón tampoco se consideraron estacionalidades.

Por configuración base se entiende la cantidad máxima de puestos de trabajo factibles de disponer en una instalación, dimensionamiento de espacios de espera, especialización de los funcionarios.

Otros supuestos son los siguientes:

- Se considera que no existen tiempos de desplazamiento desde el dispensador a las correspondientes salas de espera.
- Se considera que el tiempo de desplazamiento desde la sala de espera a los módulos de atención es de 35 segundos. Para efectos de la simulación el tiempo de desplazamiento se contabiliza como tiempo en la sala de espera.

En la Figura 2 se presenta el modelo que constituye el escenario base del análisis.

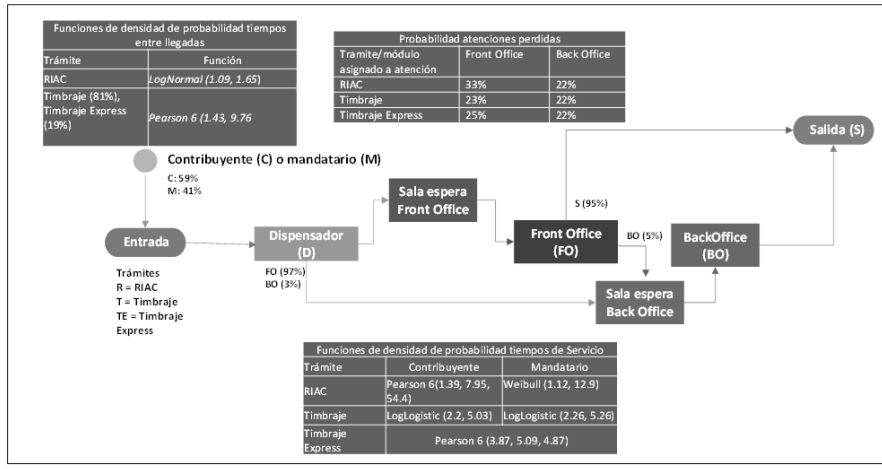


Figura 2: Modelo de simulación

4.3. Proceso de simulación

En base al modelo de probabilidades y al flujo de atención, se programó el modelo de simulación utilizando el software ProModel.

La simulación se realizó conforme a las siguientes consideraciones:

- Tiempo de simulación: Para cada réplica o corrida de simulación se simularon 7 horas. Sin embargo la entrada de entidades al sistema se restringió a las primeras 5 horas. Esto conforme a lo que sucede en la realidad donde la atención de la plataforma comienza a las 09:00 hr. y a las 14:00 hr. se cierra el acceso, de manera que no ingresen más contribuyentes o mandatarios a la plataforma. Los módulos continúan la atención sólo para los contribuyentes o mandatarios que en ese momento están en el sistema, ya sea en espera o en proceso de atención. Con lo anterior la atención en la plataforma no debería superar las 7 horas diarias.
- Cantidad de réplicas o corridas: 245, que corresponden al total de días hábiles durante año 2013, de manera tal de disponer de una muestra consistente con los objetivos del modelo que son las toma de decisiones a mediano y largo plazo.

5. Resultado y validación del modelo de simulación

5.1. Resultados del modelo

A continuación, se exponen los resultados obtenidos del proceso de simulación desarrollado, considerando el estudio de tres variables de interés:

1. Tiempo de Espera: Representa la sumatoria de los tiempos en que el contribuyente/mandatario está en estado en espera.
2. Tiempo en el Sistema: Representa el tiempo total que el contribuyente/mandatario está en el sistema, partiendo desde la emisión del número de atención hasta que abandona el sistema, considera por tanto el tiempo de espera y el tiempo de atención.
3. Tiempo de Espera *Back Office*: Representa el tiempo de espera particular del contribuyente/mandatario en la locación "Sala espera *back office*".

En la Tabla 5 se observa para cada variable de interés, el promedio ($\overline{x}$), la desviación estándar (σ) y el número de observaciones (n), para la Dirección Regional (dato real) y para la simulación.

