R E V I S T A

INGENIERÍA DE SISTEMAS

Volumen XXVII Septiembre 2013

Una licitación combinatorial aplicada a la provisión de Internet a las escuelas de Buenos Aires 9
Flavia Bonomo, Jaime Catalán, Guillermo Durán, Rafael Epstein, Alexis Jawtuschenko, Javier Marenco

Una aplicación de Web Opinion Mining para la extracción de tendencias y tópicos de relevancia a partir de las opiniones consignadas en blogs y sitios de noticias 31
Rodrigo Dueñas F., Juan D. Velásquez

Planificación del menú semanal de colaciones de un hospital de Argentina por medio de programación lineal entera 55
Sebastian Guala, Javier Marenco

Aplicación de Minería de Datos para predecir fuga de clientes en la industria de las telecomunicaciones 73
Francisco Barrientos, Sebastián A. Ríos

Programación Matemática para asesorar a un entrenador de fútbol: un juego de fantasía como caso de estudio 109
Flavia Bonomo, Guillermo Durán, Javier Marenco

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

R E V I S T A
INGENIERÍA DE SISTEMAS
ISSN 0716 - 1174

Financiado parcialmente por el Instituto Sistemas Complejos de Ingeniería.

Representante legal: Alejandra Mizala	Dirección: República 701, Santiago, Chile.
Correo electrónico: ris@dii.uchile.cl	Web URL: www.dii.uchile.cl/∼ris
Impresión: Ka2 Diseño e Impresión	Mail: contacto@ka2.cl
Diagramación: Cinthya Vergara	Portada: Gabriella Fabbri

Carta Editorial Volumen XXVII

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, de investigadores del ISCI y de académicos de las Universidades de Buenos Aires y de General Sarmiento, en Argentina.

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 la región, debido a sus características similares.

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, y su extensión no debe exceder las 30 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

Una licitación combinatorial aplicada a la provisión de Internet a las escuelas de Buenos Aires

F. Bonomo [*]
J. Catalán [***]
G. Durán [**]
R. Epstein [***]
A. Jawtuschenko [****]
J. Marenco [*****]

Resumen

Una de las clases de licitación más estudiadas en la literatura es la multi-unidades, aquella en la que se licitan varios ítems idénticos. En este trabajo, definimos una nueva subclase de las licitaciones multi-unidades, que llamamos *multi-unidades logística*. La característica central de una licitación multi-unidades logística es

[*]Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina. CONICET, Argentina..

[**]Instituto de Cálculo y Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina. CONICET, Argentina. Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile, Santiago, Chile.

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

[****]Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina. CONICET, Argentina.

[*****]Instituto de Ciencias, Universidad Nacional de General Sarmiento, Argentina. Departamento de Computación, FCEN, UBA, Argentina

que existen fuertes asimetrías de costos entre los oferentes a causa de consideraciones logísticas. Ciertas unidades pueden ser atractivas para una firma y no para otra, de acuerdo a dichas consideraciones, y lo contrario puede suceder en otras unidades a licitar. Este tipo de licitación aparece en la provisión de Internet en las escuelas de Buenos Aires, la capital de Argentina. En 2008, la ciudad de Buenos Aires debía licitar la conexión de Internet para sus 709 escuelas públicas. En este trabajo mostramos cómo fue diseñada la licitación para realizar la adjudicación minimizando el gasto de la ciudad destinado a este fin. En nuestro diseño, cada firma debe dar un precio general por el servicio mensual a ser brindado en cada escuela, descuentos por volumen en bandas prefijadas de antemano y el listado de escuelas en las que le interesa brindar el servicio. La licitación multi-unidades logística resultante se puede interpretar como una licitación combinatorial.Se implementó un modelo de programación lineal entera para obtener el conjunto de ofertas más conveniente para el estado.De acuerdo con los análisis presentados en este trabajo, se puede estimar que la ciudad de Buenos Aires obtuvo ahorros cercanos al 20 % por implementar este nuevo modelo de licitación en lugar del formato de licitación originalmente propuesto.

Palabras Clave: Licitaciones Combinatoriales, Licitaciones Multi-unidades, Programación Lineal Entera.

1. Introducción

Los procesos de licitación permiten vender y comprar bienes y servicios en forma eficiente, aumentando el bienestar de todos los actores involucrados. Para ello, es crucial entender las preferencias de los actores y la naturaleza de los productos que se compran y venden. La aparición de Internet y del comercio electrónico han abierto un inmenso campo de aplicación para estudiar y mejorar los mecanismos de licitación, donde la Investigacion de Operaciones está jugando un rol central.

En este trabajo proponemos una nueva subclase de licitaciones que se deriva de la clásica subasta conocida como *multi-unidades* (se puede consultar un resumen del estado del arte de licitaciones multi-unidades

en [6]). En la literatura se han propuesto diversos mecanismos para licitar estas unidades idénticas de modo de asignar los artículos a quienes más los valoran maximizando de este modo la recaudación del vendedor, en lo que constituye un mecanismo óptimo. En muchas licitaciones se compran y venden múltiples productos que están compuestos de una compleja mezcla de bienes y servicios, donde la logística juega un rol central. Efectivamente, la componente material de estos productos a licitar entre sí es idéntica, lo que define una licitación multi-unidades, y por lo tanto el precio final debería ser parecido para todos ellos. Sin embargo, la componente logística por lo general cambia esa apreciación porque las habilidades y especialidades de las empresas que proveen estos productos son heterogéneas, lo que provoca que algunas empresas tengan ventajas comparativas para proveer algunos de los productos a licitar, mientras que otras empresas son más competitivas en otros casos. Estas licitaciones las hemos clasificado en este trabajo como "licitaciónes multi-unidades logísticas", y suelen resolverse a través de la formulación de una licitación combinatorial.

Una *licitación combinatorial* (combinatorial auction) es una subasta en la cual los oferentes pueden armar un paquete de ítems y presentan un precio por el conjunto, que se acepta o rechaza en su totalidad. Por lo tanto, el valor de cada ítem es relativo al conjunto en el que está inserto. Esta característica define la propiedad combinatorial de la licitación. Este tipo de licitaciones tiene múltiples aristas que incluyen el diseño de la licitación, el desarrollo de modelos matemáticos que permitan determinar el mejor conjunto de ofertas para el organizador de la licitación y la implementación de algoritmos que permitan resolver estos modelos. El organizador de la licitación busca minimizar costos si es un "comprador", o maximizar su beneficio si es un "vendedor". Dada la interdisciplinariedad de los problemas de licitaciones combinatoriales, conviven en su formulación y resolución economistas, expertos en gestión, especialistas en investigación de operaciones y teoría de juegos, y profesionales de las ciencias de la computación. Para más detalle sobre licitaciones combinatoriales ver [1].

En este paper presentamos el caso de la licitación de servicios de Internet para los colegios públicos de la ciudad de Buenos Aires, la capital de la Argentina. Mostramos que esta licitación también clasifica como multi-unidades logística y diseñamos un mecanismo de carácter combi-

natorial para su resolucion. En 2008 el gobierno de la ciudad de Buenos Aires lanzó una licitación para instalar servicios de Internet en los 709 establecimientos públicos escolares de la ciudad, por un lapso de dos años. El proyecto preliminar de la licitación involucraba plantear una licitación multi-unidades tradicional. En la propuesta original del gobierno de la ciudad cada empresa debía realizar una oferta individual por cada escuela, pudiendo no ofertar el servicio en algunas de ellas. El valor de la oferta debí a corresponder al abono mensual por la provisión de Internet, servicio que estaba sujeto a condiciones técnicas estipuladas de antemano. En cada escuela ganaría la empresa que hiciera la mejor oferta.

Este diseño para la licitación exhibía algunos problemas. En primer lugar, no se daba a las empresas la posibilidad de realizar descuentos por volumen, con la consecuente suba del precio promedio del abono mensual. Por otra parte, la provisión de Internet está fuertemente restringida por la tecnología preinstalada, con lo cual la distribución geográfica de las empresas es un factor limitante a la hora de presentarse a la licitación. Por este motivo, con este diseño de licitación era posible que los precios tendieran a ser altos en las zonas de la ciudad con menos competencia (zonas con pocas empresas que ya tuvieran la tecnología instalada), e incluso habría una gran posibilidad de colusión en dichas zonas de baja competencia.

Este trabajo presenta un nuevo diseño que se propuso y se aplicó en esta licitación, con los objetivos de aprovechar descuentos por volumen por parte de las empresas proveedoras –intentando así disminuir la erogación total por parte del gobierno de la ciudad– y dificultar las posibilidades de colusión entre los oferentes. El diseño para la licitación contempla que no todas las empresas participantes pueden proveer el servicio en todas las escuelas, incluyendo también la posibilidad de proporcionar descuentos por volumen. Cada escuela no se considera como una entidad separada, sino que el precio individual que la ciudad abona por ella a una empresa depende del conjunto de escuelas asignado a la empresa. Por estos motivos, la licitación propuesta se puede interpretar como una licitación combinatorial con restricciones logísticas, aunque manteniendo características propias de un proceso multi-unidades. No estamos al tanto de trabajos previos en los que se haya presentado este diseño de licitación.

Existen diversos ejemplos de licitaciones combinatoriales realizadas desde ámbitos públicos, con el objetivo de optimizar el uso del presupuesto, aunque en estos trabajos previos no se aplicó el formato de subasta presentado por primera vez en el presente trabajo. Un ejemplo paradigmático es la licitación de alimentos que realiza el estado de Chile, a través de su agencia JUNAEB, por casi mil millones de dólares por año, para proveer alimentación a los colegios públicos. Este procedimiento se enmarca en esta nueva clasificación de licitación multi-unidades logística. En este caso se entregan dos millones de almuerzo por dia a niños que están estudiando en 5,000 colegios y el sistema opera 200 dias al año. El producto es un almuerzo que se compone de alimentos, como arroz o pollo, que son bienes muy homogéneos cuya calidad está especificada en detalle y se asemejan a un commodity. Estos alimentos deben ser transportados, almacenados, cocinados y servidos a los niños, lo que significa una operación logística de envergadura. En este caso, la componente espacial juega un rol clave que afecta la logística y por ello la calidad del producto. Es distinto operar en Santiago, una gran ciudad de mas de seis millones de habitantes, que operar en una región rural. Las empresas se han especializado y son eficientes para ciertas condiciones pero no para otras. Para optimizar este proceso de compra se diseñó una licitación combinatorial que ha operado exitosamente desde 1997 [2], donde el país se dividió en Unidades Territoriales que constituyen los objetos a licitar y los oferentes las pueden agrupar en paquetes que se aceptan o rechazan en su conjunto. Esta característica da origen a la naturaleza combinatorial de la licitación y al mismo tiempo permite que cada empresa refleje sus costos para cada Unidad Territorial en forma muy eficiente.

Otro ejemplo de uso concreto de una licitación combinatorial en el contexto de la procuración pública está dado por el mercado de omnibus de Londres [3, 5]. Este mercado comprende alrededor de 800 rutas a través de un área de 1630 kilómetros cuadrados, usadas por más de 3,5 millones de pasajeros por día. Antes de la etapa de desregulación los servicios de ómnibus en el Gran Londres eran provistos por una empresa estatal londinense. En 1984 el servicio fue privatizado, y desde entonces es otorgado a empresas de transporte a través de licitaciones combinatoriales anuales. La concreción de las licitaciones se fue haciendo de manera gradual. La primera de ellas fue realizada en 1985, pero recién en 1995 la mitad de las rutas habían sido licitadas al menos una vez. Hoy, el sis-

tema ha llegado a una situación de estabilidad, licitándose alrededor de un 20 % de la red año a año. La licitación de rutas de ómnibus de Londres es considerada un éxito, pues devino en un incremento de la calidad del servicio y en una disminución de los costos para el gobierno de Londres. En este caso nuevamente podemos separar el producto bajo licitación en dos componentes principales: primero están los buses, que son bienes homogéneos, y luego la logística que se necesita para operar estos buses, que incluye abastecerlos con combustible y aceite, realizar los cambios de aceite y otras mantenciones, corregir las fallas, y contratar, entrenar y capacitar a una dotación de conductores profesionales. Las capacidades logísticas de las empresas son diferentes, mientras algunas tendrán ventaja para operar en unos recorridos, otras serán más eficientes en otros. Estas diferencias aparecen por instalaciones físicas que están en lugares específicos de la ciudad, localización de los choferes, conocimiento y habilidad para operar rutas de gran demanda o de baja demanda, entre otros aspectos.

El presente trabajo está organizado del siguiente modo. En la Sección 2 se aborda el diseño de la licitación, mencionando distintas posibilidades para procesos licitatorios de tipo multi-unidades con descuentos por volumen y restricciones dadas por consideraciones logísticas. La Sección 3 contiene dos modelos matemáticos para resolver el problema de adjudicar las escuelas a los oferentes minimizando el costo total, y la Sección 4 describe la experiencia de aplicación de este diseño a la licitación del servicio de Internet a las escuelas públicas de la ciudad de Buenos Aires. Finalmente, la Sección 5 presenta las conclusiones del trabajo.

2. Diseño de la licitación

Se describe en esta sección el proceso de diseño de la licitación realizado por los autores. Se presentan distintos diseños alternativos y se discuten las características de cada uno, arribando al diseño finalmente implementado. El objetivo principal de esta etapa del trabajo es obtener un diseño que genere competencia entre las empresas a través de descuentos por volumen (cuantas más escuelas se asignan a la empresa, el precio por unidad debe ser menor, buscando que la ciudad realice el menor desembolso total posible), manteniendo un proceso transparente y que no

esté sesgado en favor o en contra de ninguna empresa particular. Como se mencionó en la introducción, el principal factor a tener en cuenta con relación a este último punto es que la provisión del servicio de Internet en una escuela depende de la tecnología instalada, y esto define el carácter logístico de la licitación. Para una empresa sin instalaciones en la zona es muy costoso proveer este servicio.

2.1. Licitación basada en unidades territoriales

Desde el punto de vista administrativo de su sistema educativo, la ciudad de Buenos Aires está dividida en 21 *distritos escolares*, como puede verse en la Figura 1. Para incluir dentro del diseño la posibilidad de que los oferentes realizaran descuentos por volumen, se considera primero la posibilidad de diseñar una licitación como la desarrollada en el caso de los comedores escolares de Chile [2], utilizando los distritos escolares como *unidades territoriales* para la licitación. Las empresas podrían entonces ofertar por combinaciones de distritos escolares, adjudicando a posteriori del modelo matemático la partición de la ciudad que fuera la más conveniente en costos para el estado. Cada unidad territorial se asignaría en forma completa a una empresa (seleccionada por el modelo matemático), de modo que cada empresa ganadora debería instalar el servicio en todas las escuelas de las unidades territoriales obtenidas.

Figura 1: Partición de la ciudad de Buenos Aires en distritos escolares.

Esta propuesta tenía la ventaja de considerar descuentos por volu-

men y la intención de la ciudad de obtener un menor precio total basado en hacer competir entre sí a las empresas, pero seguía adoleciendo de un problema importante: no estaba teniendo en cuenta las tecnologías ya instaladas en la ciudad por parte de las potenciales empresas participantes. Con esta partición de la licitación en distritos escolares, podía suceder que una empresa que tenía ya su tecnología en sólo una parte del distrito subiera fuertemente el precio en todo el distrito para compensar el tener que llegar a zonas donde todavía no había accedido. Las empresas en estas condiciones podían objetar el proceso licitatorio, aduciendo –con razón– que se veían desfavorecidas por el formato de la licitación.

Por lo tanto, se desechó esta posibilidad y se procedió a analizar el radio de acción donde ya tenían tecnología instalada cada una de las empresas que se suponía podían participar de la licitación. Se generó entonces una propuesta que diseñaba nuevas unidades territoriales en base a cruzar los distritos escolares con los radios de acción de las potenciales empresas participantes.

Esta nueva propuesta aglutinaba los distritos escolares en 11 unidades territoriales, teniendo en cuenta los radios de acción de las empresas (ver Figura 2). De todas maneras, la propuesta seguía presentando algunas deficiencias: por un lado, el radio de acción de cada empresa estaba siendo establecido por el gobierno de la ciudad, lo que podía dar lugar a errores en la definición del mismo, y por otra parte, seguía habiendo unidades territoriales con posiblemente muy poca competencia, lo que abría la puerta a intentos de colusión.

2.2. Una licitación multi-unidades logística

Se decidió entonces dejar de lado la idea de licitar por unidades territoriales definidas a priori y se procedió a generar la siguiente propuesta, que fue la que finalmente se utilizó en la licitación. Cada empresa tendría que fijar un precio unitario por brindar el servicio en una escuela y asimismo decidiría al ofertar en qué escuelas participar, de acuerdo a las restricciones logísticas. El precio debía ser el mismo para todas las escuelas por las que realizaba ofertas, siendo así un "precio unitario por ítem", independiente de la ubicación de la escuela. Además, cada empresa tenía la opción de ofrecer descuentos por volumen (con tramos tarifarios prefijados). Estas consideraciones ubican a este formato de licitación dentro de lo que damos en llamar en este trabajo "licitación multi-unidades

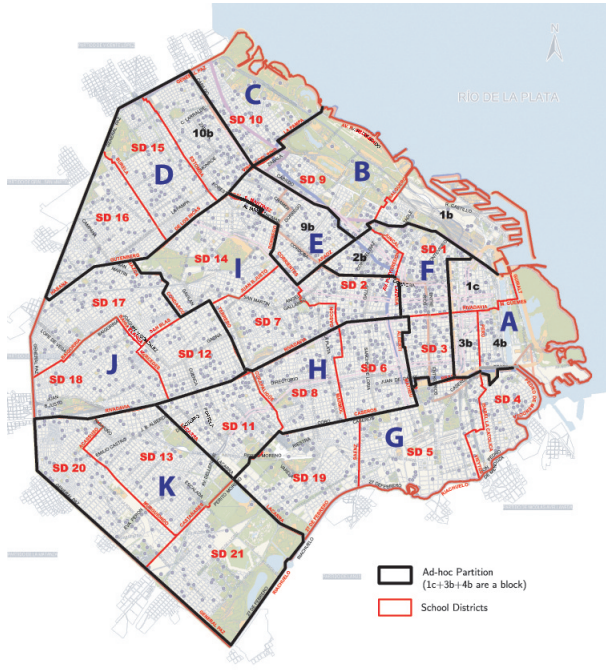

Figura 2: Partición en unidades territoriales considerando radios de acción.

logística".

A su vez, propusimos que se impusiera una cota superior de escuelas a ser asignadas a una misma empresa, para evitar monopolios. Este último punto finalmente no se aprobó, dado que el gobierno de la ciudad consideró que no era un inconveniente asignar todas las escuelas a una misma empresa, si así se conseguía el mejor precio para el estado.

Un potencial problema que tiene este mecanismo (compartido con la propuesta original del gobierno de la ciudad) es que podría suceder que para alguna escuela haya solamente una empresa interesada, y que esa empresa oferte un precio excesivamente alto (aunque en ese caso estaría obligada a ofrecer ese precio alto como su precio unitario para toda la ciudad). En estos casos, la ciudad puede declarar "desierto" este ítem de la licitación, y proveerlo de Internet por medios propios (por ejemplo, contratando a la misma empresa en forma particular al precio del mercado). Un mecanismo equivalente consiste en establecer un "precio de reserva" para cada unidad, de modo tal que si el mejor precio ofertado esta por debajo del precio de reserva, la unidad no se asigna.

Esta nueva propuesta resolvía prácticamente todos los problemas

planteados en las propuestas anteriores:

- No se puede poner un precio alto donde hay poca competencia y un precio más bajo donde hay mucha competencia, a causa de que el precio unitario es el mismo para todas las escuelas (dentro de cada tramo tarifario).

- Se capturan los descuentos por volumen a través de las rebajas según el tramo tarifario, buscando así el menor precio total para el estado.

- Es prácticamente imposible que dos empresas realicen intentos de colusión en zonas de baja competencia. Por ejemplo, si las dos empresas se ponen de acuerdo para "repartirse" entre ellas las escuelas de una zona de baja competencia a un precio alto, automáticamente estarían perjudicando sus posibilidades en las zonas de alta competencia, dado que el precio unitario es el mismo para todas las escuelas. Como se mencionó en el párrafo anterior, tampoco es conveniente para estas empresas poner un precio excesivamente alto para estas escuelas (apostando a quedarse solamente con las escuelas en zonas de baja competencia pero a un precio muy alto), porque los responsables de la licitación pueden declarar desiertas a estas escuelas en ese caso.

- Es la propia empresa quien define su radio de acción a través de la elección de escuelas.

Éste fue el diseño de licitación que finalmente se aplicó en el llamado. Este mecanismo de licitación es una combinación entre (a) la intención inicial de licitar cada escuela por separado por parte de la ciudad (evitando así que alguna empresa pudiera objetar la licitación al estar conformada por unidades territoriales poco coincidentes con su radio de acción) y (b) la necesidad de generar competencia entre las empresas ofreciendo descuentos por volumen. Dado que cada escuela se licita por separado, este mecanismo corresponde a una licitación multi-unidades, pero además cada empresa define su propio radio de acción y, en consecuencia, se siguen generando unidades territoriales como regiones elementales de competencia directa entre las empresas. La diferencia con otras licitaciones combinatoriales existentes en la literatura es que estas regiones se definen *a posteriori* de la recepción de las ofertas, puesto que

están dadas por las intersecciones maximales de los conjuntos de escuelas para los cuales cada empresa manifestó su interés. Estas intersecciones maximales las llamaremos a lo largo de este trabajo "unidades de competición".