Cabe señalar que para la variable "Tiempo de Espera", el "n" representa la cantidad de atenciones terminadas y perdidas. Por su parte para la variable "Tiempo en el Sistema", el "n" representa la cantidad de atenciones terminadas y para la variable "Tiempo de Espera *Back Office*", "n" representa la cantidad de atenciones que fueron derivadas a un módulo de atención *back office*.

VARIABLE	DRMSO (Valor Real)			Simulación		
	$\overline{x}$	σ	n	$\overline{x}$	σ	n
Tiempo de Espera (min)	26,99	24,81	233.043	29,09	28,35	243.716
Tiempo en el Sistema (min)	31,87	24,99	168.612	41,32	39,32	178.170
Tiempo de Espera *Back Office* (min)	16,90	14,59	16.880	13,00	16,66	16.750

Tabla 5: Resultados de variables de interés.

5.2. Validación del modelo

5.2.1. Validación analítica

Para hacer más exacta la validación del modelo de simulación, se realizó una interpolación a la cantidad real de atenciones, del valor promedio simulado de las variables de interés.

La interpolación indicada se realizó mediante el siguiente procedimiento:

- Con la información de las 245 réplicas de la simulación se graficó, para cada variable de interés, el valor promedio de la variable versus la cantidad diaria de atenciones.
- Utilizando la cantidad real de atenciones (n), según Tabla 5, en la Tabla 6 se calculó, dividiendo esta cantidad por 245, la cantidad promedio diario real de atenciones para cada variable estudiada.
- Por último, utilizando el gráfico, se determinó el promedio interpolado de la variable de interés, como el valor de la variable cuando la cantidad de atenciones es igual al promedio diario real según Tabla 6.

- Interpolación de la variable "Tiempo de Espera".

  La Figura 3 representa el promedio de la variable "Tiempo de Espera" versus cantidad diaria de atenciones. A partir de este gráfico se interpoló,

VARIABLE	Promedio diario real de atenciones ($\frac{n}{245}$)
Tiempo de Espera	951,20
Tiempo en el Sistema	688,21
Tiempo de Espera *Back Office*	68,90

Tabla 6: Promedio diario de atenciones

a la cantidad promedio diario real de atenciones que es 951, el tiempo promedio de espera, obteniéndose el siguiente valor:

$$Tiempo\ promedio\ de\ espera\ interpolado = 27,91\ min$$

Además se calculó para la variable el error de la simulación, medido como la diferencia porcentual del tiempo de espera promedio interpolado de acuerdo a la simulación versus el tiempo promedio de espera real, el cual es:

$$Error = 3,41\,\%$$

Figura 3: Tiempo promedio de espera.

También se observa en la Figura 3 que a medida que se incrementa la cantidad de atenciones el tiempo promedio de espera se incrementa y que cuando la cantidad de atenciones supera las 900, que es aproximadamente la cantidad promedio real de atenciones, el tiempo promedio de espera se incrementa de manera no lineal, lo cual indica que la capacidad de atención de la Dirección Regional no es suficiente para satisfacer la demanda de atenciones, motivo por el cual los funcionarios deben continuar atendiendo después del horario de cierre de la oficina.

- Interpolación de la variable "Tiempo en el Sistema".

En forma análoga se realiza la interpolación de la variable "Tiempo en el Sistema". En efecto, la Figura 4 representa el promedio de la variable "Tiempo en el Sistema" versus cantidad diaria de atenciones.

Al interpolar el valor de la variable, al valor promedio real diario de atenciones que es 688, se obtiene el siguiente resultado:

$$Tiempo\ promedio\ en\ el\ sistema\ interpolado = 39,29 min$$

Para la variable el error de la simulación, calculado como la diferencia porcentual del tiempo promedio en el sistema interpolado según la simulación versus el tiempo promedio en el sistema real, es:

$$Error = 23,29\,\%$$

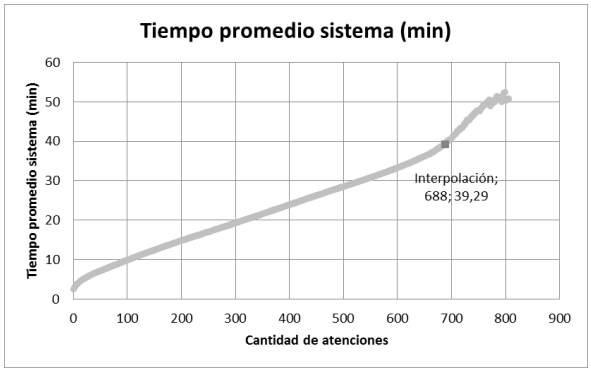


Figura 4: Tiempo promedio en el sistema.

- Interpolación de la variable "Tiempo Espera *Back Office*".

  Al interpolar el valor de la variable, al valor promedio real diario de atenciones que es 69, se obtiene el siguiente resultado:

$$Tiempo\ promedio\ espera\ back\ office\ interpolado = 16,80 min$$

El error de la simulación para la variable, calculado como la diferencia porcentual del tiempo promedio de espera *back office* interpolado según la simulación versus el tiempo promedio de espera *back office* real, es:

$$Error = 0,61\,\%$$

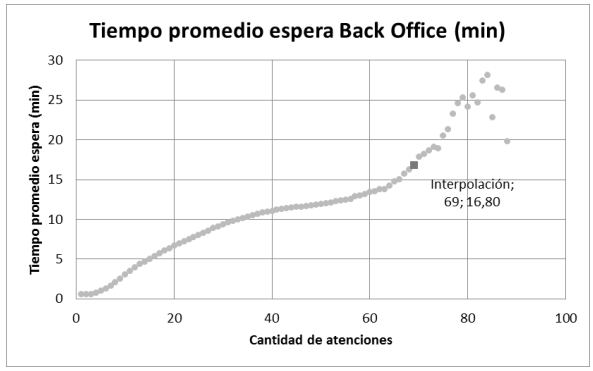


Figura 5: Tiempo promedio espera back office.

Por lo tanto, es posible concluir que el modelo de simulación constituye una buena aproximación a la realidad para los tiempos de espera, que permite evaluar y analizar potenciales escenarios de configuración de locaciones. El error en la simulación de tiempo en el sistema se puede explicar por los órdenes de magnitud de 10^{-2}, de los p-value obtenidos en la estimación de las funciones de densidad de probabilidad de los tiempos de servicio de timbraje, el cual es el trámite con mayor demanda. La Tabla 7 resume la validación del modelo.

VARIABLE	Valor Real ($\overline{x}$)	Valor Simulación (interpolado)($\overline{x}$)	Error
Tiempo de Espera (min)	26,99	27,91	3,41 %
Tiempo en el Sistema (min)	31,87	39,29	23,29 %
Tiempo de Espera *Back Office* (min)	16,90	16,80	0,61 %

Tabla 7: Resultados validación modelo.

5.2.2. Validación según variables y parámetros de negocio

Como se señaló anteriormente, el modelo provee una buena aproximación en dos de las tres variables de interés estudiadas, Sin embargo la variable "tiempo en el sistema" presenta un error de un 23,29 %. Esta diferencia se podría explicar por las siguientes variables de negocio:

- **Tipo de atenciones:** Los trámites se pueden clasificar según su estructuración, existiendo trámites más estructurados que otros. Los trámites de mayor estructuración son los trámites de *front office*, donde los trámites de timbraje son más estructurados que los trámites de RIAC. Por

otra parte los trámites menos estructurados son los trámites que se realizan en el *back office*, que son revisiones específicas efectuadas por un fiscalizador, y en que la gama de elementos a revisar es amplia. Lo anterior trae como consecuencia una alta variabilidad en los tiempos de servicio en estas locaciones, afectando las estimaciones asociadas.