3. Formulación matemática de la licitación

Describimos en esta sección dos modelos de programación lineal entera para adjudicar las ofertas ganadoras en forma óptima, minimizando los costos totales para el estado. Presentamos en primer lugar una formulación exponencial que muestra claramente la naturaleza combinatorial de la licitación, y luego una formulación compuesta por un número polinomial de variables y restricciones, que fue eficiente de resolver y fue la utilizada en la práctica. En ambos modelos se busca la solución óptima para la ciudad, adjudicando todas las escuelas.

3.1. Formulación exponencial

Sea C el conjunto de empresas y sea E el conjunto de escuelas. Para cada empresa $i \in C$, definimos $C_i \subseteq E$ como el conjunto de escuelas para las cuales la empresa realizó ofertas. Con estas definiciones, para cada empresa $i \in C$ y cada subconjunto $S \subseteq C_i$, introducimos la variable binaria x_{iS}, de modo tal que $x_{iS} = 1$ si la empresa i recibe exactamente el conjunto S de escuelas, y $x_{iS} = 0$ en caso contrario.

Finalmente, para cada $k = 0, \ldots, |E|$, llamamos γ_{ik} al precio unitario por escuela que solicita la empresa i en caso de recibir exactamente k escuelas. Con estas definiciones, se puede plantear el siguiente modelo de programación lineal entera para el problema:

$$\text{mín} \quad \sum_{i \in C} \sum_{S \subseteq C_i} \gamma_{i,|S|} \, |S| \, x_{iS}$$

$$\sum_{i \in C} \sum_{S \subseteq C_i : j \in S} x_{iS} = 1 \qquad \forall \, j \in E \tag{1}$$

$$\sum_{S \subseteq C_i} x_{iS} = 1 \qquad \forall \, i \in C \tag{2}$$

$$x_{iS} \in \{0,1\} \qquad \forall i \in C, \; \forall S \subseteq C_i \tag{3}$$

La función objetivo solicita minimizar el costo total. Las restricciones (1) especifican que cada escuela debe ser asignada a exactamente una empresa, y las restricciones (2) imponen que cada empresa debe recibir exactamente un subconjunto de las escuelas por las que ofertó. La naturaleza de las variables se define en (3).

Este modelo explicita la naturaleza combinatorial de la licitación, dado que cada empresa finalmente recibirá un subconjunto de las escuelas por las que realizó ofertas, cobrando un precio unitario por escuela que depende de la cantidad de escuelas asignadas.

Sin embargo, desde el punto de vista computacional tiene el problema de que está compuesto por un número exponencial de variables (en el número de escuelas), con lo cual es prácticamente imposible de utilizar a menos que se implemente algún mecanismo de generación de columnas. Por otra parte, este modelo presenta un elevado nivel de simetrías: si por un subconjunto $T \subseteq E$ de escuelas varias empresas realizaron ofertas y T se particiona en más de una empresa, entonces cualquier partición de T que mantenga el número de escuelas asignadas a cada empresa es una solución alternativa con la misma función objetivo. Esta propiedad es conocida como *simetría* en el contexto de la programación lineal entera, puede dificultar enormemente la resolución computacional del modelo involucrado, y ha sido altamente estudiada en la literatura (ver por ejemplo ([4, 7, 8]).

Por otra parte, este alto nivel de simetría conspira contra la pretensión de encontrar todos los óptimos alternativos del problema. En el contexto de una licitación, es crucial determinar todas las soluciones óptimas del modelo, para ponerlas en manos de los decisores a cargo de la licitación. Al haber tantas soluciones equivalentes, determinar si existe alguna solución óptima *esencialmente distinta* puede no ser una tarea sencilla con este modelo.

Por estos motivos, se diseñó el modelo que se presenta en la siguiente sección, que evita estas simetrías recurriendo a variables enteras generales.

3.2. Formulación polinomial

Este modelo surge de observar que si para un subconjunto de escuelas hay un mismo grupo de empresas interesadas, entonces no es relevante para la optimización determinar qué escuelas recibe cada empresa,

sino que alcanza con determinar *cuántas* escuelas del subconjunto son asignadas a cada firma. Llamamos *región* a un conjunto maximal de escuelas con estas características; es decir, un conjunto de escuelas para las cuales exactamente las mismas empresas realizaron ofertas, y tal que no está estrictamente contenido en otro conjunto que tiene esta misma propiedad. Estas regiones son una especie de "unidades de competencia", en las cuales las mismas empresas se disputan las escuelas, que a su vez son indistinguibles para la decisión final.

Las regiones se arman entonces a posteriori en función de las escuelas que eligió cada empresa, y el modelo determina cuántas escuelas se asignan a cada empresa en cada una de las regiones que se generaron con las ofertas de las empresas (no necesariamente una región se asigna íntegra a una empresa). A posteriori de la resolución del modelo, si hubiera regiones cuyas escuelas fueron asignadas a más de una empresa, se asigna qué escuelas van para cada empresa en cada región. Este proceso se puede realizar de manera manual o de manera algorítmica, siguiendo criterios de proximidad geográfica. Es importante mencionar que esta asignación no tiene impacto en la función objetivo final.

1. Parámetros del modelo:

   - C: conjunto de empresas;
   - R: conjunto de regiones, definido por la intersección de las ofertas de cada empresa;
   - E_r: conjunto de escuelas de la región $r \in R$;
   - p_{ji}: 1 si la empresa i ofrece el servicio en la región j, 0 si no, para toda región $j \in R$ y para toda empresa $i \in C$;
   - T: conjunto de tramos tarifarios, en este caso particular se consensuó con el gobierno de la ciudad $T = \{0\text{–}19,\ 20\text{–}39,\ \ldots,\ 80\text{–}99,\ 100\text{–}149,\ 150\text{–}199,\ 200\text{–}299,\ \ldots,\ 600\text{–}699,\ 700\text{–}709\}$;
   - $\min_t$ y $\max_t$: los límites inferior y superior de escuelas para el tramo $t \in T$;
   - c_{ti}: costo por escuela en el tramo $t \in T$ ofrecido por la empresa $i \in C$, de modo tal que si una empresa recibe entre $\min_t$ y $\max_t$ escuelas, entonces cobrará un valor de \$ c_{ti} por cada una.

2. Variables del modelo:

   - $x_{ji} \in \mathbb{Z}_{\geq 0},\ j \in R,\ i \in C$: cantidad de escuelas en la región j asignadas a la empresa i;
   - $y_{it} \in \{0,1\},\ i \in C,\ t \in T$: variable que define si a la empresa i se le aplica el tramo tarifario t;
   - $z_{it} \in \mathbb{Z}_{\geq 0},\ i \in C,\ t \in T$: cantidad de escuelas asignadas a la empresa i en el tramo tarifario t.

3. Formulación del modelo:

$$\begin{aligned}
\text{mín} \quad & \sum_{i \in C} \sum_{t \in T} c_{ti} z_{it} & & \\
\sum_{i \in C} x_{ji} &= |E_j| \qquad \forall\, j \in R & & (4)\\
\sum_{j \in R} x_{ji} &\geq \min_t - M(1 - y_{it}) \qquad \forall\, i \in C,\ \forall\, t \in T & & (5)\\
\sum_{j \in R} x_{ji} &\leq \max_t + M(1 - y_{it}) \qquad \forall\, i \in C,\ \forall\, t \in T & & (6)\\
\sum_{t \in T} y_{it} &= 1 \qquad \forall\, i \in C & & (7)\\
z_{it} &\geq \sum_{j \in R} x_{ji} - M(1 - y_{it}) \qquad \forall\, i \in C,\ \forall\, t \in T & & (8)\\
x_{ji} &\leq p_{ji}|E_j| \qquad \forall\, i \in C,\ \forall\, j \in R & & (9)\\
x_{ji} &\in \mathbb{Z}_{\geq 0} \qquad \forall\, j \in R,\ \forall\, i \in C & & (10)\\
y_{it} &\in \{0,1\} \qquad \forall\, i \in C,\ \forall\, t \in T & & (11)\\
z_{it} &\in \mathbb{Z}_{\geq 0} \qquad \forall\, i \in C,\ \forall\, t \in T & & (12)
\end{aligned}$$

La función objetivo busca minimizar el costo total. Las restricciones (4) especifican que se deben cubrir todas las escuelas de cada región. Las restricciones (5) y (6) vinculan las variables x con las variables y, de modo tal que $y_{it} = 1$ si la empresa i recibe un número de escuelas incluido dentro del tramo tarifario t. Para nuestro caso particular, tomamos $M = 709$. Las restricciones (7) especifican que cada empresa debe estar asociada a un único tramo tarifario. Las restricciones (8) fuerzan a que z_{it} sea a lo menos la cantidad total de escuelas asignadas a la empresa, siempre que $y_{it} = 1$ (es decir, siempre que la empresa deba utilizar el tramo tarifario t). Las restricciones (9) indican que no se pueden asignar a una empresa más escuelas que las existentes en una región, si es que

dicha empresa participa en esa región, y que no se puede asignar a una empresa una escuela de una región en la cual no participa. Finalmente, las restricciones (10)-(12) especifican la naturaleza de las variables. Es interesante mencionar que las variables z se pueden definir como reales no negativas (es decir, $z_{it} \in \mathbb{R}_{\geq 0}$ para $i \in C$ y $t \in T$), ya que en la solución óptima van a resultar enteras por las restricciones del modelo.

Notar que el número de variables y de restricciones de esta formulación está acotada superiormente por el número de escuelas (dado que el número de regiones está acotada por el número de escuelas, si asumimos como sucede en la práctica que el número de empresas oferentes y el número de tramos tarifarios es mucho menor que el número de escuelas). Esta formulación es más eficiente que aquella más natural (y también polinomial en el número de escuelas) que se obtiene considerando a cada escuela en forma individual; es decir, con una variable binaria por cada escuela y por cada empresa, que determine si la escuela es asignada a la empresa o no. Además de estar compuesta por un número mayor de variables y restricciones, esta formulación tendría serios problemas de simetría, y esta última característica impactaría negativamente en el procedimiento considerado en la próxima sección.

3.3. Búsqueda de óptimos múltiples

Una característica interesante de los modelos de programación lineal entera aplicados a licitaciones es que no sólo se debe encontrar una solución óptima, sino que se deben poner a disposición de los decisores *todas* las soluciones óptimas, con el objetivo de realizar un proceso transparente y que no perjudique a ninguna empresa. En caso de que exista más de una solución óptima, la decisión sobre cómo asignar queda a criterio de los funcionarios responsables del proceso licitatorio.

Por ello, una vez obtenido el óptimo del modelo anterior, agregamos nuevas restricciones al modelo de modo que la solución obtenida deje de ser factible y no se pierda ninguna otra solución factible. Volvemos a correr entonces el modelo para ver si obtenemos el mismo valor de la función objetivo o uno mayor. Repetimos este procedimiento mientras sigamos obteniendo el mismo valor de la función objetivo original, generando así todos los óptimos alternativos.

Describimos a continuación las restricciones que deben ser agregadas a fin de excluir sólo al punto óptimo del conjunto de las soluciones factibles.

Sea $x_{ji} = a_{ji}$ la solución óptima, para cada empresa $i \in C$ y cada región $j \in R$. Para cada valor $a_{ji} > 0$, agregamos dos variables binarias w_{ji} y w'_{ji}, que tomarán valor 1 si $x_{ji} < a_{ji}$ y $x_{ji} > a_{ji}$, respectivamente (y 0, en caso contrario). Para expresar esta situación y lograr que al menos una de las variables x cambie su valor en la nueva solución óptima, agregamos las siguientes restricciones:

$$\begin{aligned} x_{ji} &\geq (a_{ji}+1)w_{ji} && \forall\, i \in C, j \in R \text{ tales que } a_{ji} \neq 0 \\ 709 - x_{ji} &\geq (709-(a_{ji}-1))w'_{ji} && \forall\, i \in C, j \in R \text{ tales que } a_{ji} \neq 0 \\ \sum_{a_{ji}\neq 0} (w_{ji} + w'_{ji}) &\geq 1 \end{aligned}$$

El agregado de estas nuevas variables y restricciones potencialmente puede complicar los tiempos de resolución del modelo, en especial luego de varias iteraciones de eliminación de óptimos alternativos. Esto puede suceder si hay varias empresas que realizan ofertas similares, dando origen a muchas soluciones óptimas pero esencialmente distintas. Finalmente, es interesante mencionar que este proceso de eliminación de óptimos alternativos es posible porque el modelo determina cantidades de escuelas "equivalentes" a ser asignadas a cada empresa, y no contiene un nivel de detalle por escuela. Si éste fuera el caso, la cantidad de óptimos alternativos al modelo (pero correspondientes a soluciones esencialmente equivalentes) sería un número excesivamente grande para todo fin práctico.

4. Ofertas y resultados

Cuatro empresas participaron del proceso licitatorio. La empresa A ofertó por las 709 escuelas, lo que era previsible dado que era sabido que contaba con cobertura al momento de la licitación en toda la ciudad. La empresa B ofertó por 348 escuelas, abarcando toda la zona central de la ciudad. La empresa C ofertó por 99 escuelas en la zona norte de la ciudad, mientras que la empresa D ofertó por 97 escuelas también en el norte de Buenos Aires (la zona de mayores recursos de la ciudad y donde se esperaba que hubiera más competencia). En la Figura 3 pueden verse los sectores ofertados por cada empresa y en la Figura 4 se observan las

seis regiones (o "unidades de competencia") que quedaron determinadas a posteriori de las ofertas. Notar que en la zona sur de la ciudad quedaron determinadas 248 escuelas donde sólo ofertó la empresa A. Es importante destacar que ni la empresa ni el gobierno de la ciudad sabían previo a las ofertas que esta empresa iba a ser la única en presentarse en la región sur de la ciudad.

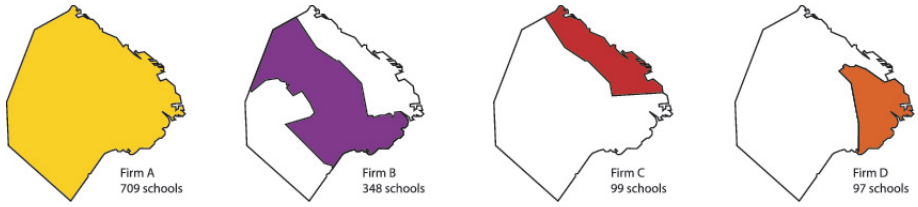

Figura 3: Ofertas de las 4 empresas.

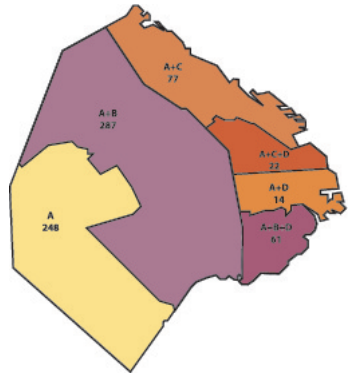

Figura 4: Regiones determinadas una vez conocidas las ofertas de cada empresa.

En la Figura 5 presentamos los valores ofertados por cada empresa para cada tramo tarifario. Cabe notar que el gobierno de la ciudad estimaba como un buen precio un valor del abono mensual cercano a los U$S 250 y que las cuatro empresas ofertaron en su mejor precio valores en torno de esa cifra. La Figura 6 muestra estos mismos valores en forma gráfica, poniendo en evidencia que la empresa A armó su oferta con el propósito de entregar el servicio a las 709 escuelas.

El modelo arrojó como resultado que la empresa A se adjudica el servicio para las 709 escuelas a un costo total mensual para la ciudad de Buenos Aires de U$S 166.501 (y este es el único óptimo del problema), lo

	Firm A		Firm B		Firm C		Firm D	
	Unit price: $ 1,174.18		Unit price: $ 665.50		Unit price: $ 497.92		Unit price: $ 422.51	
Interval	Discount	Unit price	Discount	Unit price	Discount	Unit price	Discount	Unit price
1 - 19	0%	$ 1,174.18	0%	$ 665.50	0%	$ 497.92	5%	$ 401.38
20 - 39	0%	$ 1,174.18	18%	$ 545.71	0%	$ 497.92	10%	$ 380.25
40 - 59	0%	$ 1,174.18	28%	$ 479.16	0%	$ 497.92	20%	$ 338.00
60 - 79	0%	$ 1,174.18	32%	$ 452.54	33%	$ 268.88	25%	$ 316.88
80 - 99	0%	$ 1,174.18	40%	$ 399.30	45.02%	$ 222.52	33%	$ 283.08
100 - 149	10%	$ 1,056.76	50%	$ 332.75	---	---	---	---
150 - 199	15%	$ 998.05	59%	$ 272.86	---	---	---	---
200 - 299	20%	$ 939.34	61%	$ 259.55	---	---	---	---
300 - 399	30%	$ 821.92	70.5%	$ 196.32	---	---	---	---
400 - 499	40%	$ 704.51	---	---	---	---	---	---
500 - 599	50%	$ 587.09	---	---	---	---	---	---
600 - 699	60%	$ 469.67	---	---	---	---	---	---
700 - 709	80%	$ 234.84	---	---	---	---	---	---

Figura 5: Ofertas de cada empresa para cada tramo tarifario.

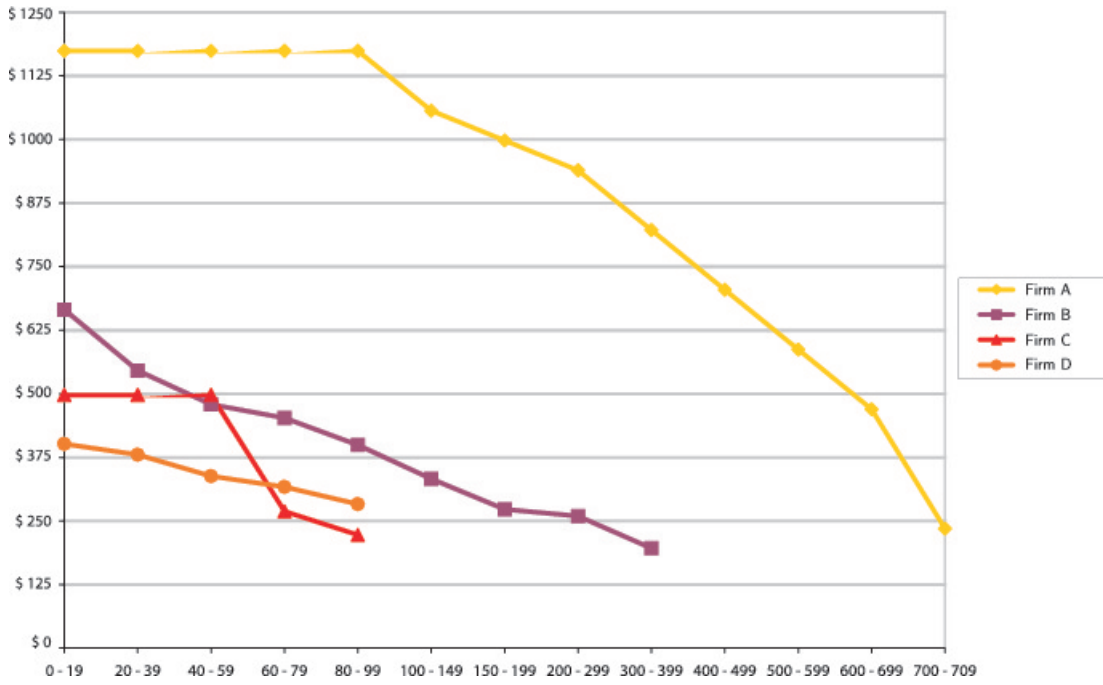

Figura 6: Gráfico de las ofertas de las 4 empresas.

que da una erogación total para los dos años de U$S 3.996.024. El costo promedio mensual por escuela es de U$S 234,84 (la oferta hecha por la empresa A para el último tramo tarifario).

5. Conclusiones

Consideramos que el principal aporte de este trabajo es el de proponer un nuevo formato de licitación, que llamamos *licitación multi-unidades logística*, que es útil para el caso en el que los ítems licitados tienen interés individual, se buscan descuentos por volumen, y tal que la provisión del servicio para un ítem depende de su ubicación geográfica y esta valoración es distinta para cada oferente. Estas consideraciones de orden logístico determinan "unidades de competencia" entre las empresas (llamadas también *regiones* en la Sección 3.2), cuya definición permite la formulación de un modelo de programación lineal entera compacto y que elimina las simetrías presentes en los modelos que identifican individualmente a cada ítem de la licitación.