- **Cantidad de atenciones:** El modelo consideró para la estimación de los tiempos entre llegadas un mes promedio, como lo fue junio del 2013, Sin embargo existen meses con situaciones especiales, como abril, que es el mes de la "Operación Renta", el cual se caracteriza porque llega una mayor cantidad de contribuyentes a la oficinas del SII, sobre todo, a efectuar consultas.

- **Parámetros anualizados:**Se utilizaron parámetros en el modelo anualizados, como por ejemplo, relación mandatarios/contribuyentes, porcentaje de números perdidos y porcentaje de trámites en "*back office*". Esto podría generar ciertas diferencias en el modelo al combinar estos parámetros con los tiempos de llegada y de atención de un mes "promedio", explicados en el punto anterior.

6. Evaluación de escenarios

A continuación, luego de validar el modelo de simulación, se procedió a generar distintos escenarios de mejoramiento del proceso de atención, de manera de evaluarlos utilizando simulación.

La generación de los escenarios de simulación se enfocó en el aumento de la cantidad de locaciones que dan prioridad a la atención de Timbraje, considerando que la frecuencia en la llegada de contribuyentes/mandatarios que vienen a realizar trámites de Timbraje/Timbraje Express es mayor a la frecuencia en la llegada de contribuyentes/mandatarios que vienen a realizar trámites de RIAC.

La configuración de los escenarios consideró como restricción el espacio físico disponible para atención de contribuyentes con lo cual la cantidad total de locaciones se mantuvo constante.

De acuerdo a lo anterior, en la Tabla 8 se establecen las distintas configuraciones para evaluación, partiendo del escenario de control o base.

Con estas configuraciones se procedió a ejecutar nuevamente el modelo de simulación.

Escenario	**LOCACIONES**				**Total**
	Locaciones *Front Office*			**Locaciones *Back Office***	
	Locaciones Timbraje /RIAC (prioridad Timbraje)	**Locaciones Timbraje /RIAC (prioridad RIAC)**	**Locaciones Timbraje Express**		
Base	9	7	3	3	22
1	10	7	2	3	22
2	11	5	3	3	22
3	11	6	2	3	22

Tabla 8: Configuración de escenarios

Para potenciar la evaluación de escenarios se agregaron al análisis las siguientes variables:

1. Tiempo total atención: Tiempo total utilizado para completar la atención, con tope de 7 horas diarias que es el tiempo de simulación.

2. Cantidad de atenciones: Cantidad total de atenciones en el horizonte de simulación de 245 días.

3. Porcentaje ocupación locaciones: Tiempo efectivo en que una locación es utilizada respecto al tiempo disponible para atención.

En las Tablas 9 y 10, se observan las variables que se utilizarán para la evaluación de cada uno de los escenarios.

VARIABLE	Escenario			
	Base	1	2	3
Tiempo promedio de espera (min)	29,09	21,65	26,01	21,49
Tiempo promedio en el sistema (min)	41,32	33,71	37,06	33,10
Tiempo promedio de espera *back office* (min)	13,00	15,09	14,34	15,38
Tiempo total atención (hora)	1.611	1.564	1.610	1.579
Cantidad de atenciones	243.716	244.085	242.833	243.912

Tabla 9: Desempeño del sistema para los distintos escenarios.

LOCACIÓN	Escenario			
	Base	1	2	3
Front Office	74,24 %	76,49 %	74,26 %	76,43 %
Back Office	66,05 %	69,12 %	66,50 %	67,96 %
% Promedio ponderado de ocupación	73,13 %	75,48 %	73,20 %	75,28 %

Tabla 10: Porcentaje ocupación de locaciones.

Como se aprecia en las Tablas 9 y 10, no existe ninguna configuración que tenga el mejor resultado para todos los conceptos medidos. De esta forma, la configuración a seleccionar dependerá del criterio que se plantee.