Este tipo de licitación es interesante para el caso en el cual los costos marginales de la provisión del servicio son bajos o nulos, y estas consideraciones típicamente se originan cuando la provisión del servicio ofertado depende de tecnología instalada (que denominamos "consideraciones logísticas" en este trabajo). En el caso particular de la provisión del servicio de Internet, una empresa que ya tiene tecnología instalada en las inmediaciones de una escuela puede proporcionar el servicio con un costo de instalación mínimo, correspondiente a las horas de trabajo del técnico que realiza la instalación (y que a los efectos del precio total, puede considerarse prácticamente nulo). Por el contrario, si la empresa no tiene tecnología instalada cercana a la escuela, el costo de proveer el servicio es muy alto, y en ese caso el precio por el servicio también lo será debido a que debe incluir los costos de instalación de tecnología. En este sentido, los costos de la provisión del servicio son casi en su totalidad "costos hundidos", y debido a estas consideraciones la definición de las escuelas por las cuales cada empresa está interesada en proveerle el servicio pasa a ser un componente crucial de la licitación. El formato de licitación propuesto en este trabajo es útil para aprovechar estas características de la estructura de costos de las empresas participantes.

Como se mencionó en la Sección 2.2, este formato impide que un oferente ponga un precio alto donde hay poca competencia y un precio más bajo donde hay más competencia, a causa de los precios unitarios iguales para todos los ítems. Por otra parte, es difícil que dos empresas realicen intentos de colusión. Finalmente, el formato de la licitación incluye descuentos por volumen, y a pesar de que se trata de una licitación con características combinatoriales, no se definen conjuntos de ítems (en nuestro contexto, las "unidades territoriales") de antemano, sino que los propios oferentes definen los conjuntos de ítems por los cuales están interesados.

En cuanto al potencial ahorro que este modelo de licitación le generó a la ciudad de Buenos Aires en contraposición a una licitación multiunidades estándar, podemos hacer el siguiente análisis. Supongamos de manera optimista que en todas las escuelas donde hubo más de un postulante se hubiera podido conseguir el mejor precio ofrecido (U$S 196,32, el mejor precio ofertado por la empresa B), y que en las escuelas donde la empresa A terminó siendo monopólica, esta empresa hubiera ofertado su segundo mejor precio (U$S 469,67). En ese caso el costo total mensual para la ciudad hubiera sido de U$S 206.982,83, lo que significa en base a esta estimación que se obtuvo un 20 % de ahorro para la ciudad. Este 20 % llevado a los dos años de contrato implica un ahorro global cercano a U$S 800.000.

Es interesante analizar cuál habría sido el costo para la ciudad si hubiera sido posible contratar dentro de cada región a la empresa con el mejor precio final, correspondiente a su último tramo tarifario (aunque no se correspondiera con la cantidad de escuelas asignadas a la empresa). Por ejemplo, en la región determinada por las empresas A y B, se contrataría a la empresa B por su mejor precio, que es de U$S 196,32. Esta asignación no es realista porque cada empresa ofertó su mejor precio por un número alto de escuelas y puede no estar recibiendo esa cantidad en este caso, pero proporciona una cota inferior muy optimista del precio total que podría haber pagado la ciudad. Si se realizara esta asignación, el precio total por mes abonado por la ciudad sería de U$S 151.704,44, que es sólo 8.8 % más bajo que el precio finalmente obtenido por la ciudad. Este porcentaje es relativamente bajo, y da una idea de que el diseño de la licitación generó una alta competencia entre las empresas participantes.

Dadas las ofertas recibidas, la solución final asigna todas las escuelas

al mismo oferente (la empresa A, en nuestro caso) por un precio unitario de U$S 234.84. Es interesante analizar cuál hubiera sido el precio unitario más alto que podía haber ofertado la empresa A en el último tramo tarifario, de modo tal que siguiera recibiendo la provisión del servicio para todas las escuelas. Este valor puede calcularse corriendo el modelo de la Sección 3.2 con diferentes valores para el último tramo tarifario de la empresa A (realizamos una búsqueda binaria sobre el rango de valores posibles para dicho precio, hasta obtener el valor límite buscado). En este caso, el valor máximo resulta de U$S 401.38, es decir un 71 % por encima del valor ofertado. Este valor sugiere que la licitación fue efectivamente competitiva y que no hubo colusión entre los oferentes.

Agradecimientos: El presente trabajo fue parcialmente financiado por los proyectos ANPCyT PICT-2012-1324 (Argentina), CONICET PIP 112-200901-00178 (Argentina), UBACyT 20020100100980 (Argentina), y por el Instituto Milenio "Sistemas Complejos de Ingeniería" (Chile). El tercer autor es parcialmente financiado por el proyecto FONDECyT 1110797 (Chile). Los autores quieren agradecer a la Agencia en Sistemas de Información (ASI) del gobierno de la ciudad de Buenos Aires, responsable de la organización de la licitación, y en particular a Julián Dunayevich y Eduardo Terada, funcionarios de la ASI durante la concreción de este proyecto, por su colaboración para la realización de este proyecto. También agradecen a los dos revisores anónimos por sus sugerencias que permitieron mejorar la versión final del trabajo.

Referencias

[1] P. Cramton, Y. Shoham, and R. Steinberg, Combinatorial Auctions, *MIT Press*, 2006.

[2] R. Epstein, L. Henríquez, J. Catalán, G. Weintraub, and C. Martínez, A Combinatorial Auction Improves School Meals in Chile, *Interfaces* **32(6)** (2002), 1–14.

[3] S. Glaister, and M. Beesley, Bidding for Tendered Bus routes in London, *Transportation Planning and Technology* **15** (1991), 349–366.

[4] R.G. Jeroslow, Trivial integer programs unsolvable by branch-and-bound, *Mathematical Programming* **6** (1974), 104-109.

[5] D. Kennedy, London bus tendering: an overview, *Transport Reviews* **15(3)** (1995), 253–264.

[6] A. Kwasnica, and K. Sherstyuk, Multi-Unit Auctions, manuscript (2012), (http://www2.hawaii.edu/˜ katyas/pdf/Kwasnica_Sherstyuk_multiunit_101912_wp.pdf).

[7] F. Margot, Symmetry in Integer Linear Programming, *in 50 Years of Integer Programming*, Springer (2009).

[8] P. A. Rey, Eliminating Redundant Solutions of Some Symmetric Combinatorial Integer Programs, *Electronic Notes in Discrete Mathematics* **18** (2004), 201–206.

Una aplicación de Web Opinion Mining para la extracción de tendencias y tópicos de relevancia a partir de las opiniones consignadas en blogs y sitios de noticias

Rodrigo Dueñas F. [*]
Juan D. Velásquez[*]

Resumen

El análisis de tendencias se ha abordado tradicionalmente a través de la realización de encuestas, las cuales poseen un alto contenido de subjetividad y las respuestas se ven constantemente afectadas por factores exógenos al evento bajo estudio. Este exceso de factores exógenos y subjetividad puede conducir a errores significativos, basta con ver los resultados de la última encuesta para la elección de alcaldes, la que predijo de manera errónea qué candidatos ganarían en las comunas mas emblemáticas de Santiago. En este trabajo, presentamos una metodología alternativa para detectar tendencias en la Web, a través del uso de técnicas de recuperación de la información, modelamiento de tópicos y minado de opiniones. Dado un conjunto de sitios semilla, se procede a extraer los tópicos que se mencionan en los documentos recuperados desde ellos y posteriormente se acude a las redes sociales para obtener la opinión por parte de sus usuarios en relación a estos. Usando esta metodología de detección de tendencias es posible complementar la información extraída a través de metodologías tradicionales para predecir eventos y reducir los efectos de los factores exógenos introducidos por los medios tradicionales.

Palabras Clave: *Opiniones*, *Tendencias*, *Web Opinion Mining*, *Tópicos*, *Blogs*, *Noticias*.

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

1. Introducción

Con la aparición de las aplicaciones web que permiten la creación de contenido y la colaboración por parte de los usuarios, como lo son *wikis* y *blogs*, (las cuales darían el puntapié inicial a lo que ahora se conoce como la Web 2.0) la función de la Web en la sociedad mundial se vio fuertemente potenciada. Esta dejó de ser tan sólo un repositorio de información y se transformó en un canal interactivo entre todas las entidades que la componen, tanto usuarios como proveedores de información. Este cambio de paradigma permitió que todos ellos pudiesen contribuir activamente a la creación de contenido, provocando un explosivo aumento en la participación de sus usuarios en la Web, y por consiguiente de la cantidad de información y conocimiento disponible en ella.

Junto a ello, Internet trajo consigo cambios drásticos en la manera que se interactúa en el mercado empresarial. Estos cambios provocan que una empresa pueda crecer en muchas direcciones y no sólo aumentando la cantidad de productos que produce o el número de personas a las que ofrece sus servicios. Así, una vez que una empresa decide crecer, ya sea expandiendo negocios hacia nuevos mercados u ofreciendo nuevos productos y servicios, la cantidad de información externa que debe abarcar para poder realizar una buena toma de decisiones se vuelve cada vez mayor, por lo que debe analizar un conjunto siempre creciente de fuentes de información para poder recuperar el conocimiento necesario para que este análisis sea valioso para la empresa.

En esta misma línea, es cada vez más necesario ser capaz de manejar grandes volúmenes de datos para gestionar de la mejor manera posible los recursos que se disponen, y al mismo tiempo anticiparse a cada movimiento que realizará la competencia en busca de obtener ventajas competitivas, o impedir que otros las obtengan, para ser líderes en el mercado. El primer problema al que se debe enfrentar una empresa sumergida en el mundo globalizado, es el más complejo desde el punto de vista de la gestión de operaciones, por lo que varias metodologías y herramientas han nacido proponiendo soluciones, entre las cuales se encuentran los *Data Warehouses*, la *Business Intelligence* y el recientemente acuñado término de *BigData*. El segundo problema no sólo involucra a la gestión de operaciones, ya que es necesario tener un equipo multidisciplinario encargado constantemente de monitorear el mercado, las acciones de las otras empresas, los anuncios presentes en los medios y cualquier indicio que permita anticiparse a los lanzamientos de productos y servicios de la competencia.

Una posible solución al problema planteado es minar la web en busca de

esos indicios de manera automática, con foco en qué tópicos se habla en la web, y analizando las redes sociales para estimar que percepción poseen los cibernautas sobre ellos. Es factible plantear la hipótesis de que a través de realizar un análisis de gran parte del conocimiento objetivo generado por los usuarios y los medios se puede atisbar aquellos indicios claves a la hora de plantear una planificación estratégica y operacional. Un sistema capaz de realizar esto de manera aproximada es realizable utilizando técnicas de recuperación de la información, modelamiento de tópicos y minado de opiniones sobre fuentes cuidadosamente seleccionadas que sean capaces de otorgarle al sistema una muestra significativa de todo lo que se habla sobre los mercados en los cuales se ve inmersa la empresa a nivel competitivo.

Por lo mencionado anteriormente, se plantea como hipótesis de investigación que es posible extraer tendencias y obtener una representación aproximada del comportamiento de estas a través del análisis de los documentos presentados en sitios de noticias y las opiniones consignadas en las redes sociales por parte de sus usuarios.

En la segunda sección de este artículo se da a conocer el estado del arte respecto de las técnicas de recuperación de información, modelamiento de tópicos en documentos y finalmente sobre algoritmos de minado de opiniones. En la sección 3, se da a conocer en detalle el modelo propuesto para la detección de tendencias en la Web, en particular la detección de tópicos y el minado de opiniones referentes a estos, los cuales serán evaluados a través de los experimentos presentes en la sección 4. Para finalizar, en la quinta sección de este artículo se presentan las conclusiones relevantes al trabajo desarrollado y posibles ramas de investigación a futuro.

2. Trabajo relacionado

2.1. Modelos de Tópicos

Un modelo de tópicos tiene como objetivo identificar las relaciones latentes entre documentos pertenecientes a una colección, con el fin de dar una descripción sucinta de esta sin perder información desde el punto de vista estadístico.

El precursor de los modelos de tópicos es David Blei, el cual en [4] describe de manera detallada los modelos de tópicos y las aplicaciones de estos. En ella, se define un tópico como el conjunto de elementos que pueden representar una temática presente en una colección de documentos sin pérdida de información estadística. Por ejemplo, si existe una colección de documentos textuales

que abarca múltiples temas, un tópico es un conjunto de palabras que logra describir estadísticamente uno de estos temas.

Entre los modelos de tópicos existente, los más utilizados son los desarrollados por Blei *et al.* Entre ellos, los más populares son el modelo LDA (Latent Dirichlet Allocation) [4] y el modelo CTM (Correlated Topic Model) [3].

Estos modelos de tópicos se cimentan sobre las siguientes definiciones:

- Una *palabra* w es la unidad elemental de un documento textual y se define como un elemento de un vocabulario indexado V. Para efectos de estos modelos, para representar una palabra se hace uso de vectores unitarios en donde la n-ésima palabra de V se representa con un vector de largo $|V|$ en el cual sólo su componente n-ésima es igual a 1.
- Un *documento* es un arreglo de palabras descrito como $\mathbf{w} = (w_1, w_2, ..., w_N)$, donde w_n es la n-ésima palabra de este.
- Un *corpus* es una colección de documentos descrita como $D = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$.
- Un *tópico* es una distribución de probabilidad sobre un vocabulario V fijo. Por ejemplo, el tópico *política* está descrito por palabras como *partido, diputado, senado, ley* de manera frecuente y palabras como *guerra, marcador, gol* con probabilidad casi nula.

A continuación se da a conocer una descripción de cada uno de los modelos mencionados anteriormente, dando a conocer las diferencias entre estos y las principales características de cada uno de ellos.

2.1.1. Latent Dirichlet Allocation

El modelo llamado *Latent Dirichlet Allocation*[4] es considerado el más sencillo de los modelos de tópicos presentes hoy en día, y por ello es utilizado frecuentemente en aplicaciones que requieran obtener información sobre colecciones de documentos de manera rápida y eficiente.

El modelo LDA trabaja bajo el supuesto de que los tópicos presentes en la colección de documentos que se está analizando no necesariamente están relacionados y por consiguiente no dependen entre ellos.

Para extraer la estructura de tópicos presente en una colección, este modelo hace uso de un modelo estadístico de generación de documentos, tópicos y palabras a lo largo del tiempo que abarque esta. El siguiente proceso se realiza para cada documento presente en una colección:

1. Definir una distribución aleatoria para la presencia de los tópicos en la colección y una distribución para la presencia de las palabras para cada tópico que se desea encontrar.

2. Luego, por cada palabra presente en el documento bajo análisis se debe:

   a) Escoger un tópico aleatoriamente haciendo uso de la distribución generada en el paso 1.

   b) Escoger una palabra del documento aleatoriamente a partir de la distribución del vocabulario en relación al tópico escogido.

Formalmente, para determinar la estructura de tópicos existente luego del proceso de generación, es necesario calcular las distribuciones condicionales entre los tópicos y sus documentos, el cual es un problema NP completo debido a que la cantidad de estructuras que pueden representar una colección de documentos crece exponencialmente en relación a la cantidad de documentos y palabras presente en esta. Este proceso es descrito formalmente como sigue:

1. Escoger $N \sim Poisson(\xi)$
2. Escoger $\theta \sim Dirichlet(\alpha)$
3. Para cada palabra w_n en $\mathbf{w}$

   a) Escoger un tópico $z_d \sim Multinomial(\theta)$

   b) Escoger una palabra w_d a partir de $p(w_n|z_n, \beta)$, la distribución multinomial de probabilidades condicionada sobre el tópico z_n.

Donde cada variable del proceso corresponde a:

- β es la matriz de probabilística de que el documento contenga la palabra w^j dado que discute el tópico z^i, con $B_{ij} = p(w^j = 1|z^i = 1)$.
- θ_d es la distribución de tópicos para el documento d, es decir, el conjunto de probabilidades $\theta_{d,k}$ donde esta corresponde a la probabilidad de que el documento d trate del tópico k.
- z_d son las asociaciones de tópicos para el documento d con $z_{d,n}$ es el tópico asociado a la palabra n-ésima del documento d
- w_d es el conjunto de palabras presentes en el documento d.
- $w_{d,n}$ es la palabra n-ésima del documento d.

A partir de esto, es posible definir el proceso generativo de documentos a través de la distribución conjunta de variables observables y no observables como se define en la ecuación 1. La solución a esta ecuación puede ser obtenida haciendo uso de algoritmos de inferencia estadística como el algoritmo *Sampleo*

de Gibbs, los que además de estimar la estructura de tópicos de una colección, permiten inferir la estructura de tópicos presente en otros *corpus* que estén compuestos de documentos que hablen de temas similares a los utilizados para entrenar el modelo.

$$p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}\right) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n|\beta_{1:K}, z_{d,n}}) \right) \tag{1}$$

2.2. Modelos de extracción de opiniones

Con el nacimiento de las redes sociales y la llegada de la Web 2.0, los usuarios comenzaron ser capaces de generar nuevo contenido en la web y también de dar a conocer sus opiniones sobre variados hechos, productos, servicios y cualquier otro tema que sea susceptible de generar un sentimiento o una opinión en ellos.

Para aprovechar esta nuea información que está siendo generada en la web se han desarrollado una serie de metodologías, algoritmos y técnicas para recuperar información desde documentos opinados. Esta nueva rama de la recuperación de la información es llamada *Web Opinion Mining*, la cual tiene como objetivo principal extraer información a partir de las opiniones que se encuentran en los documentos opinados publicados en la web [9].

Los modelos de opinión son utilizados frecuentemente en donde es necesario hacer uso de las opiniones de los usuarios para evaluar u obtener información sobre productos y servicios. En [15] se menciona que los algoritmos de minado de opiniones son utilizados frecuentemente en detección de spam en review de productos, creación y mejoramiento de sistemas de recomendación de productos y servicios o de avisaje online, evaluación de nuevos productos en la web, evaluar el impacto que tienen las reviews en las utilidades de un negocio o un producto, etc.

Una opinión se define como una creencia subjetiva por parte de un sujeto sobre algún objeto, tema o situación en particular, que nace de una interpretación emocional por parte de éste del objeto bajo análisis o una característica de este [6]. Por consiguiente, una opinión es una creencia subjetiva de un *emisor* sobre el *receptor* o una característica de este, y posee una polaridad que señala el tipo de emoción (positiva o negativa) que da paso a la opinión propiamente tal.

Los modelos de extracción o minado de opiniones trabajan sobre *documentos opinados*, los cuales son definidos en [9] como todo documento que contenga una o más oraciones que expresan una opinión. Por lo tanto, se puede decir que

los modelos de extracción de opiniones buscan determinar qué tipo de emoción motiva la emisión de una opinión en un documento [6] o que polaridad es la predominante en este [18].

Para dar a conocer un modelo de opiniones es necesario presentar una serie de definiciones que dan sustento a la gran mayoría de los modelos utilizados en la actualidad. Estas definiciones son las que siguen:

- **Objeto**: Un *objeto* o es una entidad que puede ser un producto, un servicio, un individuo, una organización, un evento, etc. descrito por la dupla (T, A) donde T es la jerarquía que describe cada una de las componentes del objeto y A es el conjunto de atributos de este. A su vez, cada componente posee su propio conjunto de sub-componentes y atributos.

- **Opinión**: Una *opinión* sobre una característica f objeto o es una evaluación emocional que realiza un *emisor* sobre este o una característica de él.

- **Emisor**: El *emisor* de una opinión es aquella persona u organización que la expresa.

- **Polaridad**: La *polaridad* de una opinión indica si la opinión es *positiva*, *negativa* u *objetiva*.

Además, en el modelo de análisis de opiniones basado en características, un objeto o se describe como un conjunto de características $F = f_1, f_2, ..., f_n$ donde también se incluye el objeto en cuestión como una característica particular. En este caso, cada característica f_i puede ser descrita por el conjunto de palabras o frases $W_i = w_{i1}, w_{i2}, ..., w_{im}$, donde cada w_{ij} es un sinónimo de la característica f_i; además, f_i también puede ser expresada a través del conjunto de indicadores de característica $I_i = i_{i1}, i_{i2}, ..., i_{iq}$.

Bajo este modelo, un documento d que contiene opiniones es descrito como aquel que contiene opiniones sobre un conjunto de objetos $o_1, o_2, ..., o_q$ emitidas por un conjunto de emisores $h_1, h_2, ..., h_p$. En este caso, cada opinión o_j se enfocan en un subconjunto F_j de características del objeto en cuestión y puede ser clasificada en uno de los siguientes tipos:

- **Opinión directa**: Es la quintupla $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$, donde o_j es el objeto sobre el cuál consiste la opinión, f_{jk} es la característica del objeto o_j que está siendo analizada, oo_{ijkl} es la polaridad de la opinión sobre la característica f_{jk}, h_i es el emisor de la opinión y finalmente, t_l es el momento en el cuál h_i expresó la opinión.

- **Opinión comparativa**: Expresa la relación, sea esta de similitud o de diferencia entre dos o más objetos y las preferencias del emisor de la opinión sobre un conjunto común de características entre los objetos.

Toda opinión se basa en las emociones que guían al emisor a emitirla en el momento que este acto sucede. De acuerdo a lo expresado en [9], las emociones son ***sentimientos y pensamientos subjetivos***, y estas se dividen en 6 tipos primarios: *amor, alegría, sorpresa, rabia, tristeza* y *temor*.