Para efectuar un análisis que permita establecer cuál es la mejor alternativa de configuración, se establecieron cuatro medidas de desempeño de la eficiencia y calidad del proceso de atención, las cuales son:

- **Servicio:** Se asocia a la variable "Tiempo promedio de espera".
- **Cobertura:** Se asocia a la variable "Cantidad de atenciones".
- **Costo:** Se asocia a la variable "Tiempo de atención".
- **Riesgo de operación:** Se asocia a la variable "Porcentaje promedio ponderado de ocupación". Se considera que a medida que se incrementa el nivel de ocupación de las locaciones se incrementa el riesgo que el sistema se vea afectado, en los tiempos de espera, ante un aumento marginal en la demanda de atenciones.

La Tabla 11 resume para cada escenarios de evaluación los valores de cada una de las medidas de desempeño definidas y las variable asociadas.

VARIABLE	Medida de Desempeño	Escenario			
		Base	1	2	3
Tiempo promedio espera (min)	Servicio	29,09	21,65	26,01	21,49
Cantidad de atenciones	Cobertura	243.716	244.085	242.833	243.912
Tiempo de atención (horas)	Costo	1.611	1.564	1.610	1.579
% Promedio ponderado ocupación	Riesgo	73,13	75,48	73,20	75,28

Tabla 11: Porcentaje ocupación de locaciones.

En la Tabla 12, se visualiza el valor correspondiente al mejor desempeño y al escenario y concepto asociado. Como se observa, sólo el escenario N°2 no cuenta con el mejor valor en alguna de las medidas de desempeño analizadas.

También se observa que el escenario N°1 cuenta con el mejor valor en dos de las cuatro medidas de desempeño. Mediante una simple inspección se podría concluir que el mejor escenario sería el N°1. No obstante, esto va a depender de la ponderación que tengan las medidas de desempeño en la estrategia que se defina y de la magnitud de las diferencias entre cada medida de desempeño.

Escenario	**MEDIDAS DE DESEMPEÑO**			
	Servicio	**Cobertura**	**Costo**	**Riesgo**
Variable asociada	**Tiempo promedio espera (min)**	**Cantidad de atenciones**	**Tiempo de atención (horas)**	**%promedio ponderado ocupación**
Valor asociado	21,49	244.085	1.564	73,13
Base				X
1		X	X	
2				
3	X			

Tabla 12: Mejores desempeños y escenario asociado

Para este caso en particular, la estrategia a seguir estará enfocada en "Servicio" y en "Cobertura".

De esta forma, se definieron los siguientes factores de ponderación para la función objetivo:

- Servicio = 40 %
- Cobertura = 40 %
- Costo = 10 %
- Riesgo de operación =10 %

Para identificar que configuración de locaciones es mejor, dado los factores antes indicados, se normalizará para cada alternativa de configuración los valores de las medidas de desempeños, de manera tal que den cuenta de un puntaje estándar único.

Luego, escalando entre [0-100] para cada medida de desempeño, es decir, asignando el valor 100 a la mejor alternativa, se tiene:

Por lo tanto, la mejor alternativa para estos criterios, corresponde al escenario N° 3. Lo anterior, dado que es el escenario que presenta el mayor puntaje.

VARIABLE	Medida de Desempeño	Escenario (valor normalizado)			
		Base	**1**	**2**	**3**
Tiempo promedio espera	**Servicio**	73,87	99,26	82,62	100
Cantidad de atenciones	**Cobertura**	99,85	100	99,49	99,93
Tiempo de atención	**Costo**	97,08	100	97,14	99,05
% promedio ponderado ocupación	**Riesgo**	100	96,89	99,90	97,14
Puntaje Ponderado Final		**91,81**	**99,16**	**94,28**	**99,31**

Tabla 13: Comparación de resultados normalizados por medida de desempeño.

No obstante, se podría decir que entre la configuración N° 1 y la N°3, prácticamente no existirían diferencias significativas, por lo que la decisión podría moverse entre estos dos escenarios.

También se observa que todas las configuraciones analizadas, mejoran la situación actual, lo que representa una real oportunidad para generar un incremento en la performance del proceso de atención.