Si bien todas las opiniones nacen de una emoción, la manera en que estas son expresadas por el emisor de ellas permite clasificarlas en dos tipos: las opiniones *explícitas*, aquellas en que el emisor expresa claramente la opinión a través de una frase subjetiva; y las *implícitas*, donde la opinión en cuestión es expresada a través del uso de una frase objetiva. Un ejemplo de opinión explícita es *"me encanta el sabor de este helado"* y de opinión implícita es *"la linterna explotó a la semana de haberla comprado"*.

2.2.1. Aplicaciones de los algoritmos de minado de opiniones

Entre las aplicaciones que tienen los algoritmos de minado de opiniones podemos encontrar:

1. **Análisis de reviews de productos**: En [2] se discuten distintas aplicaciones de estos algoritmos en el análisis de reviews de productos, entre ellas se destacan el resumen de opiniones, detectar reviews falsos o spam y la evaluación monetaria de las características de un producto.

2. **Sistemas de recomendación**: En [7] se estudia mejorar sistemas de recomendación de productos a través del uso de las opiniones emitidas por usuarios de estos mismos sistemas.

3. **Política**: En [13] se muestran distintos enfoques para analizar campañas políticas y la percepción de la gente sobre leyes y candidatos políticos.

2.2.2. Algoritmos para extracción de polaridad de opiniones

En la plataforma de detección de tendencias se hace uso de algoritmos de detección de polaridad para determinar qué es lo que se opina en la web sobre los tópicos que son extraídos desde los documentos recuperados. A continuación se dará a conocer las dos afluentes más utilizadas de algoritmos de detección de polaridad en opiniones.

Algoritmos de clasificación a través de aprendizaje supervisado: La mayoría de los algoritmos de aprendizaje supervisado existentes (Naive-

Bayes, Support Vector machines, etc.) pueden ser aplicados a la clasificación de polaridad de documentos tal como se muestra en [16, 11].

El algoritmo más utilizado debido a su simplicidad es un clasificador Naive-Bayes, el cual busca obtener las probabilidades de que un documento d posea la polaridad p $\Pr(p \mid d)$. Este tipo de clasificador obtiene estas probabilidades al resolver el siguiente problema de maximización: $arg \max_{p \in P}\{\Pr(p \mid d)\}$.

Los clasificadores de Naive-Bayes hacen uso de la regla de Bayes para poder simplificar el problema de maximización que deben resolver:

$$p_d = arg \max_{p \in P} \left\{ \frac{\Pr(d \mid p) \cdot \Pr(p)}{\Pr(d)} \right\} \tag{2}$$

Debido a que sólo se busca conocer la probabilidad de que un documento tenga una polaridad y no obtener un puntaje específico para el nivel de polaridad que posee, el denominador de la ecuación 2 puede ser eliminado. Esto junto con el hecho de que uno de los supuestos del clasificador de Naive-Bayes es la independencia condicional entre todas las polaridades, se puede decir que:

$$\Pr(d \mid p) = \prod_{i=1}^{m} \Pr(w_i \mid p) = \prod_{i=1}^{m} \frac{\#(w_i, p)}{\#(w_i)} \tag{3}$$

Con $\#(w_i, p)$ el número de veces que la palabra w_i se ha encontrado en documentos de polaridad p en el conjunto de entrenamiento y $\#(w_i)$ el número de veces que la palabra w_i aparece en este último. Para evitar que existan probabilidades 0, se realiza un proceso llamado "suavización de Laplace"que consiste en lo siguiente:

$$\Pr(d \mid p) = \prod_{i=1}^{m} \frac{\#(w_i, p) + 1}{\#(w_i) + m} \tag{4}$$

Con estas ecuaciones basta resolver el problema de maximización planteado para obtener las probabilidades de que cada documento posea una polaridad en particular.

En general, las características utilizadas por los algoritmos de aprendizaje supervisado se dividen en las siguientes categorías:

- *Frecuencia y presencia de términos*: Si bien el uso de frecuencia de aparición de términos, por ejemplo a través del modelo *tf-idf*, en la recuperación de la información siempre ha sido de mucha utilidad, en [16] se muestra que en el caso de la extracción de opiniones desde documentos la *presencia* de un término es más importante que la frecuencia con que este aparece.

- *Partes del discurso*: Los adjetivos han sido utilizados con frecuencia [11] en el uso de algoritmos de aprendizaje supervisado ya que existe una alta correlación entre la presencia de adjetivos en una oración y la subjetividad de esta.
- *Sintáxis*: En [12] se hace uso de la relación entre las palabras como características en algoritmos de aprendizaje supervisado.

Algoritmos de clasificación a través de aprendizaje no supervisado: En [19] se propone un algoritmo de aprendizaje no supervisado para la clasificación de polaridad de documentos que se compone de tres etapas:

1. Se extraen todas las frases con verbos o adjetivos, ya que tal como se menciona en [11], estas partes del discurso se han mostrado muy útiles a la hora de detectar opiniones en documentos. Sin embargo, a pesar de que un adjetivo por si solo puede demostrar subjetividad, puede que no exista la información suficiente para determinar la polaridad de la opinión. Debido a esto, este algoritmo trabaja con pares de palabras, una de ellas siendo un adjetivo y la otra una palabra contextual que facilita la determinación de la polaridad de la oración en cuestión. Estos pares de palabras son extraídos siempre y cuando, considerando las dos palabras y la que les sigue, correspondan a alguno de los patrones conocidos.
2. Se estima la polaridad de las frases extraídas, haciendo uso de la métrica de dependencia estadística entre términos llamada *pointwise mutual information* (PMI) que se presenta en la ecuación 5

$$PMI(w_1, w_2) = log_2 \left(\frac{\Pr(w_1 \wedge w_2)}{\Pr(w_1)\Pr(w_2)} \right) \tag{5}$$

   Luego, la polaridad de una frase puede ser calculada basándose en el nivel de asociación entre ella y las palabras de referencia *pobre* y *excelente* a través de la ecuación 6

$$oo(frase) = PMI(frase, "excelente") - PMI(frase, "pobre") \tag{6}$$

3. Finalmente, el algoritmo calcula la polaridad *oo* promedio de todas las frases en el documento y lo clasifica dependiendo de si el promedio es positivo o negativo.

Algoritmos basados en lexicones de opinión: Los algoritmos basados en lexicones de opinión son los algoritmos más sencillos y a su vez los que buscan ser de uso más general debido a que la información utilizada para determinar la polaridad de una opinión no está restringida a ningún dominio en particular. Un lexicón es un conjunto de palabras rotuladas con polaridad de sentimientos, es decir, cada palabra perteneciente al lexicón tiene asociado un puntaje de polaridad.

Estos algoritmos trabajan bajo la hipótesis de que una palabra es considerada la unidad elemental de una opinión y por lo tanto la polaridad de una opinión puede reconstruirse a partir de la polaridad de cada una de las palabras que la componen. Ejemplos de algoritmos que hacen uso de lexicones para determinar la polaridad de una opinión se pueden encontrar en [14, 16]. En relación al minado de opiniones desde documentos de microblogging, Kouloumpis et al. dan a conocer en [8] que los algoritmos de basados en lexicones pueden dar buenos resultados.

El lexicón utilizado por la plataforma de detección de tendencias es *SentiWordNet* el cual está disponible públicamente para ser usado en este tipo de aplicaciones de minado de opiniones.

Cada palabra presente en un lexicón tiene asociado un puntaje por cada polaridad positiva, negativa y objetiva que representan el aporte de esta palabra para la polaridad de una opinión. En el caso de *SentiWordNet* [14] se tiene que cada palabra tiene asociado sólo los puntajes de polaridad positiva w^p y negativa w^n, y además el puntaje de objetividad $w^o = 1 - w^p + w^n$.

Los algoritmos basados en lexicones de opinión hacen uso de las siguientes metodologías para reconstruir la polaridad de la opinión contenida en un documento a partir de sus palabras:

- **Conteo de palabras**: los puntajes de polaridad de un documento se obtiene a través de la fracción de palabras cuya que posee una polaridad predominante p. En este caso, una palabra será considerada de una polaridad p si su mayor puntaje es el de aquella polaridad.

- **Promedio de palabras**: En un algoritmo de promedio de palabras, el puntaje asociado a una polaridad p es el promedio de los valores de polaridad p de todas las palabras presentes en el documento.

A partir de estas metodologías básicas se pueden realizar diversas variaciones tales como: modificar los puntajes de cada palabra en base al conjunto de palabras que la rodean en el documento, hacer uso de las negaciones y la capitalización, y finalmente incorporar al puntaje la existencia de intensificadores y disminuidores de adjetivos.

2.3. Modelos de detección de Tendencias

Los modelos de detección de tendencias buscan modelar el comportamiento de los tópicos tanto desde el punto de la cobertura que este tenga en la Web, como también la percepción que los usuarios de esta tengan sobre él. Por esto, los modelos de detección de tendencias van un paso más allá que los modelos de *topic tracking*, buscando analizar también la percepción que tiene la sociedad del tópico en particular y en cómo ambas componentes se relacionan para convertir un tópico en una tendencia.

En los últimos años se han realizado variados acercamientos a la detección de tendencias en la Web. Aplicaciones enfocadas en el uso de *key-words* para extraer tendencias en Web-usage mining se presentan en [21], política [13], finanzas [17] y sistemas de recomendación [7]. Sin embargo, la contribución de este trabajo se asemeja más a metodologías genéricas que van más allá de un dominio en particular, como la presentada en [20], cuyo foco es principalmente el cómo construir una plataforma de detección de tendencias sobre una arquitectura de *cloud computing*, y no en como recuperar la información necesaria ni como decidir si es que un tópico discutido en un conjunto de documentos a lo largo del tiempo refleja una tendencia.

3. Detección de Tendencias en la Web

En la actualidad, el análisis de tendencias se ha abordado tradicionalmente a partir de encuestas, las cuales poseen un alto contenido de subjetividad, y puede conducir a errores significativos a la hora de representar los hechos que sucederán en el futuro. Estos errores pueden darse debido al contexto en que las encuestas son realizadas, las motivaciones que la gente tiene a la hora de responder y otros factores exógenos al instrumento en si.

Por otro lado, en las redes sociales, las opiniones consignadas por sus usuarios son una expresión neta de sus sentimientos. Al ser estas no obligadas ni apresuradas, es posible complementar los resultados obtenidos a través de las encuestas con un análisis de estas, reduciendo el ruido producido debido a los factores previamente mencionados.

Para comprobar la hipótesis mencionada en la primera sección de este artículo, se diseñó una plataforma de detección de tendencias que analiza la información presente en la Web en dos ejes: el primero se enfoca en el análisis de eventos, que viene dado por los documentos presentes en los sitios de noticias; y aquel que trata de los sentimientos que expresan los usuarios de las redes sociales sobre aquellos eventos. Cabe destacar que esta plataforma hace uso de

técnicas existentes de algoritmos de recuperación de la información y también modelos de tópicos y minado de opiniones para modelar cómo los tópicos se comportan a lo largo del tiempo en busca de identificar tendencias en la Web.

Inicialmente se describirá el enfoque utilizado para recuperar noticias y extraer qué tópicos están siendo discutidos en la blogosfera y en los sitios de noticias. Posteriormente se dará a conocer la metodología para extraer las opiniones a partir de los documentos publicados por los usuarios de las redes sociales, y finalmente la metodología utilizada para juntar ambos conjuntos de información con el fin de identificar tendencias en la Web.

3.1. Plataforma de Detección de Tendencias

Tal como se menciona en la sección de trabajo relacionado, los modelos de tendencias no sólo buscan detectar qué tópicos se discuten a lo largo del tiempo en un corpus de documentos, también tienen como objetivo analizar las reacciones sociales que provocan estos tópicos . En el caso de este trabajo de investigación, se acotan a las opiniones consignadas por los usuarios de redes sociales a lo largo del periodo de análisis. Así, la plataforma propuesta debe estar formada por dos pilares fundamentales: la detección de tópicos a lo largo del tiempo, y el análisis de dichos tópicos en las redes sociales a través del minado de las opiniones que les competen.

3.1.1. Minado de noticias

Se considera que una fuente de documentos presente en la web es un *feed* si cada elemento que esta contenga es desplegado de manera cronológica y pertenecen todos una misma temática. Si una fuente de documentos dispone de un punto de acceso donde se puedan recuperar cada uno de los documentos existentes en ella se dice que es un *feed sindicable*, un ejemplo de esto son todos aquellos sitios web que tienen la opción de suscribirse a su contenido a través de RSS.

Una limitante a considerar a la hora de trabajar con feeds sindicables es que el conjunto de documentos presentes cuando se accede a esta depende del tiempo. Esto implica que el conjunto de documentos $\{d_i^F\}_{i\in\mathbb{N}}$ que se obtienen al solicitar todos los documentos desde la fuente F se ve limitada por el momento t en el cual se realice esta petición. En este caso, se define $\{d_i^{F^t}\}_{i\in\mathbb{N}}$ como el conjunto de documentos recuperados desde una fuente F en un instante de tiempo t. Además, se define $\{F_i\}_{i\in\mathbb{N}}$ como el conjunto de *feeds* que recorrerá el módulo de recuperación de documentos a través de su *crawler* para alimentar el módulo de extracción de tópicos.

Para este proyecto, sólo se trabajará con feeds sindicables, por lo que, en

base a lo anterior es posible definir un algoritmo de recuperación de documentos a partir de una lista de fuentes $\{F_i\}_{i\in\mathbb{N}}$ sindicables (sean estas *RSS* o *Atom*) como se describe en el algoritmo 3.1:

Algoritmo 3.1: Recuperación de documentos

Data: $\{F_i\}_{i\in\mathbb{N}}, t$
Result: $\bigcup_i \{d_j^{F_i^t}\}_{j\in\mathbb{N}}$

```
1 documents := [];
2 for i ← 1 to ‖{F_i}_{i∈ℕ}‖ do
3 |   document ← retrieveDocument(F_i);
4 |   documents ← documents ⋃ document;
5 return documents;
```

Para extraer qué tópicos se discuten a lo largo del tiempo en la Web, se propone utilizar un enfoque basado en modelo de tópicos, debido a que permiten de manera directa obtener las *keywords* necesarias para posteriormente extraer las opiniones presentes en las redes sociales, y además, permiten monitorear la evolución de los tópicos a lo largo del tiempo. El hacer uso de técnicas de *topic tracking* o *topic detection* no es recomendable debido a las limitaciones que estos imponen para la posterior recolección de opiniones asociadas a cada tópico.

Una vez recuperados los documentos desde los sitios de noticias, se procede a utilizar el modelo LDA para recuperar qué tópicos se están tratando en ellos. Este modelo permite, dada una colección de documentos $\{d_i\}_{i=1...\mathbb{N}}$, obtener un conjunto de tópicos t asociados a documentos, los cuales están descrito por la probabilidad $P(topic = t|document = d)$ de que un documento d pertenezca al tópico t y además, para cada tupla (w, t) la probabilidad $P(topic = t|word = w)$ de que una palabra w describa al tópico t. Así, es posible obtener los tópicos que se tratan a lo largo del tiempo en los feeds que se están minando y las palabras que los describen para luego utilizar esta información con el fin de recuperar documentos opinados desde las redes sociales.

Para cada periodo t_i, se toman todos los documentos de los dos periodos anteriores t_{i-1}, t_{i-2} y se entrena un nuevo modelo LDA con estos. Luego, para los documentos del periodo t se realiza inferencia con el modelo LDA sobre estos para descubrir el modelo de tópicos subyacente en estos.

Una vez que se tengan los documentos de los periodos t_i, t_{i-1}, t_{i-2}, es posible enlazar dos tópicos T y T', con vectores de probabilidades de palabras $\vec{w}_T$ y $\vec{w}_{T'}$ través de una función de distancia de tópicos que se define como sigue:

$$d(T, T') = \sum_{w \in \vec{w}_T} \sum_{\vec{w}_{T'}} w_i - w_j \tag{7}$$

Y luego, dado toda dupla T y T' de tópicos, se enlazan sí y sólo si el resultado la función $d(T, T')$ está bajo un umbral ϕ que se define a la hora de comenzar el análisis.

3.1.2. Minado de Opiniones

Una vez que se han extraído los tópicos a partir de los documentos presentes en sitios de noticias, se procede a extraer las opiniones sobre cada uno de ellos en las redes sociales a través del uso de modelos de minado de opiniones. Con esto es posible es posible obtener un puntaje de opinión para cada tópico a lo largo del tiempo a partir de una colección de documentos. La metodología a utilizar es la siguiente:

Para definir qué documentos serán recuperados desde las redes sociales, se tiene el algoritmo 3.2, que dado un tópico T obtiene todos los n-gramas de largo n que caracterizan a ese tópico en particular en el periodo t. El parámetro N consiste en la cantidad de palabras que deben ser utilizadas para la obtención de los unigramas que describen al tópico, además, el método T.`words`(t, N) obtiene las N palabras más relevantes del tópico T en el periodo t.

Algoritmo 3.2: Método `generateQueries`

Data: T, t, n, N
Result: $\{query_i\}_{i \in \mathbb{N}}$

```
1 queries := [];
2 words = T.words(t, N);
3 forall p ∈ permutaciones(words, n) do
4     queries.append(p);
5 return queries;
```

Luego, para cada tópico T, se obtienen todas las queries que le correspondan, y se obtienen documentos opiniones en las redes sociales que se determinen utilizando sus APIs. En el caso particular de este experimento, sólo se trabajará con la red social de microblogging Twitter. Para cada documento, se procede a obtener su polaridad de la siguiente manera:

Algoritmo 3.3: Clasificación de documentos opinados

Data: $\{d_i\}_{i=1...\mathbb{N}}$
Result: $\{\vec{d_i}\}_{i\in\mathbb{N}}$
documents := [];
for $i \leftarrow 1$ **to** $\|\{d_i\}_{i=1...\mathbb{N}}\|$ **do**
    $\vec{d_i} \leftarrow$ `polaridad`(d_i);
    documents.append$(\vec{d_i})$;
return documents;

3.1.3. Visualización de Tendencias

Una vez obtenidas las noticias y las opiniones relacionadas a los tópicos en discusión, se procede a generar un gráfico como el presentado en la figura 1 que representa el comportamiento de cada tópico a lo largo del tiempo. En donde las barras corresponden a la cantidad de documentos en los cuales se hace mención del tópico para cada periodo de tiempo, y la opinión de los usuarios de las redes sociales a lo largo del tiempo con respecto a este se representa como una línea.

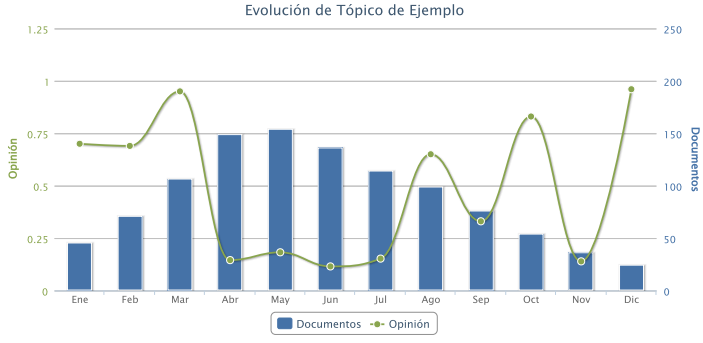

Figura 1: Ejemplo de gráfico por tópico

3.2. Diseño del experimento

Sobre una temática en particular, se visitará periódicamente un conjunto de 20 sitios de noticias que publiquen documentos sobre esta, y se ejecutará la metodología presentada a lo largo de un mes.

3.2.1. El entorno

Los sitios de noticias. En cuanto a los sitios, se requiere satisfacer tres requerimientos: en primer lugar, debe tener una frecuencia de publicación adecuada. Además, la cantidad de documentos por sitio no puede ser excesivo y, por último, estos deben estar en inglés para facilitar el procesamiento de los documentos.

El tema a analizar. El tema a analizar debe ser capaz de generar discusiones en las redes sociales, o al menos muestras de apreciación o desagrado, ya que de manera contraria no será posible realizar la última etapa del proceso de detección de tendencias y por lo tanto, el experimento se verá invalidado.

3.2.2. Captura y transformación de datos

Sitios de noticias. Una vez elegidos los sitios, estos serán visitados periódicamente para recuperar los artículos publicados en ellos. Cada artículo será almacenado con su contenido original, y para su procesamiento se procederá a remover todo el contenido que no sea texto plano (por ejemplo, *tags* de **html**) y todas las stopwords que se encuentren.

Opiniones. Al igual que los artículos recuperados de los sitios de noticias, estos serán almacenados tal como fueron extraídos desde su fuente.

3.3. Soluciones existentes para detección de tendencias

En el ámbito académico, múltiples investigaciones [10, 1, 5] han abordado la detección de tendencias en la web, principalmente en las redes sociales, destacándose entre ellas dos tipos distintos, aquellas que tienen como objetivo detectar de manera temprana aquellos tópicos que serán tendencia en el corto plazo, y las que buscan detectar aquellos tópicos que están siendo tendencia y su presencia va en aumento a lo largo del tiempo.

En aplicaciones comerciales, la plataforma web *NewsWhip*[1] ofrece prestaciones similares a las presentes en la plataforma de detección de tendencias presentada, sin embargo, su enfoque es lograr ser un agregador de noticias con características sociales, como la medición de menciones en las redes sociales de una noticia en particular o el análisis de noticias de una empresa en particular en la web. Además, *NewsWhip* ofrece la herramienta *Spike*, que permite a los generadores de contenido analizar cómo sus noticias se esparcen por la web.

La empresa *Sysomos*[2] se enfoca en monitorear las redes sociales en búsqueda de información relevante para una empresa en particular, sin embargo, no

[1] `http://www.newswhip.com/`
[2] `http://www.sysomos.com/`

hacen uso de la información presente en las noticias y no tienen como objetivo hacer un análisis extenso de las tendencias en la Web, si no monitorear las conversaciones que se están realizando en las redes sociales.

Otra iniciativa que busca detectar tendencias en la Web es Google Trends, la cual toma un enfoque distinto a los ya mencionados al analizar el comportamiento de búsqueda de los usuarios de su motor de búsqueda, sin embargo, no hacen uso de los datos presentes en su red social Google+ para complementar las tendencias obtenidas con información sobre las opiniones de la gente sobre ellas.

4. Aplicación del experimento y análisis de resultados

4.1. Captura de datos

Para recuperar los documentos existentes en los sitios de noticias o blogs que se analizaron, se implementó un *crawler* hecho en Java capaz de parsear y recuperar información desde fuentes *RSS*. Para cada fuente *RSS*, se solicita periódicamente la lista de artículos presente en ella, y en caso de que se encontraran nuevos elementos en relación a la última extracción de documentos se procede a almacenar esta diferencia en la base de datos. En el caso de los documentos opinados recuperados desde las redes sociales, también se desarrolló un *crawler* en Java para recuperar los documentos opinados asociados a un tópico en particular.

4.2. Aplicacion del Modelo de Detección de Tendencias

4.2.1. Entorno

La temática escogida: Los experimentos se desarrollaron con el fin de analizar lo sucedido en la temática de la tecnología y sus ramificaciones, en particular, se enfocó el estudio sobre noticias y opiniones en inglés. Ambas elecciones se realizaron en base a la alta cantidad de información disponible sin importar el periodo en el cual se realizara en el estudio.

Los sitios analizados: Se escogió de manera manual una muestra de 20 blogs o sitios de noticias en inglés que traten la temática de la tecnología. Cada uno de estos debe disponer de su contenido en formato RSS para una más fácil recuperación de sus artículos.

experimento	10	20	30
primero	66 %	58 %	49 %

Tabla 1: *Precision*

experimento	10	20	30
primero	37 %	46 %	59 %

Tabla 2: *Recall*

El periodo de análisis: Para el desarrollo del análisis se analizaron sitios de noticias entre Abril del 2011 y Enero del 2012, analizando los tópicos tratados por ellos en dicho periodo.

4.2.2. Experimentos

Como primer experimento, se aplicó la metodología presentada en el entorno previamente descrito, a partir del cual se procedió a analizar los tópicos extraídos y los gráficos temporales para cada uno de estos, y se determinó si la información presentada en ellos correspondía a lo que se podía observar a partir del análisis de los hechos ocurridos en este periodo. Por otro lado, el segundo experimento consistió en el análisis experto de estos gráficos para ver si dichos tópicos podían ser categorizados como tendencias.

4.3. Resultados Obtenidos

Luego de analizar los sitios de noticias previamente elegidos durante el periodo de análisis con un número variable de tópicos por periodo, y considerando una semana por iteración de la metodología, se encontró que la cantidad de tópicos extraídos por el modelo LDA en cada periodo que ofrecía mejores resultados correspondía a 10 y además, se determinó hacer uso de periodos de 7 días de largo.

4.3.1. *Precision y recall*

En la tabla 1 se muestra la precisión lograda en el primer experimento para las tres cantidades de tópicos por semana que fueron seleccionadas. Se puede observar que a medida que la cantidad de tópicos por periodo aumenta, la *Precision* del algoritmo disminuye, ya que a medida que esta variable aumenta, la granularidad del modelo LDA aumenta, provocando que un tópico descubierto por inspección sea dividido en dos tópicos más pequeños pero altamente relacionados. Este suceso ocurre en todo dominio que se quiera analizar, sin embargo, a medida que el dominio bajo análisis es más amplio, la cantidad

óptima de tópicos por periodo aumenta. Por lo tanto, es necesario ajustar el modelo dependiendo del dominio bajo análisis.

En la tabla 2, se observa que el *Recall* aumenta a medida que la cantidad de tópicos por periodo aumenta. Esto se debe a que si bien hay una mayor fragmentación de macrotópicos, se incluyen tópicos pequeños independientes que son absorbidos por ellos cuando la cantidad de tópicos por periodo disminuye.

En el caso del segundo experimento, se observó que un 63 % de los tópicos observados tuvieron un comportamiento similar al esperado por los expertos consultados, lo que indica, que a pesar de que la herramienta es propuesta como un apoyo a la detección de tendencias, esta realiza un buen trabajo en modelar el comportamiento de los tópicos a lo largo del tiempo.

5. Conclusiones

En este trabajo de investigación se demostró que es posible hacer uso de una herramienta de detección de tendencias basada en datos presentes en la web, para mejorar la calidad de la información provista por medios tradicionales de detección de tendencias como lo son las encuestas de opinión.

Para lograr este resultado se realizó un amplio estudio de cuáles de los datos originados en la web pueden complementar la información presente en los medios tradicionales, junto con los modelos matemáticos que se usan para describir tópicos en colecciones de documentos y la manera en que los usuarios de la web expresan sus opiniones en las redes sociales.

Si bien esta metodología es un complemento para los medios tradicionales, una de sus limitaciones es que la demografía de los usuarios de Internet, y aquellas personas accesibles a través de encuestas no siempre coinciden, por lo que si se desean realizar estudios enfocados en ciertos sectores de la población es posible que esta metodología no logre aportar suficiente valor. Por otro lado, al realizar el análisis de los datos de manera periódica, no es posible dar alerta temprana de sucesos que ocurren en el día a día. Por ello, los resultados entregados por esta herramienta deben ser considerados como un apoyo a decisiones de negocio enfocadas en un mercado en particular y también como un complemento a metodologías tradicionales de detección de tendencias.

Como trabajo futuro, se plantea considerar nuevas técnicas de minado de opiniones que se especialicen en documentos obtenidos desde sitios de microblogging y además características de estos como la ironía y los acrónimos de expresiones populares. Por otro lado, se plantea modificar el modelo de tópicos usado para que sea capaz de detectar reapariciones de tópicos después de

un tiempo prolongado. Finalmente, se propone la evaluación del impacto de implementar un sistema de alerta temprana.

Agradecimientos: Este trabajo fue parcialmente financiado por el Proyecto FONDEF project D10I- 1198: WHALE: Web Hypermedia Analysis Latent Environment y por el Instituto Milenio Sistemas Complejos de Ingeniería (ICM: P-05-004-F, CONICYT: FBO16).

Referencias

[1] F. Alvanaki, M. Sebastian, K. Ramamritham, and G. Weikum. Enblogue: emergent topic detection in web 2.0 streams. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 1271–1274, Athens, Greece, 2011. ACM.

[2] Nikolay Archak, Anindya Ghose, and Panagiotis G. Ipeirotis. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 56–65, San Jose, California, USA, 2007. ACM.

[3] David M. Blei and John D. Lafferty. Correlated topic models. In *NIPS*, 2005.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

[5] Irena Pletikosa Cvijikj and Florian Michahelles. Monitoring trends on facebook. In *Proceedings of the 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, DASC '11, pages 895–902, Sydney, Australia, 2011. IEEE Computer Society.

[6] T Damer. *Attacking faulty reasoning: a practical guide to fallacy-free arguments*. Wadsworth/Cengage Laerning, Australia Belmont, CA, 2009.

[7] Shay David and Trevor John Pinch. Six degrees of reputation: The use and abuse of online review and recommendation systems. *First Monday*, July 2006. Special Issue on Commercial Applications of the Internet.

[8] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis : The good the bad and the omg ! *Artificial Intelligence*, 70(2):538–541, 2011.

[9] Bing Liu. Sentiment analysis and subjectivity. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, Florida, USA, 2010. ISBN 978-1420085921.

[10] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, Indianapolis, Indiana, USA, 2010. ACM.

[11] T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, volume 4 of *EMNLP '04*, pages 412–418. ACL, 2004.

[12] V. Ng, S. Dasgupta, and SM Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 611–618. Association for Computational Linguistics, 2006.

[13] B. O'Connor, R. Balasubramanyan, B.R. Routledge, and N.A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, ICWSM '10, pages 122–129. AAAI Press, 2010.

[14] Bruno Ohana and Brendan Tierney. Sentiment classification of reviews using sentiwordnet. *Discovery*, page 13, 2009.

[15] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January 2008.

[16] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[17] V. Sehgal and C. Song. Sops: stock prediction using web sentiment. In *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*, ICDMW '07, pages 21–26, Omaha, Nebraska, USA, 2007. IEEE Computer Society.

[18] Edison M. Taylor, Cristián Rodriguez, Juan D. Velásquez, Goldina Ghosh, and Soumya Banerjee. Web opinion mining and sentiment analysis. In

Juan D. Velásquez, Vasile Palade, and Lakhmi C. Jain, editors, *Advanced Techniques in Web Intelligence-2*, pages 105–126. Springer, 2012.

[19] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Philadelphia, Pennsylvania, 2002. Association for Computational Linguistics.

[20] Athena Vakali, Maria Giatsoglou, and Stefanos Antaris. Social networking trends and dynamics detection via a cloud-based framework design. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 1213–1220, New York, NY, USA, 2012. ACM.

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

Planificación del menú semanal de colaciones de un hospital de Argentina por medio de programación lineal entera

Sebastián Guala*
Javier Marenco*

Resumen

La programación de colaciones en centros de salud siempre ha sido un área compleja que involucra varios factores: estándares alimentarios, variedad, costos y aspectos culturales. En este trabajo presentamos el desarrollo de un modelo de programación entera para la planificación de un menú semanal de colaciones para un hospital de Argentina. En este contexto, la programación del menú contempla los almuerzos y las cenas, y cada colación se compone de entrada, plato de fondo más acompañamiento y postre. El objetivo de la planificación es proponer un menú semanal que minimice los costos respetando las exigencias de variedad de platos que se ajusten al gusto local. Dentro del marco de los estándares alimentarios se plantearon dos escenarios: uno restringido a los estándares recomendados por organismos internacionales y otro más cercano a las características culturales de los pacientes. Los resultados muestran mejoras de un 21 % a un 25 % comparados con los costos obtenidos por los métodos manuales utilizados actualmente.

Palabras Clave: Problema de la Dieta, Programación Lineal Entera

*Instituto de Ciencias, Universidad Nacional de General Sarmiento, Buenos Aires, Argentina.

1. Introducción

La alimentación en centros de salud es un área compleja y delicada por sus implicancias en los tiempos de recuperación de los pacientes y en el bienestar tanto de pacientes como del personal médico y administrativo [1]. En este contexto, es necesaria una correcta programación de los menús y planes de alimentación, que garantice una dieta equilibrada y con un impacto positivo en la salud de los pacientes. Determinar el menú semanal o mensual que debe ser elaborado implica una tarea compleja debido al gran número de variables que deben ser tenidas en cuenta, entre las cuales se encuentran los requerimientos nutricionales máximos y mínimos recomendados (hidratos de carbono, proteinas, grasas, colesterol, calcio, hierro, etc.), las influencias culturales en las características gastronómicas de los platos e ingredientes utilizados (tipos y preparaciones "culturalmente aceptadas" por los pacientes), la variedad de platos en lo referente a ingredientes y presentación, y los costos del menú.

Existen destacados trabajos relacionados con estos aspectos, en los que se desarrollan modelos de planificación alimentaria que satisfacen los requerimientos nutricionales diarios al menor costo [1, 2, 3, 4, 10, 13]. Sin embargo, la planificación propuesta en estos modelos es realizada al nivel de los ingredientes disponibles, sin considerar platos y preparados. En otras palabras, combinando alimentos en general tales como leche, cereales, carnes, verduras, etc. obtienen una propuesta que satisface los requerimientos diarios, pero sin considerar las formas en que esos alimentos son combinados o presentados al paciente. En consecuencia, no se obtiene en estos trabajos una propuesta gastronómica factible, sino que se obtiene una cota inferior ideal sin contemplar las preferencias del paciente.

El objetivo de este trabajo es desarrollar un modelo de programación entera que permita programar menús semanales de almuerzos y cenas donde se indiquen los platos preparados dentro de un conjunto de platos propuestos, de acuerdo al tipo y cantidad de cada ingrediente y sus cualidades nutricionales. El modelo es aplicable a servicios de alimentación en general, pero este desarrollo fue motivado a propuesta de los requerimientos específicos de un hospital de la Provincia de Buenos Aires, Argentina. Esta programación debe tener en cuenta los estándares alimentarios establecidos y demás consideraciones culturales y de variedad minimizando los costos del menú. Con este modelo se pretende optimizar el costo semanal de alimentación del personal y de los pacientes internados que no tengan restricciones alimentarias (como los pacientes

diabéticos, cardíacos y celíacos, entre otros, que reciben dietas especiales).

El presente trabajo está organizado de la siguiente manera: en la Sección 2 se define el problema, los tipos de platos que componen cada colación y una descripción general de las restricciones que deben incluirse en el modelo. En la Sección 3 se formula el modelo de programación entera. La Sección 4 presenta las características de la implementación del modelo y se muestran los resultados obtenidos según se sigan las recomendaciones de organizaciones internacionales o se flexibilicen levemente ciertos parámetros nutricionales para adaptarse cuantitativamente al gusto local. Finalmente, la Sección 5 presenta las conclusiones del trabajo.

2. Definición del problema

Se deben planificar las combinaciones de platos a incluir en el almuerzo y la cena durante un período de 7 días. Tanto los funcionarios del hospital (médicos, enfermeros y personal administrativo) como los pacientes internados que no requieren una dieta especial reciben el mismo menú para el almuerzo y la cena. Salvo los pacientes, los turnos laborales hacen que los funcionarios que almuerzan no sean los mismos que los que cenan en el hospital. Ambas colaciones deben estar compuestas por una serie de platos ordenados en "entrada", "plato de fondo", "acompañamiento" y "postre". El plato de fondo y el acompañamiento pueden ser reemplazados por un solo plato que no lleve acompañamiento (por ejemplo, pastas) y que a los efectos de este trabajo llamaremos "plato contundente". El menú se genera sobre la base de un listado de platos prefijados –acompañados por los datos de ingredientes necesarios, sus cantidades y costo– y cada ingrediente a su vez está acompañado por su información nutricional. Cada plato pertenece a sólo uno de los tipos posibles.

La planificación debe respetar las siguientes restricciones:

- Cada colación debe estar compuesta por exactamente una entrada, un plato de fondo con acompañamiento y un postre. Se puede reemplazar el plato de fondo y el acompañamiento por un solo plato contundente.

- La entrada, acompañamiento y postre pueden aparecer en el menú semanal a lo más dos veces y, si aparecen dos veces, deben pasar a lo menos dos colaciones entre la primera y la segunda aparición del mismo plato. Esto evita que dentro de la estadía media de internación un paciente repita estos platos. A su vez, dado que el personal no es el mismo en el almuerzo que en la cena, cada turno no repetiría estos platos antes de dos días.

- No se pueden repetir platos de fondo ni platos contundentes dentro de las 14 colaciones planificadas. Esta condición fuerza una mayor variedad de platos, priorizando la variedad de platos de fondo y platos contundentes por sobre los demás (que pueden aparecer hasta dos veces).

- El aporte nutricional total de los platos servidos por día debe encontrarse dentro de los valores diarios máximos y mínimos recomendados. Al no considerar el desayuno y la once, a los requerimientos nutricionales diarios recomendados se les descuentan los valores típicos aportados por el desayuno y la once.

En lo respectivo a la variedad de platos y exigencias nutricionales, estas restricciones en principio son suficientes para obtener un menú acorde con los estándares nutricionales. Sin embargo, los funcionarios del hospital a cargo de la planificación del menú solicitaron ciertas exigencias cualitativas asociadas a una cultura de alimentación saludable en general, que exceden el mero cumplimiento de valores nutricionales:

- A lo menos uno de los dos postres de cada día debe ser una fruta o contener frutas frescas.

- A lo más un plato al día debe estar compuesto por pastas, arroz o polenta. Estos platos se encuentran dentro de la categoría "contundente" puesto que no llevan acompañamiento. Adicionalmente, si se sirve uno de estos platos contundentes en una de las colaciones, la entrada debe ser a base de verduras.

Finalmente, existe otro grupo de condiciones que fueron establecidas para satisfacer las costumbres gastronómicas locales. Se consideraron dos escenarios, según el menú se ajuste estrictamente a los valores nutricionales máximos y mínimos diarios recomendados por la Organización Mundial de la Salud (OMS) [5, 6, 14] y otras organizaciones de referencia internacional [7, 8, 9] o a una versión más "tradicional" que permita que los valores máximos de ciertos parámetros nutricionales puedan ser moderadamente excedidos y que determinados platos sean evitados, con el fin de adaptarse al gusto local. Las proteínas y el colesterol, asociados principalmente al consumo de carne están entre los parámetros excedidos, mientras que el pescado forma parte de los platos evitados:

- No se sirve pescado en el almuerzo y a lo más un plato con pescado en la cena. Esto obedece a la poca cultura gastronómica basada en pescados y mariscos en la región pampeana de Argentina y especialmente en la Provincia de Buenos Aires, donde se aplicó el modelo. Particularmente

se evitó el pescado en el almuerzo debido a las repetidas quejas de los funcionarios del turno diurno.

- En el almuerzo debe haber un plato con no menos de 100 gramos de carne de ave, ovina, bovina o porcina (excluido pescado), y en la cena a lo más un plato con 40 o más gramos de carne (incluido pescado entre las posibilidades). Esto obligó a que se incrementaran los límites máximos recomendados de consumo diario de proteínas y de colesterol dentro de valores considerados seguros, dado que la incorporación de esta condición podría llevar a que no hubiera soluciones factibles para las recomendaciones de la OMS, que sugiere que el consumo diario de carnes debe estar alrededor de 80 gramos.

En resumen, las restricciones impuestas a la planificación se dividen en tres grupos: el primero es el grupo básico que caracteriza el tipo de servicio y los parámetros nutricionales internacionalmente recomendados, el segundo grupo aporta buenas prácticas alimentarias generales no contempladas explícitamente en los parámetros nutricionales y el tercero ajusta el menú resultante a la gastronomía local. A partir de la omisión o aplicación de este tercer grupo de restricciones se generan dos propuestas alternativas: una propuesta considera un menú estrictamente ajustado a las recomendaciones de la OMS y otra propuesta refleja un menú más cercano al consumo tradicional de la región, lo que implica algunas "licencias" alimentarias.

Vale aclarar que el conjunto total de platos disponibles para confeccionar el menú ya puede interpretarse como parte de una cultura gastronómica local, lo que significa que cualquier combinación de estos platos dará como resultado un menú socialmente aceptable desde el punto de vista cualitativo. Sin embargo, desde el punto de vista de la *etno-gastronomía*, el último grupo de condiciones le da una perspectiva localista al consumo de proteinas de origen animal.

3. Desarrollo del modelo

Presentamos en esta sección un modelo de programación lineal entera para el problema descrito en la sección anterior. Como parte de la definición se explicitará el manejo que se realiza de los datos de ingredientes, grupos de ingredientes y aportes nutricionales, que se describió en términos cualitativos en la sección anterior. Llamamos *ingredientes* a los elementos que el hospital debe comprar para preparar los distintos platos (por ejemplo, algunos ingredientes son manzanas, leche, tomates, carne de vaca, etc.). Cada ingrediente realiza un aporte de distintos *atributos nutricionales* (especificados en el Cuadro 1).

Atributo (t)	$\min_t$	$\max_t$ (OMS)	$\max_t$ (local)
Hidratos de Carbono (g)	200	400	400
Proteínas (g)	30	70	80
Grasas (g)	30	70	70
Calorías (Kcal)	1000	2500	2500
Sodio (mg)	300	2000	2000
Colesterol (mg)	0	300	400
Hierro (mg)	2	N/A	N/A
Calcio (mg)	400	N/A	N/A
Fibra (g)	7	N/A	N/A
Fósforo (mg)	300	N/A	N/A
Potasio (mg)	1000	N/A	N/A

Tabla 1: Listado de atributos nutricionales, con sus requerimientos diarios mínimos y máximos para ambos escenarios. Las entradas "N/A" indican que en una dieta regular no hay riesgos asociados a un consumo excesivo (siempre dentro de valores razonables).

Los ingredientes se combinan entre sí para formar *platos* (por ejemplo, carne al horno, pollo al verdeo, tallarines con crema, etc.).