7. Conclusiones

- Con la utilización de técnicas y herramientas de simulación en el modelamiento del proceso de atención de contribuyentes, se es factible generar un modelo, que puede predecir el comportamiento del sistema en un periodo de tiempo.
- El modelo de simulación desarrollado permite analizar y evaluar el comportamiento de variables y medidas de desempeño ante distintos escenarios de atención, de manera que dada una estrategia de atención se pueda seleccionar la configuración más conveniente.
- La utilización de herramientas de simulación en la evaluación de escenarios, reduce los riesgos asociados con la implementación de nuevas configuraciones de atención. Mediante esta metodología, se pueden conocer de manera previa estimaciones de las medidas de desempeño.
- Uno de los supuestos importantes utilizados en el desarrollo del modelo fue la utilización de información de un mes con operaciones “promedio”,

como es el mes de junio, en el cual la afluencia de público no se ve afectada por situaciones extraordinarias, como lo es la operación renta en el mes de abril. Esto debido a que el modelo busca ser de utilidad en la simulación de configuraciones base asociadas a decisiones de tipo táctico y estratégico, en las cuales no se consideran las contingencias del día a día.

- De la evaluación de escenarios realizada, se observa que el proceso actual de atención es perfectible, al encontrarse una brecha de mejoramiento en el desempeño. Esto, sin necesidad de aumentar la cantidad de locaciones y por ende efectuar inversiones que impliquen un incremento de costos.

- Al observar los porcentajes de ocupación que se obtienen de la simulación, se infiere que se está ante un escenario de riesgo. Esto se traduce en que, ante un evento inesperado, (por ejemplo, la falta de personal en algún momento de la jornada) se apreciaría un aumento considerable en los tiempos de espera para los contribuyentes, afectando de esta manera la calidad del servicio. Luego, es importante evaluar la incorporación de nuevas locaciones de atención que permitan contar con holguras.

- Para poner en práctica el modelo de simulación en la plataforma de atención de la DRMSO es recomendable comenzar con la implementación de configuraciones que no requieran incurrir en inversiones o costos adicionales, tales como las que se evaluaron en este documento, es decir, sobre la base de la infraestructura actual. De esta manera es simple volver atrás si es necesario.

- En la implementación del modelo se deberá tener en cuenta variables de negocio, tales como los tipos de atención en relación al grado de estructuración que tengan e incrementos en la cantidad de atenciones producto de situaciones puntuales asociadas a las obligaciones de los contribuyentes.

- Por otra parte, el modelo tiene gran potencial en la evaluación de la creación de nuevas Direcciones Regionales, cambios de edificios en las oficinas actuales, fusión o división de Direcciones Regionales, aumento o disminución de dotación en determinados trámites, entre otros. También se visualiza potencial de la herramienta de simulación en la definición y evaluación de factibilidad de metas institucionales relacionadas con la atención de contribuyentes, como por ejemplo tiempos máximos de espera.

Referencias

[1] R. Akhavian y A.H. Behzadan. Evaluation of queuing systems for knowledge-based simulation of construction processes. *Automation in Construction*, 47:37–49, 2014.

[2] L. Birta y G. Arbez. Modelling and simulation: Exploring dynamic system behaviour. *Springer-Verlag London*, 2007.

[3] N. Boccara. Modeling complex systems. *Springer-Verlag New York*, 2010.

[4] C. Meng, S. Nageshwaraniyer, A. Maghsoudi, Y. Son, y S. Dessureault. Data-driven modeling and simulation framework for material handling systems in coal mines. *Computers & Industrial Engineering*, 64 (3):766–779, 2013.

[5] ProModel. User guide, 2010.

[6] G.A. Pugh. Agent-based simulation of discrete-event systems. *American Society for Engineering Education*, 2006.

[7] S.M. Ross. *Simulation*. Academic Press, 2012.

[8] STAT::FIT. User guide.

Programas de Postgrado y Postítulos DII

DOCTORADO

MAGÍSTERES

Programas de Postgrado y Postítulos DII

Programas de Postgrado y Postítulos DII

Programas de Postgrado y Postítulos DII

EDUCACIÓN EJECUTIVA