Para la formulación del modelo, contemplamos los siguientes conjuntos:

- I: conjunto de *ingredientes*. Utilizamos habitualmente el índice $k \in I$ para referirnos a los ingredientes, y asumimos que este conjunto está particionado en $I = \text{Verdura} \cup \text{Fruta} \cup \text{Carne} \cup \text{Pescado} \cup \text{Pollo} \cup \text{Harinas} \cup \text{Otros}$.

- P: conjunto de *platos disponibles*. Utilizamos habitualmente el índice $i \in P$ para referirnos a los platos, y asumimos que este conjunto está particionado en $P = \text{Entrada} \cup \text{PFondo} \cup \text{Acomp} \cup \text{Cont} \cup \text{Post}$, del siguiente modo:

  1. Entrada: Entradas,
  2. PFondo: Platos de fondo que llevan acompañamiento,
  3. Acomp: Acompañamientos,
  4. Cont: Platos contundentes (platos de fondo que no llevan acompañamiento),
  5. Post: Postres.

- $J = \{1, \ldots, 2n\}$: conjunto de *colaciones* a lo largo de los n días del horizonte de planificación. Para $d = 1, \ldots, n$, el almuerzo del d-ésimo

día es la colación $2d-1$, y la cena del d-ésimo día es la colación $2d$. Utilizamos habitualmente el índice $j \in J$ para referirnos a las colaciones. De acuerdo con esta definición, los almuerzos tienen índice impar y las cenas tienen índice par.

- T: conjunto de *atributos nutricionales* dado por el Cuadro 1. Utilizamos habitualmente el índice $t \in T$ para referirnos a los elementos de este conjunto.

Además, asumimos como datos de entrada los dados por los siguientes parámetros:

- bp_k: Cantidad base para el cálculo proporcional de los atributos del ingrediente k, prácticamente todos los ingredientes tienen $\text{bp}_k = 100$ gramos.
- prop_{kt}: Cantidad del atributo t (en gramos o miligramos) por cada bp_k unidades del ingrediente k.
- min_t: Consumo mínimo diario recomendado del atributo t (expresado en las unidades correspondientes).
- max_t: Consumo máximo diario recomendado del atributo t (expresado en las unidades correspondientes).
- precio_k: Precio unitario del ingrediente k (expresado en pesos por kilogramo).
- bruto_{ik}: Peso bruto comprado del ingrediente k destinado a cada plato i (expresado en gramos). Esta cantidad representa cuánto lleva el plato con el ingrediente tal y como se compra: con cáscaras, semillas, piel o huesos dependiendo del ingrediente, que debe quitarse para cocinar pero que forma parte del peso total y, por lo tanto, del costo del plato
- neto_{ik}: Peso neto utilizado del ingrediente k en cada plato i (expresado en gramos). Esta cantidad representa el peso del ingrediente que efectivamente se sirve en el plato y por lo tanto es la fracción que debe ser considerada para el cálculo nutricional.

Para cada plato $i \in P$ y cada colación $j \in J$, introducimos la variable binaria x_{ij}, de modo tal que $x_{ij} = 1$ si el plato i se sirve en la colación j, y $x_{ij} = 0$ en caso contrario. Con estas definiciones, el modelo se puede formular del siguiente modo:

1. La función objetivo solicita minimizar el costo total:

$$\text{mín} \sum_{i \in P} \sum_{j \in J} x_{ij} \Big(\sum_{k \in I} \text{precio}_k \, \text{bruto}_{ik} / 10^3 \Big).$$

2. Exactamente una entrada por colación:

$$\sum_{i \in \text{Entrada}} x_{ij} = 1, \qquad \forall j \in J.$$

3. O bien un plato contundente o bien un plato de fondo y acompañamiento por colación:

$$\sum_{i \in \text{Cont}} 2x_{ij} + \sum_{i \in \text{PFondo}} x_{ij} + \sum_{i \in \text{Acomp}} x_{ij} = 2, \quad \forall j \in J.$$

4. A lo más un plato de fondo por colación:

$$\sum_{i \in \text{PFondo}} x_{ij} \leq 1, \qquad \forall j \in J.$$

5. A lo más un acompañamiento por colación:

$$\sum_{i \in \text{Acomp}} x_{ij} \leq 1, \qquad \forall j \in J.$$

6. Exactamente un postre por colación:

$$\sum_{i \in \text{Post}} x_{ij} = 1, \qquad \forall j \in J.$$

7. No se pueden repetir platos de fondo ni platos contundentes:

$$\sum_{j \in J} x_{ij} \leq 1, \qquad \forall i \in \text{PFondo} \cup \text{Cont}.$$

8. Los postres, entradas y acompañamientos se repiten a lo más dos veces durante las $2n$ colaciones:

$$\sum_{j \in J} x_{ij} \leq 2, \qquad \forall i \in \text{Entrada} \cup \text{Acomp} \cup \text{Post}.$$

9. Los postres, entradas y acompañamientos no se repiten en una ventana de tres colaciones:

$$x_{ij} + x_{i,j+1} + x_{i,j+2} \leq 1, \qquad \forall j \in J, j < 2n-1,\ i \in \text{Entrada} \cup \text{Acomp} \cup \text{Post}.$$

10. Se deben respetar los consumos mínimos y máximos de cada atributo nutricional:

$$\min_t \leq \sum_{i \in P} (x_{ij} + x_{ij+1}) \Big(\sum_{k \in T} \frac{\text{prop}_{kt}\, \text{neto}_{ik}}{\text{bp}_k} \Big) \leq \max_t,$$

$$\forall t \in T,\ j \in J, j \text{ impar.}$$

11. Se debe tener exactamente un plato con carne en el almuerzo. Definimos el conjunto C de los *platos con carne* como los platos cuyos ingredientes de carne aportan al menos 100 gramos al total, es decir $CA = \{i \in P : \sum_{k \in \text{CarneUPollo}} \text{neto}_{ik} \geq 100\}$. Es importante observar que el conjunto Carne incluye platos con carne ovina, bovina y porcina. Esta es la restricción más significativa de las impuestas para que el menú se adapte al consumo local de carnes, lo que obligó a elevar el margen superior del consumo de proteínas y colesterol, aunque manteniéndose dentro de parámetros de seguridad:

$$\sum_{i \in CA} x_{ij} = 1, \qquad \forall j \in J, j \text{ impar.}$$

12. Se debe tener a lo sumo un plato con carne en la cena. En este caso, alcanza con que los ingredientes aporten 40 gramos o más de carne para que un plato ingrese a esta categoría, y definimos el conjunto $CN = \{i \in P : \sum_{k \in \text{CarneUPolloUPescado}} \text{neto}_{ik} \geq 40\}$:

$$\sum_{i \in CN} x_{ij} \leq 1, \qquad \forall j \in J, j \text{ par.}$$

13. No se debe incluir pescado en el almuerzo. Para esto, definimos el conjunto PP de los platos con pescado como $PP = \{i \in P : \sum_{k \in \text{Pescado}} \text{neto}_{ik} > 0\}$:

$$\sum_{i \in PP} x_{ij} = 0, \qquad \forall j \in J, j \text{ impar.}$$

14. A lo más un plato de pescado en cada cena:

$$\sum_{i \in PP} x_{ij} \leq 1, \qquad \forall j \in J, j \text{ par.}$$

15. No puede haber solamente verdura en la cena. Definimos el conjunto de platos que no tienen exclusivamente verdura como $PV = \{i \in P \setminus \text{Post} : \sum_{k \notin \text{Verdura}} \text{neto}_{ik} > 0\}$:

$$\sum_{i \in PV} x_{ij} \geq 1, \qquad \forall j \in J, j \text{ par.}$$

16. No puede haber pastas ni polenta ni arroz sin verduras (al menos 50 gramos en la entrada). Para esto, definimos el conjunto EV de entradas con verdura como $EV = \{i \in \text{Entrada} : \sum_{k \in \text{Verdura}} \text{neto}_{ik} > 50\}$, y definimos el conjunto CH de platos contundentes con harina como $CH = \{i \in \text{Cont} : \sum_{k \in \text{Harinas}} \text{neto}_{ik} > 0\}$:

$$\sum_{i \in EV} x_{ij} \geq \sum_{i \in CH} x_{ij}, \qquad \forall j \in J, j \text{ par.}$$

17. A lo más un plato con pasta, polenta o arroz (igual o más de 20 gramos de harinas) por colación. Definimos el conjunto PH de platos con harina como $PH = \{i \in P : \sum_{k \in \text{Harinas}} \text{neto}_{ik} \geq 20\}$:

$$\sum_{i \in PH} x_{ij} \leq 1, \qquad \forall j \in J.$$

18. No menos de una fruta por día de postre. Definimos el conjunto PF de postres con fruta como $PF = \{i \in \text{Post} : \sum_{k \in \text{Fruta}} \text{neto}_{ik} > 0\}$:

$$\sum_{i \in PF} x_{ij} + x_{ij+1} \geq 1, \qquad \forall j \in J, j \text{ impar.}$$

19. Naturaleza de las variables:

$$x_{ij} \in \{0, 1\}, \qquad \forall i \in P, j \in J.$$

4. Resultados

En esta sección se presenta un resumen de los resultados obtenidos para el caso de los dos escenarios de menús planteados (siguiendo los requerimientos de la OMS y siguiendo un criterio más localista) y una comparación con la situación actual. Para tener en cuenta exclusivamente los requerimientos de la OMS, se eliminan las restricciones (12) a (16). Para la ejecución se utilizó un total de 63 platos clasificados en 13 entradas, 10 platos de fondo, 9 acompañamientos, 14 platos contundentes y 17 postres. A su vez, los platos están compuestos por un total de 75 ingredientes clasificados por tipo y clase. Todos los datos utilizados fueron aportados por el personal encargado del servicio de colaciones del hospital.

Para estos datos, el modelo de programación lineal entera está compuesto por 882 variables binarias y 838 restricciones (incluyendo las restricciones localistas). El modelo se codificó en el lenguaje de modelado ZIMPL [11] y se resuelve por medio del paquete SCIP 3.0.0 [12]. A modo ilustrativo, la resolución de este modelo en un computador con procesador Intel Core 2 Duo con procesadores de 2 GHz y 4 GB de memoria RAM finaliza luego de 3 minutos con un 8 % de gap de optimalidad, valores que son aceptables para los usuarios. Se implementó una aplicación en Java para el manejo de los datos, la resolución del modelo y la visualización adecuada de los resultados (ver Figura 1), con la intención de que esta herramienta pueda ser utilizada por personal no especializado en investigación de operaciones. Se utiliza tecnología Java para la interfaz y SCIP para la resolución del modelo dado que se espera distribuir la aplicación en distintos hospitales, y no es factible utilizar software comercial que involucre licencias de alto costo en estas instalaciones.

El Cuadro 2 contiene el menú obtenido a partir de los platos disponibles y de los costos y atributos nutricionales de sus ingredientes, ajustado a las recomendaciones alimentarias diarias de la OMS. Como se mencionó, este menú no es una formulación "abstracta" dado que el hecho de planificar a partir de una lista de platos preestablecidos ya implica una primera adaptación al gusto local dado por la disponibilidad de los ingredientes utilizados y por la forma de preparar los platos.

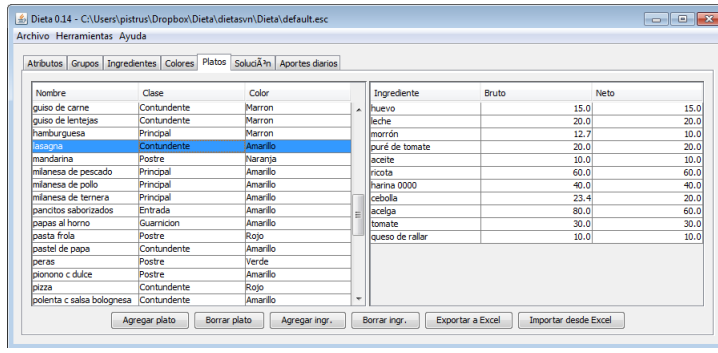

Figura 1: Interfaz principal de la aplicación computacional implementada para manejar los datos y visualizar los resultados.

Día	Almuerzo	Cena
1	sopa de verdura tallarines con salsa de pollo pionono con dulce	empanada de verdura milanesa de ternera ensalada jardinera compota de manzana
2	salchichas envueltas hamburguesa papas al horno mandarina	ensalada primavera pizza gelatina bicolor
3	croquetas de arroz pollo a la portuguesa puré de papa compota de manzana	sopa de verdura tallarines con crema de verdeo flan con vainillas
4	empanada de verdura arroz con pollo pionono con dulce	budín de anco fuccile con salsa de pollo duraznos
5	revuelto de zapallitos tallarines con salsa bolognesa vigilante	salchichas envueltas tarta de pollo ensalada jardinera gelatina con frutas
6	pancitos saborizados pollo al verdeo papas al horno mandarina	budín de anco lasagna vigilante
7	ensalada primavera polenta con salsa bolognesa duraznos con crema	pancitos saborizados milanesa de pollo puré de papa flan

Tabla 2: Planificación de menúes obtenida por el modelo respetando los estándares nutricionales solicitados por la OMS.

Día	1	2	3	4	5	6	7
Costo	11.16	9.32	9.11	9.04	9.54	10.16	9.92
Hidratos	206.48	201.52	201.03	209.77	226.20	200.52	202.45
Proteinas	62.30	48.66	41.98	56.1	50.26	51.14	57.71
Grasas	33.75	44.30	32.05	30.75	47.35	35.84	30.69
Calorías	1379.35	1399.45	1260.45	1340.75	1532.00	1329.16	1316.81
Sodio	594.78	751.68	335.55	555.83	787.50	436.13	474.13
Colesterol	288.00	292.30	299.30	284.0	295.30	294.20	298.80
Hierro	15.24	11.92	23.60	13.89	9.98	9.69	11.15
Calcio	430.00	538.64	451.25	405.2	426.64	515.49	479.79
Fibra	19.67	13.93	19.19	18.44	15.88	16.38	17.31
Fósforo	934.68	747.74	772.60	761.13	950.96	951.89	957.64
Potasio	2701.05	2098.05	2405.20	1924.6	2220.15	2373.53	2799.68

Tabla 3: Aportes nutricionales diarios de la solución del Cuadro 2.

El Cuadro 3 muestra los aportes diarios del menú propuesto en el Cuadro 2. Se puede ver que se ha obtenido una combinación de platos que logra un balance nutricional satisfactorio ya que mantienen un aporte equilibrado que, en general, no va de un extremo al otro de los rangos de recomendaciones nutricionales. Este hecho era uno de los puntos a evaluar debido a que los modelos previos citados en las referencias resuelven problemas con variables continuas asociadas con los ingredientes, teniendo más flexibilidad para el cumplimiento de las restricciones nutricionales (aunque –como se mencionó– con el inconveniente de no generar menús atractivos).

De igual manera, el Cuadro 4 y el Cuadro 5 muestran el menú obtenido y sus aportes diarios ajustados a las costumbres locales, respectivamente. Este menú evidencia el aumento efectivo de platos con alto contenido de carnes, lo que se refleja en el mayor nivel de proteínas y colesterol diarios consumidos en comparación con el Cuadro 3. Sin embargo, el mayor consumo de proteinas de origen animal y de colesterol se mantiene dentro de niveles considerados seguros para adultos que no tengan prescripciones alimentarias específicas. El criterio de consumo que es considerado seguro se valora a partir de la apreciación de los especialistas puesto que no hay valores exactos sino que se basan en las experiencias de la especialidad.

Desde el punto de vista de los costos, el Cuadro 6 muestra el mejor uso de los recursos mediante la planificación del menú con las herramientas propuestas en este trabajo en comparación con la planificación manual. Como era de esperar, la versión con las recomendaciones de la OMS logra menores costos al tener condiciones menos restrictivas respecto del consumo de carnes que la versión localista. No obstante, es importante tener en cuenta que el menú actual se elabora considerando la versión local del consumo de carnes.

Día	Almuerzo	Cena
1	rollitos de jamón y queso milanesa de pollo ensalada jardinera pasta frola	tortilla de verdura tallarines con salsa bolognesa mandarina
2	croquetas de arroz carne al horno papas al horno flan con vainillas	ensalada primavera tallarines con crema de verdeo gelatina con frutas
3	pancitos saborizados hamburguesa puré de papa flan	budín de anco polenta con salsa bolognesa duraznos con crema
4	croquetas de arroz pollo al verdeo ensalada de tomate pionono con dulce	ensalada primavera pizza gelatina con frutas
5	empanada de jamón y queso pollo a la portuguesa papas al horno pasta frola	sopa de verdura lasagna compota de manzana
6	empanada de verdura pastel de papa mandarina	tortilla de verdura tallarines con salsa de pollo vigilante
7	budín de anco milanesa de ternera ensalada jardinera compota de manzana	pancitos saborizados tarta de pollo puré de papa flan con vainillas

Tabla 4: Planificación de menúes obtenida por el modelo ajustado a las preferencias locales.

Día	1	2	3	4	5	6	7
Costo	10,56	10,72	9,42	8,18	11,78	11,48	8,93
Hidratos	235,95	211,48	211,05	205,35	219,43	201,68	200,13
Proteinas	65,15	72,27	49,41	46,70	55,75	62,00	57,51
Grasas	43,50	35,85	32,29	34,60	62,65	32,05	36,44
Calorías	1595,90	1257,65	1332,41	1320,10	1664,55	1343,15	1358,46
Sodio	1067,50	314,30	366,38	517,75	1147,27	549,58	470,10
Colesterol	388,40	388,90	351,49	380,40	392,20	387,30	359,90
Hierro	12,41	25,37	11,05	10,26	11,16	15,08	25,45
Calcio	413,15	442,58	463,20	575,63	483,15	448,00	464,80
Fibra	15,38	15,60	16,63	11,35	16,94	18,42	22,23
Fósforo	976,70	777,25	894,69	660,35	937,08	1031,08	1039,17
Potasio	2442,00	2199,05	2646,83	1417,55	2395,05	2773,05	3142,58

Tabla 5: Aportes nutricionales diarios de la solución del Cuadro 4.

Criterio	Costo semanal	Diferencia	Mejora
Internacional(OMS)	\$68,24	\$22,21	24,56 %
Local	\$71,07	\$19,38	21,43 %

Tabla 6: Costos semanales por paciente de los menús obtenidos con respecto al costo de \$90,45 del menú confeccionado manualmente.

5. Conclusiones

El modelo desarrollado permite generar propuestas gastronómicas factibles y aplicables para servicios de colaciones. Este modelo permite una mejora de los costos de entre un 21 % y un 25 % sobre la situación presente del hospital que presentó la inquietud. Estas mejoras dependen, respectivamente, según se trate del modelo ajustado a la versión local o a la OMS. Sin embargo, es importante tener en cuenta en la comparación que la planificación actual se hace sobre la base de los patrones de consumo local, lo que lleva la mejora más cerca del 21 %.

Con relación a los tiempos de ejecución, cabe señalar que el tiempo que el personal encargado de la gestión del servicio de colaciones debe dedicarle actualmente a la planificación manual del menú semanal, teniendo en cuenta la complejidad del caso, es del orden de varias horas para obtener una planificación razonable. En este sentido, la disminución de los tiempos de procesamiento al utilizar la herramienta propuesta en este trabajo permite que los funcionarios encargados de esta tarea puedan analizar múltiples escenarios (que además son óptimos para el costo y no solamente "razonables"). Además, la herramienta permite reaccionar rápidamente ante cambios en los precios de los ingredientes.

El modelo presentado está siendo implementado actualmente para la planificación de las colaciones del hospital como menú base. Los convincentes resultados han animado a los funcionarios responsables de la planificación de los menús del hospital a solicitar nuevas características sobre la solución, que se comentan a continuación.

La revisión más importante y compleja consiste en incorporar la combinación de los colores de los platos que se sirven en una colación como una nueva característica a tener en cuenta. Una posible restricción puede ser que cada colación contenga platos de a lo menos tres colores distintos. Esta propiedad no afecta la calidad nutricional de la colación, pero puede mejorar la experiencia gastronómica del paciente al recibir una combinación de platos más colorida.

El agregado de estas restricciones implica la incorporación de un nuevo juego de variables binarias al modelo, que hace que los tiempos de resolución se resientan. La complejidad de esta incorporación radica en que no existe un criterio claro y uniforme sobre la combinación de colores que debe tener una colación, y esta discusión entre los propios funcionarios del hospital demora la implementación de esta característica.

Paralelamente, los funcionarios del hospital encargados de la programación del servicio plantearon el interés de extender el número de días que abarca el menú a dos semanas e incluso un mes. Esto requiere la inclusión de nuevos platos para ampliar la posibilidad de combinaciones y va a implicar la revisión de algunas restricciones centrales del modelo como la imposibilidad de repetir platos de fondo y la limitación de repetir postres a los sumo dos veces, entre otras. Es posible que los tiempos de resolución también se vean afectados por esta ampliación del horizonte de planificación.

Finalmente, otra inquietud que despertó el interés del personal del hospital es la posibilidad de que el modelo contemple la diferencia en la cantidad de personas que reciben el servicio de colaciones al mediodía y en la cena. El hecho de que la población diurna del hospital duplica a la nocturna abre otra posibilidad de lograr un uso más eficiente de los recursos.

Agradecimientos. Los autores agradecen a la Lic. Cecilia Volk por su valioso aporte a este trabajo acercando datos y comentarios sobre los resultados.

Referencias

[1] J. L. Balintfy. A Mathematical Programming System for Food Management Applications. *Interfaces*, 6:13-31, 1975.

[2] J. L. Balintfy. The Cost of Decent Subsistence. *Management Science*, 25(10):980-989, 1979.

[3] G. Dantzig. The Diet Problem. *Interfaces* 20:43-47, 1990.

[4] G. Dantzig. Linear Programming and Extensions, Princeton Press, Princeton, New Jersey, 1963.

[5] FAO/WHO Expert Consultation. Carbohydrates in human nutrition. FAO Food and Nutrition paper No. 66. Rome, 1998.

[6] FAO/WHO/UNU Expert Consultation. Report on Human Energy Requirements. Interim Report, Rome, 2004.

[7] Food and Nutrition Board-Commission on Life Sciences-National Research Council. Recommended Dietary Allowances (10th ed). Washington DC: National Academy Press, 1989.

[8] Food and Nutrition Board/Institute of Medicine. Dietary Reference Intakes (DRI) and Recommended Dietary Allowances (RDA) for energy, carbohydrate, fiber, fats, fatty acids, cholesterol, proteins and amino acids. Institute of Medicine of the National Academies. Washington DC. The Nacional Academy Press, 2002.

[9] Food and Nutrition Board/Institute of Medicine. Dietary Reference Intakes (DRI) for Calclium, Phosphorus, Magnesieum, Vitamin D and Fluoride. Institute of Medicine of the National Academies. Washington DC. The Nacional Academy Press, 2002.

[10] S. G. Garille and S. I. Gass. Stigler's Diet Problem Revisited. *Operations Research*, 49:1-13, 2001.

[11] T. Koch. Rapid Mathematical Programming. ZIB-Report 04-58, 2004.

[12] T. Achterberg. SCIP: solving constraint integer programs. *Mathematical Programming Computation*, 1(1):1-41, 2009.

[13] G. J. Stigler. The Cost of Subsistence. *Journal of Farm Economics*, 27:303-14, 1945.

[14] WHO/FAO Joint WHO/FAO Expert Consultation. Diet, Nutrition and the Prevention of Chronic Diseases. Technical Report Series No. 916. Geneva, 2003.

Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicaciones

Francisco Barrientos *
Sebastián A. Ríos *

Resumen

Hoy día, la minería de datos está cobrando relevancia creciente en empresas u organizaciones para resolver problemas complejos de negocio. Por ejemplo, el procesamiento de volúmenes masivos de datos donde se esconde información valiosa respecto del comportamiento de compra de productos o servicios, hasta generar nuevos productos usando dichos comportamientos de los clientes. Esto es particularmente cierto en el mercado de negocios de las telecomunicaciones donde el número de clientes normalmente llega a varios millones y un analista humano es incapaz de realizar su labor sin metodologías y algoritmos que permitan automatizar o semi-automatizar el descubrimiento de conocimiento. El problema es aún más complejo, si tomamos en consideración que estas empresas cuentan con muchos sistemas internos desde los cuales debe ser obtenida la información necesaria y suficiente para poder realizar cualquier modelo predictivo. Este paper tiene por objetivo mostrar una metodología para poder realizar predicción de fuga de clientes ó Churn en un ambiente multiplataforma en la industria de las telecomunicaciones. Además, se usaron diversos algoritmos como Redes Neuronales, Support Vector Machines y Árboles de Decisión y se evaluó la calidad como el porcentaje de aciertos en la variable predicha. Esta fue aplicada a una empresa de telecomunicaciones real; donde se utilizaron los resultados para poder generar estrategias de retención de clientes y así realizar la evaluación de la calidad de los resultados obtenidos.

Palabras Clave: Proceso KDD, Minería de Datos, Modelos Predictivos, Predicción de Churn

*Centro de Investigación en Inteligencia de Negocios (www.ceine.cl). Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile, Santiago, Chile.

1. Introducción

El estudio del churn o fuga de clientes es un área en la cual año a año se invierten grandes recursos: equipos de especialistas, consultoras externas, software especializado, etc. Siempre, con la intención de poder descubrir de manera anticipada, si es que un cliente va a decidir cambiarse de una compañía a su competencia. Por ejemplo, pasar del Retail A al B, o de la Farmacia X a la Y. En particular, en el área de las telecomunicaciones, se ha hecho cada vez más necesario estudiar la fuga de clientes, dada la alta competitividad que se está desarrollando a nivel mundial y las nuevas legislaciones que están surgiendo en Chile (como, la portabilidad numérica). En la industria de las telecomunicaciones durante los años 2008 a 2010, la fuga de clientes llegó a ser del 30 % anual [28, 37] (estudios previos a la portabilidad numérica). A partir de lo anterior, se desprende que la fuga de clientes es un problema de la industria y donde se hace necesaria la aplicación de herramientas avanzadas que permitan predecir y describir de algún modo, qué clientes tienen mayor potencial de riesgo de cambiarse de compañía.

Las líneas de negocio presentes que encontramos comúnmente en empresas de telecomunicaciones son: tráficos de larga distancia, telefonía local, internet, cargos de accesos, servicios privados, facturas y cuentas corrientes atención de clientes (referente a las solicitudes de atención y reclamos), clientes y contratos, modelos de operaciones, participación de mercado y suscriptores. A nivel mundial, el tamaño de esta industria es de 4,03 trillones de dólares, cifra que se pronostica que crezca a un 6 % hasta el año 2013 [4]. Respecto al tamaño local en cuanto a inversiones realizadas en este mercado, la cifra rodea 778.153 millones de dólares [13].

Dentro de las principales razones para que un cliente deje de comprar productos de una compañía se destacan la disconformidad y la falta de políticas de retención efectivas expresadas en un mejor trato hacia ellos [27]. Para posicionar lo anterior en un contexto local se detectan cerca 11.809 reclamos en el organismo regulador de los cuales 5.242 corresponderían a disconformidades, ya sea con la suscripción, continuidad y calidad del servicio o bien con cobros irregulares [14]. Otra de las características del sector es la gran cantidad de información que generan y almacenan sus empresas [1], por ende, nuevas tendencias e ideas han aparecido para hacer uso de la minería de datos en una gran cantidad de áreas, en especial, en marketing, detección de fraude y control de calidad [1], donde la fidelización juega el rol más importante. Las empresas del sector están

especialmente sensibilizadas con la pérdida de clientes que escogen una compañía de la competencia, varios autores señalan que es más costoso conseguir un nuevo cliente que mantener uno antiguo [1, 28, 31, 43]. Por ello la finalidad de este paper es introducir el procedimiento KDD dentro de la empresa de telecomunicaciones para que pueda ser aplicado en todas sus áreas, en particular, con el objeto de detectar la fuga de clientes. Es así como se define el problema a tratar el cual consiste en predecir la fuga de clientes en esta empresa para un producto particular (NGN) cuyo segmento objetivo son las pequeñas y medianas empresas (PYMES). Una de las principales dificultades es que la información necesaria para predecir la fuga de clientes se encuentra distribuida en un sistema multiplataforma, el cual consiste en múltiples plataformas con información de clientes, transaccional, reclamos, entre otras. Estas plataformas no disponen de transferencias automáticas de información lo que dificulta aun más este trabajo. Dentro de este problema el término churn hace su aparición, el cual se es usado en el sector de telecomunicaciones para describir el cese de servicios de la suscripción de un cliente. Se habla de churner o fugado para denominar a aquel cliente que ha dejado la compañía [5].

2. Revisión de literatura

2.1. Conceptos básicos y tipos de fuga de clientes

La fuga de clientes, dentro de las telecomunicaciones, se produce cuando un cliente cancela el servicio prestado por la compañía [31]. En dicha cancelación, el cliente puede decidir renunciar a la empresa (voluntaria), o bien, la empresa puede expulsarlo (involuntaria). En particular, la connotación de churn hace referencia la fuga de los clientes, por lo que, para efectos de este estudio, se cuenta el churn en base a la decisión del cliente en abandonar la empresa por medio de la cancelación de un servicio. También, se puede entender como churn a aquel término *"usado para describir colectivamente el cese de servicios de la suscripción de un cliente...donde el cliente es alguien que se ha unido a la compañía por al menos un período de tiempo...un churner o fugado es un cliente que ha dejado la compañía"* [5]. Los principales tipos de fuga acorde a [5, 40] son:

- Absoluta: suscriptores que se han desligado sobre la base de datos total en un período.
- De línea o servicio: Este tipo de churn el número de servicios disconti-

nuados sobre la base de datos total

- Primaria: Referente al número de fallas
- Secundaria: Descenso en el volumen de tráfico
- Fuga de paquete: Esta fuga se caracteriza por el hecho de que cambian de planes y/o productos dentro de la compañía [5].
- Fuga de la compañía: Sin lugar a dudas el más costoso, en este caso el cliente se fuga hacia la competencia, por ende, no solamente se pierde el ingreso no percibido, sino que también el prestigio de la compañía expresado en la participación de mercado de la competencia.

2.2. El proceso KDD

La minería de datos se establece como una de las etapas de un proceso más genérico denominado Knowledge Discovery in Databases (KDD), el cual, es el proceso de análisis de bases de datos que busca encontrar relaciones inesperadas que son de interés o valor para el poseedor de dicha base de datos [20]. En términos simples es encontrar relaciones no triviales en los datos. Este proceso iterativo consiste en cinco etapas, en donde la minería de datos es definida como una fase más de este procedimiento. Al ser un proceso iterativo es posible volver a una etapa previa en caso de que no se tengan resultados satisfactorios al final de una. A continuación se describe dichas etapas o fases bajo la perspectiva de [11, 16]:

- Integración o Selección: Aquí se escogen las variables y las fuentes a considerar en el proceso completo, por lo que se refiere a la creación del conjunto de datos como la base de datos de estudio en el proceso. Dentro de este paper, esta etapa adquiere gran importancia, pues está sujeta a la interacción de múltiples plataformas como parte de selección de fuentes de información. La sección 2.2.1 presenta una descripción detallada de esta etapa.
- Preprocesamiento: El análisis y limpieza de los datos son las líneas principales a seguir en esta sección, donde se produce el tratamiento de valores ausentes (missing), los valores fuera de rango (outliers). Para ello, se emplean distintas técnicas de imputación de datos que van desde un tratamiento valor a valor (simple imputation) hasta un reemplazo contemplando múltiples variables y sus valores (multiple imputation). La sección 2.2.2 presenta una descripción detallada de esta etapa.

- Transformación: Acá se generan nuevas variables a partir del estudio de la naturaleza de las variables originales; desde la perspectiva de la escala, nominal o continua, o bien de la distribución de los valores presentes. La sección 2.2.3 presenta una descripción detallada de esta etapa.

- Minería de datos: Este paso en el proceso de KDD, consiste en la aplicación de análisis de datos para descubrir un algoritmo ad-hoc que produzca una particular enumeración de patrones a partir de los datos y que los produzca considerando restricciones de capacidad computacional [16]. Por ende, se selecciona el modelo y algoritmo a utilizar, bajo los supuestos que mantienen los objetivos primarios del estudio. La sección 2.2.5 presenta una descripción detallada de esta etapa.

- Interpretación y Evaluación: Esta última fase involucra las medidas de evaluación y la trasposición de resultados técnicos a niveles comerciales, de tal manera, que la aplicación del procedimiento converja a acciones correctivas en el negocio, que solucionen el fenómeno estudiado. Respecto a la evaluación, ésta se puede aplicar desde dos aristas: técnica y comercial. La primera se subdivide acorde al tipo de validación y sus métricas que se aplican al modelo, mientras que la evaluación comercial no se encuentra estandarizada y generalmente se puede utilizan encuestas o grupos de blindaje para medir la efectividad práctica del procedimiento. Las principales técnicas de evaluación técnica son el "holdout" y la validación cruzada. La sección 2.2.6 presenta una descripción detallada de esta etapa.

2.2.1. Data Warehousing

En la primera etapa del KDD (Integración), se requieren fuentes de información consolidadas, por ello, es que generalmente se aplica este procedimiento posterior a la implementación de un data warehouse (DWH) en la compañía. Este concepto se define como la colección de tecnologías de soporte decisivo que permite al trabajador tomar buenas y rápidas decisiones [9]. Esta colección debe ser orientada al sujeto, integrada, variante en el tiempo y estable, por ende, generalmente, se mantiene apartada de las bases de datos operacionales, pues se busca la consolidación de los datos [9, 41]. Por lo tanto, es lógico pensar que un data warehouse contiene datos consolidados a partir de múltiples bases de datos operacionales, durante extensos períodos de tiempo, por lo que es común que su tamaño alcance varios gigabytes o terabites [9]. Es importante destacar que cada estructura en un data warehouse posee una dimensión temporal [41].

Por lo anterior, la implementación de un data warehouse en la empresa es un proceso largo y complejo [9]. En ocasiones, las compañías optan por usar data marts en vez de construir un data warehouse, estos son subconjuntos de datos orientados a un departamento o área determinada, por lo mismo, no requieren de un consenso general a nivel de empresa, sin embargo, si no se incorpora estratégicamente la utilización del data warehouse se pueden producir problemas complejos de integración en el largo plazo [9]. Las principales diferencias se muestran a continuación en la Tabla 1:

DATA WAREHOUSE	DATA MART
Implementación a nivel de corporación o empresa	Implementación a nivel departamental
Aplicación a la unión de todos los data marts	Aplicación a un proceso de negocios singular
Consultas en recurso de presentación	Tecnología óptima para el acceso de datos y análisis
Estructura orientada a una vista empresarial de los datos	Estructura orientada a una vista departamental de los datos
Organización en base a un modelo entidad relación	

Tabla 1: Diferencias entre un Data warehouse y un Data mart

El diagrama general de la confección de un data warehouse puede observarse tal y como procede [9] en la Figura 1:

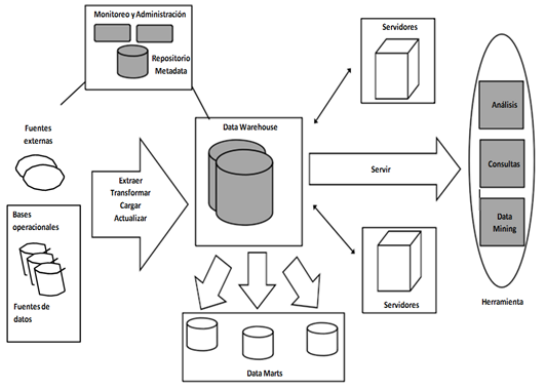

Figura 1: Diagrama estructural de un Data Warehouse

Ahora bien, la importancia de esta tecnología en las telecomunicaciones se debe a que generan un cambio en la forma en que la gente ve el desarrollo en sistemas de los [38], es decir, permite acercar el modelo de negocios a los empleados. Eventualmente a partir de esta colección de datos se requiere extraer información útil a los usuarios finales para lo cual se puede aplicar data mining [38].

2.2.2. Imputación de datos

Posterior a la integración de datos facilitada por un data warehouse, se analizan los conjuntos respectivos, cuyos primeros fenómenos observables son la aparición de valores anómalos, ausentes, o con variabilidad en el tiempo. Los valores ausentes son aquellos que no se encuentran en la base de datos analizada, su subdivisión se efectúa en base a las posibles razones de dicho fenómeno y es la siguiente [42]:

- Missing Completely at random (MCAR): Son aquellos datos perdidos que son completamente al azar, es decir, no poseen ningún tipo de relación con los datos presentes en otras variables. En este escenario, se asume que las distribuciones de probabilidad de los datos faltantes y de todos los datos son idénticas.

- Missing at Random (MAR): Estos datos o valores a diferencia de los anteriores, presentan relaciones con el resto de las variables, por lo que la distribución de probabilidad de los valores perdidos puede tener una distribución distinta a la variable en general. Los valores de este tipo pueden ser predichos usando datos completos[12, 30].

- Not Missing at random (NMAR): Son valores perdidos que tienen significado, es decir, el hecho de que esté ausente no es un error, sino información relevante para la variable

Otro tipo de fenómeno son los valores fuera de rango u outliers que se definen como aquellos valores en el conjunto de datos cuyo comportamiento es anómalo con respecto a lo observado en la mayoría de los registros [2].

Finalmente, es usual la existencia de las variables temporales, las cuales son variables cuyos valores pueden ser diferentes en distinto momentos del tiempo [8].

El trato que se le da a los datos afectos a dichos fenómenos es distinto dependiendo del tipo descrito previamente. Para el tratamiento de los valores ausentes existen distintas alternativas, dentro de las cuales, las más relevantes son [12]:

- Descarte de los registros con datos faltantes: Esta alternativa suele utilizarse cuando la información que aporta la variable es baja, o bien, la cantidad de valores perdidos es baja y la variable tiene poca varianza.

- Reemplazo de los datos faltantes con otro valor: Esta alternativa suele usarse para identificar al valor ausente.

- Imputación de los datos faltantes: Esta alternativa es factible cuando la cantidad de atributos con datos faltantes es relativamente pequeña en relación al número de registros que presentan dicha condición. Este método, sí influye en la información de la variable.

Otra perspectiva o alternativa de solución, frente a los datos perdidos, la presentan Roderick Little et al. [30], que establece las siguientes técnicas para tratar los valores ausentes:

1. Procedimientos basados en instancias completas: Esto se refiere a que cuando algunas variables no están guardadas para determinadas instancias, se sugiere simplemente eliminar éstas y analizar solamente las instancias con atributos completos. Los análisis usuales de este tipo son:

   - Listwise deletion: Este análisis sugiere que al momento de realizar la predicción, se debe trabajar con las observaciones que disponen de la información completa para todas las variables [17].
   - Parwise deletion: Este análisis considera la información completa, pero usando distintos tamaños de muestra. A diferencia del análisis listwise, solamente se eliminan aquellas observaciones que no poseen ningún dato, y los cálculos se realizan con diferentes tamaños de muestra lo que limita comparación de resultados [17].

2. Procedimientos basados en la imputación: En estos procedimientos *"los valores perdidos son llenados y la base de datos completada es analizada por métodos estandarizados. Los métodos comúnmente usados incluyen Hot Deck, Imputación por promedio e imputación por regresión"* [30], además, se agregan otro métodos como el cold deck. El hot deck imputa para cada ejemplo que contenga un valor perdido, se encuentra el ejemplo más similar y los valores perdidos son imputados de dicho ejemplo [12]. El método de la media consiste en una imputación de los valores anómalos por la media de la variable. La regresión sugiere *"imputar la información de una variable Y a partir de un grupo de covariables $X_1, X_2, \ldots, X_n$"* [29]. El cold deck consiste en seleccionar los valores o relaciones de uso obtenidos de fuentes distintas al conjunto de datos actual [44].

3. Procedimiento de asignación de pesos: Este método es usado cuando se desea adquirir una probabilidad de selección, pues "las inferencias aleatorias de la muestra de una encuesta sin respuesta comúnmente están basadas en pesos de diseño que suelen ser inversamente proporcionales a la probabilidad de selección" [30]. Usualmente se usa para estimar la población promedio a partir de una muestra. Reponderación: Este tipo de

imputación adquiere importancia cuando se tienen muy pocas respuestas o valores de alguna categoría de interés. Las ponderaciones se interpretan como el número de unidades de la población [17].

4. Procedimientos basados en modelos: Este procedimiento se aplica cuando se intuye una relación, lineal o no lineal, entre un subconjunto de variables. No obstante, el procedimiento completo se sintetiza en *"definir un modelo para los valores parcialmente perdidos y basando las inferencias en la similitud del modelo, con parámetros estimados bajos procedimientos tales como el de máxima verosimilitud"* [30].

2.2.3. Transformaciones especiales: ACP, Segmentación, RFM

- Transformaciones para variables temporales: Estas variables son referenciadas como secuencias temporales, las cuales *"se forman con los datos recopilados en una base sobre la evolución en el tiempo de un conjunto de características"* [39]. Las transformaciones que se utilizan en estas variables son: Índices, por funciones que preserven la ortonormalidad o promedio ponderados.

- Análisis factorial: Este análisis tiene el propósito de *"simplificar las numerosas y complejas relaciones que se puedan encontrar en un conjunto de variables cuantitativas observadas"* [29]. Esto quiere decir que no se encarga de reducir las variables, sino que busca encontrar el significado de los nuevos factores generados producto del análisis de componentes principales. Por ende, su definición converge a *"un procedimiento matemático mediante el cual se pretende reducir la dimensión de un conjunto de p variables obteniendo un nuevo conjunto de variables más reducido, pero capaz de explicar la variabilidad común encontrada en un grupo de individuos sobre los cuales se han observado las p variables originales"* [29].

- Modelo Recency, Frequency, Mount (RFM): este modelo data antes del año 2000 en el marketing cualitativo, como forma de medir el comportamiento del consumidor. Esta medición se hace desde tres perspectivas [23]. La primera es Recency, que indica hace cuánto el cliente respondió [23], que en otras palabras, significa el tiempo transcurrido desde la última vez que el cliente registró un accionar con la compañía. La segunda perspectiva es la Frecuencia, que provee una métrica de cuán seguido el cliente ha respondido a recibir mails, que pretende indicar, el tiempo transcurrido entre interacciones del cliente con la compañía. Finalmente, la tercera perspectiva es el valor monetario, que mide el monto en dinero o

el número de productos que el cliente ha gastado o consumido en respuesta a los mails enviados [23], lo que expresado de forma distinta, indica el valor monetario o cantidad que el cliente gasta, emplea o consume en cada accionar con la compañía. De esta manera, se puede reformular el modelo, para el caso de los reclamos en las telecomunicaciones las siguientes variables:

- Recency (R): La última vez o mes que el cliente emprendió un reclamo hacia la compañía
- Frecuencia (F): El número de meses en las que el cliente reclamó.
- Monto (M): El número de reclamos promedio involucrado en cada ocasión.

Generalmente M va asociado a un valor monetario, no obstante, acorde con J.-J. Jonker en [23], también se puede ocupar como una variable de monto de acciones.

2.2.4. Problema de clases desbalanceadas

De vez en cuando, dependiendo del tipo de mercado, aparece este problema expresado en la base de datos respecto a las clases, es decir, a las categorías o valores de la variable objetivo. Esta rareza de clases o desbalanceo de clases se da cuando existe escasez de una de las clases, esta escasez puede ser de dos tipos [18]:

- Rareza de clases: se define como la ocasión en que un valor de la variable objetivo se encuentra fuera del común denominador o en los extremos de la distribución de la variable objetivo.
- Rareza de casos: el segundo tipo corresponde a un conjunto de datos significativo pero a su vez pequeño [18], en otras palabras, son aquellas instancias que escapan al común, en cuanto a su comportamiento. Ambos tipos de escasez son consideradas una desbalanceo interno de las bases de datos. Este tipo de problema conlleva a dificultar la labor de la implementación del KDD, debido a que existen consecuencias asociadas a la ignorancia de dicha problemática, entre las cuales destacan [6, 18]:
  - Métricas de evaluación inapropiadas: En ocasiones, la métrica que ayuda a construir el modelo se basa en obtener una certeza adecuada, sin embargo, en el caso de que existe una rareza de clases del 1 %, estos modelos se construirán para obtener el 99 % de certeza, dejando de lado la clase rara que puede ser aquella de interés para el

proyecto. En otras palabras, esto indica que las clases raras tienen menor impacto en el accuracy (o certeza) que las clases comunes [18].

- Escasez de datos (Rareza absoluta y relativa): Esta consecuencia se da en las bases de datos en donde la cantidad de instancias que pertenecen a la clase rara es mucho menor con respecto al resto de las clases.
- Fragmentación del conjunto de datos: es un problema adjudicable al momento de aplicar un algoritmo en la etapa de minería de datos, porque las regularidades pueden ser solamente encontradas en particiones individuales que contienen menos datos [18], esto quiere decir que los patrones finales terminan bajo la influencia de los patrones internos de cada partición.
- Tendencia inducida: En la minería de datos para comprender el patrón general subyacente en el problema, se tiende a inducir tendencias, de hecho, muchos algoritmos de aprendizaje usan una tendencia general de manera de encontrar la generalización y evitar el sobreajuste, por lo tanto, la tendencia puede impactar la habilidad de aprender de casos o clases raras [6].
- Ruido: El ruido, cuando es consecuencia de rareza, tiene un mayor impacto sobre los casos raros que sobre los casos comunes, puesto que los casos raros tienen menos instancias para empezar, por lo tanto, requerirán menos ejemplos ruidosos para impactar el subconcepto aprendido [18].

2.2.5. Técnicas de Minería de Datos

En la etapa de minería de datos, se encuentra una variedad de modelos y perspectivas a aplicar, es en la sección modelos donde entran los algoritmos. De los cuales, se proceden a describir los más usados en el mundo de la investigación:

- K-Nearest Neighboor (KNN): El algoritmo del K vecino más cercano o KNN es uno de los algoritmos más simples. Este algoritmo no requiere de ningún parámetro fuera del número de vecinos a considerar. En pocas palabras, el algoritmo puede resumirse en que "reúne los K vecinos más cercanos y los hace votar, la clase con más vecinos gana, ..., mientras más vecinos consideramos, menor la tasa de error" [22]. Dicha cercanía, generalmente se mide en base a alguna distancia, por lo que se pueden obtener distinto resultados dependiendo de la distancia escogida, pues

diferentes métricas definirán diferentes regiones [26]. Su esquema general se propone a continuación:

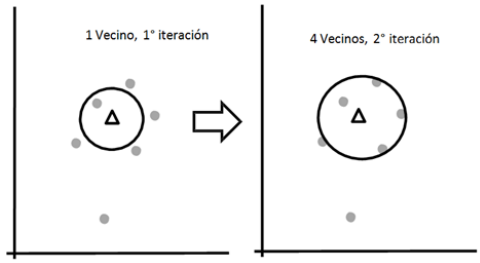

Figura 2: Modelo estructural del KNN

- Naive Bayes: En general los algoritmos de clasificación que utilizan el aprendizaje bayesiano resultan complejos en el sentido de la cantidad de parámetros. Sin embargo, el método de naive bayes convierte dicha complejidad en una simpleza factible, *"debido a que hace un supuesto de independencia condicional que reduce el número de parámetros a estimar, cuando se modela* $P(x \mid y)$*"* [36]. De forma cuantitativa, si la variable a predecir tiene dos valores pasa de estimar $2(2n-1)$ parámetros a $2n$. La utilidad de los algoritmos de aprendizaje bayesiano es que *"da una medida probabilística de la importancia de esas variables en el problema, y, por lo tanto, una probabilidad explícita de las hipótesis que se formulan"* [32]. Una explicación de la matemática subyacente de este algoritmo se encuentra en cite36.

- Árboles de Decisión: Los árboles de decisión son modelos que usualmente se representan en forma de grafos. Es *"un modelo predictivo que puede ser usado para representar tanto modelos regresivos como aquellos de clasificación, se refiere a un modelo jerárquico de decisiones y sus consecuencias"* [33]. Dentro de un esquema general, el árbol de decisión consiste en un grafo donde existe un nodo único o parental, el cual, contiene las instancias a contemplar en el modelo. Un ejemplo de este tipo de modelos es el LADTree, que es un tipo de árbol de decisión que itera sobre el ADTree que es un árbol que en vez de establecer criterios y dividir la muestra, asigna una puntuación a las categorías relevantes de determinadas variables.

- Support Vector Machines (SVM): A diferencia de los algoritmos anteriores, la máquina de soporte vectorial, o bien, Support Vector Machines, utilizan planos complejos para encontrar la mejor división de las instancias que permita clasificarlas de manera óptima. Cuya formulación es un

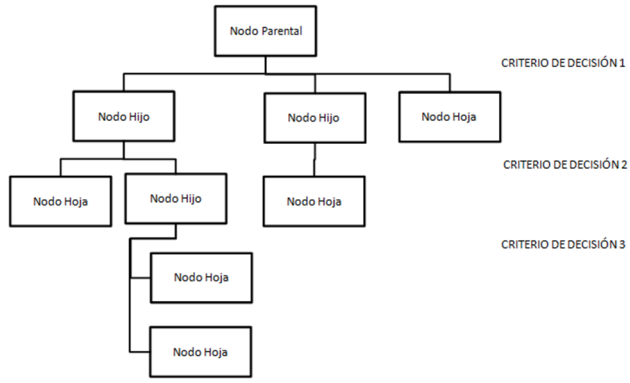

Figura 3: Modelo estructural de un Árbol de Decisión

problema de minimización cuadrática con un numero de variables igual al número de casos de entrenamiento [7, 35]: Existen además formulaciones para entrenar SVM usando programación lineal, estas formulaciones están basadas en la consideración de las normas L_1 y L_∞ en lugar de la norma L_2 [3]. Por lo tanto para grandes números de datos, se debe usar un equipo de gran capacidad. Además, las SVM utilizan la rama de optimización de la matemática, puesto que el problema que abordan involucra optimización de una función convexa [45], esto quiere decir que no contiene mínimos locales. Otra particularidad que comprende este algoritmo es que no requiere información acerca de la distribución del conjunto de datos. Es decir, se busca el balance entre certeza y cantidad de datos a aceptar. Es esta combinación la que induce el origen del problema de minimización del riesgo estructural. Cuya solución se traduce posteriormente en un problema de optimización sobre encontrar el hiperplano separador entre las instancias, mostrado en la siguiente figura:

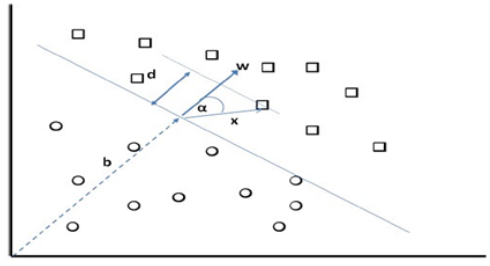

Figura 4: Modelo estructural de una SVM linealmente separable

Donde la formulación para un caso linealmente separable es:

$$\min_{w} \frac{1}{2}\vec{w}^t * \vec{w}$$

$$\text{Sujeto a } y_i(\vec{x} * \vec{w} + b) - 1 \geq 0$$

$$\forall i = 1, \ldots l$$

- Redes Neuronales: Este modelo de minería de datos es una de las estrategias más populares para aprendizaje supervisado y clasificación. Sin embargo, debido a la complejidad que posee, no se puede saber con exactitud el origen de sus resultados, lo que es una dificultad a la hora de explicar su funcionamiento. En un sentido directo, una red neuronal artificial (o denominada simplemente red neuronal, o ANN) *"consiste en procesar elementos (llamados neuronas) y las conexiones entre ellos con coeficientes(pesos) ligados a las conexiones, las cuales constituyen una estructura neuronal, y un entrenamiento y algoritmos recordatorios adjuntos a la estructura"* [24], lo que en palabras simples puede ser descrito como *"una piscina de unidades simples de procesamiento que se comunican enviando señales entre ellas sobre un gran número de conexiones ponderadas"* [25]. Un esquema general de este modelo se presenta a continuación:

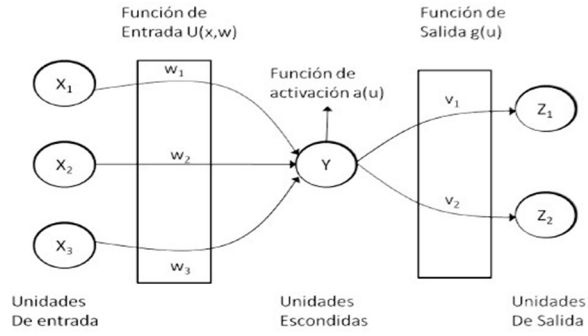

Figura 5: Modelo estructural de una red neuronal

- Regresión: La regresión, en la actualidad, consiste en *"el estudio de la dependencia de la variable dependiente, respecto a una o más variables (las variables explicativas), con el objetivo de estimar y/o predecir la media o valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en muestras repetidas) de las últimas"* [19]. Por lo que este modelo sirve para predecir y clasificar, donde su uso típico

es el de predecir la demanda o el inventario futuro de una empresa. En el caso de la clasificación, la regresión que se utiliza comúnmente no resulta muy efectiva, puesto que la variable a predecir posee una connotación nominal, sin embargo, existe un tipo de regresión que se encarga de predecir variables nominales y se denomina regresión logística.

- Multiclasificadores: Los multiclasificadores, a diferencia de los modelos anteriores, busca encontrar formas o combinaciones de volver una predicción o clasificación efectiva en una predicción eficiente. Para ello se pretende explorar la mayor cantidad de caminos abordables, abarcando toda la información posible. Sin embargo, al buscar esta eficiencia, este tipo de modelos suele caer en algoritmos de gran complejidad, además, puede suceder que el modelo no sea válido a un nivel de fundamentos matemáticos.

2.2.6. Métricas de evaluación

Las medidas de evaluación técnica que generalmente se usan, se basan en una tabla de contingencia que describe las instancias predichas acertadas y erróneas. Esta tabla de contingencia se denomina matriz de confusión que *"contiene información acerca de las clasificaciones actuales y las predichas, realizadas por un sistema de clasificación"* [21]. El esquema de ésta para un caso de clasificación binaria es:

		Clase Actual	
Categorías		0	1
Clase Hipotética	0	TN	FN
	1	FP	TP
Columnas Totales		N=FP+TN	P=TP+FN

Tabla 2: Esquema de tabla de confusión caso binario

En base a esta tabla se definen las siguientes métricas de carácter técnico [6]:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{P}$$

$$Accuracy = \frac{TP + TN}{P + N}$$

$$F - Measure = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

$$Lift = \frac{Precision}{\frac{P}{P+N}}$$

Curvas Roc: Son curvas que muestran la habilidad del clasificador para posicionar las instancias verdaderas respecto a las falsas[44]. En una definición más acertada se puede decir que las Curvas ROC son las que miden la relación de la tasa de verdaderos positivos (predicciones acertadas) versus la tasa de falsos positivos (predicciones erradas). Siendo el positivo el referente a la clase de fuga cuando se trata de un problema de clasificación binario. Estas curvas no tienen una fórmula asociada. No obstante, sí tienen una métrica, la cual llamada *"Area Under the curve"(AUC)*, que se define como el rea bajo la Curva ROC, además, tiene la siguiente propiedad estadística: *"La AUC de un clasificador es equivalente a la probabilidad que el clasificador posicionará una instancia aleatoria positiva mejor que una instancia aleatoria negativa"* [15].

3. Metodología del proceso

El procedimiento KDD ejecutado se basa en una experiencia previa existente en la empresa, que permitió identificar de forma rápida y efectiva gran parte de las fuentes de información utilizadas. Un breve esquema presenta el KDD aplicado con las bases de datos que contempla:

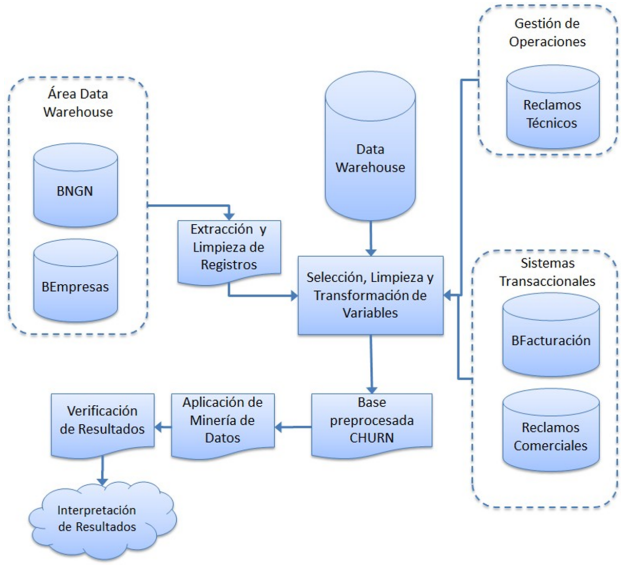

Figura 6: Esquema de procedimiento KDD Implementado

En donde las bases de datos explicitadas son descritas tal como sigue:

- Reclamos comerciales (RC): Esta base de datos contiene los registros con fecha de las solicitudes comerciales de los clientes ya sea de término de contrato, este tema no se tiene en su totalidad debido a que en esta base de datos sólo se registran las postventas en los call center, por lo tanto, es aquella que comprende las transacciones que el cliente efectúa vía telefónica con la empresa, es decir, la postventas, el cambio del servicio, término de contrato entre otros. Su periodicidad es mensual.

- BFacturación: Es la base de datos generada por el área del data warehouse destinada a temas de facturación y consumo. Su periodicidad es mensual, no obstante, su almacenamiento se mantiene cada tres meses (es decir, para obtener la base de datos correspondiente a Abril, se "borra" Enero del año en curso), es decir, si se desea averiguar la facturación del año 2009, se debe consultar directamente al data warehouse comercial.

- Reclamos Técnicos(RT): Base de datos destinada a hacer reportes acerca de los reclamos desde una perspectiva técnica. En otras palabras, en ella se encuentran todas las reparaciones a fallas en que el cliente ha avisado a la compañía. Por ende, su origen es el área de operaciones. Su periodicidad es mensual.

- BEmpresas: Esta base de datos contiene la totalidad de clientes y sus características descriptivas, es decir, su tamaño, clasificación, categoría, plan de retención, entre otros. Su origen es el área del data warehouse. Su periodicidad es mensual.

- BNGN: Esta base de datos contiene solamente los detalles de los clientes del producto NGN y sus planes correspondientes para cada cliente (con la vigencia respectiva referente a un contrato que puede contener múltiples planes). Su periodicidad es trimestral.

- Suscriptores: Proveniente directamente del DWH, es una base de datos que consta en sus registros de todos los teléfonos fijos existentes visibles en el mercado de las telecomunicaciones nacionales, por lo tanto, es una base de datos de 7,5 millones de registros. Debido a esto, se toma como una base de datos cuya variación es despreciable, es decir, estática y se obtiene por petición al área del data warehouse.

El conjunto de datos usado entregado desde distintas fuentes, poseía un total de 208 variables distribuidas en las múltiples bases de datos utilizadas. La cantidad de registros era aproximadamente 9000 clientes totales (vigentes y no

vigentes). Cabe destacar que las relaciones entre las bases de datos descritas en la figura 6 no necesariamente son de 1 a 1, puesto que un cliente puede registrar múltiples reclamos o solicitudes (RT y RC), así como también, tiene varios planes para un mismo contrato (BNGN sección paquetes), varios tickets de facturación (BFacturación) y teléfonos (Suscriptores). Las aristas que comprenden estas bases de datos son el comportamiento comercial y técnico del cliente, la información demográfica, las transacciones efectuadas y los equipos instalados. El horizonte de toma de datos contemplado para el estudio de fuga del servicio fue de 6 meses de historia.

El churn que se calcula en este paper es aquel referente al servicio, debido a que para que sea aplicable a nivel de cliente se debe implementar el KDD como procedimiento relevante en la compañía, además, ésta debe presentar una cultura más orientada a la retención en vez de a la fuerza de venta. Dicho churn en la compañía es de un 1 % aproximadamente, lo cual implica un problema de rarezas. La forma en que se trabajó dicho obstáculo comprendió desde el sobremuestreo que conllevó a caer en el overfitting de los modelos, la eliminación de los registros catalogados como fuera de rango y la segmentación de clientes y su predicción individual. Es esta última idea la que se aplicó como solución final. De las 9000 instancias se llegó a 5730 clientes, en una primera instancia, contemplados como vigentes. Otro punto relevante a destacar es que la ejecución del KDD contempló un total de 7 experimentos realizados en varios instantes de tiempo buscando medir la veracidad, la robustez y los resultados del procedimiento, dentro de estos experimentos se destacan aquel referente a un estudio histórico sobre la certeza del modelo en cuestión en ese momento (LADTree) y su evaluación técnica y comercial, así como también, el último experimento, el cual se describe primordialmente en este paper. En el experimento final declarado como cierre se ejecutó sobre 5692, una disminución que indica la migración de los clientes del servicio, lo cual se debió primordialmente a la aparición de un producto superior interno en la compañía cuyos clientes objetivos eran las empresas de mediano y gran tamaño que se encontraban insatisfechas con el producto NGN por la capacidad.

Respecto a las variables contempladas en el experimento final, se concretaron 43 variables, 17 catalogadas como nominales (incluyendo a la variable objetivo y el identificador) y 26 continuas.

Para el preprocesamiento se interpolaron variables de facturación posicionadas en la base de datos BEmpresas para obtener un indicador de los productos aparte que el cliente consumía aparte del NGN, se reemplazó por ceros para los valores ausentes de la BFacturación y para el caso de las bases de datos RT y RC. Se eliminaron variables en base a la alta varianza que poseían en el caso nominal, para el caso continuo, se eliminaron las variables con alta corre-

lación También se usó la moda para algunas variables nominales y tablas de contingencia para efectuar un reemplazo evitando la propagación de error en las relaciones inter-variables. En la base de datos de Suscriptores los valores perdidos se reemplazaron con un valor 0 solamente a modo de identificación para su posterior tratamiento.
Respecto a las transformaciones, las más relevantes, son la segmentación de planes NGN (cuya ubicación está en la sección paquetes de la base de datos BNGN) con un algoritmo de clusterización denominado Two-Step Clúster. De esta forma, se agruparon los planes y posteriormente la variable que se pudo relacionar directamente con el ID del cliente fue la cantidad de planes de un tipo determinado. Otra de las transformaciones fue la utilización del ACP para reducir la dimensionalidad de las variables provenientes de la BFacturación. Una tercera transformación se ejecutó para las variables de la RT bajo el uso del modelo RFM. Respecto a la base de datos de los Suscriptores la transformación que se utilizó detalla si el cliente tiene algún teléfono (ANI) en alguna de las compañías competidoras de la empresa que ejecutó este estudio.
En lo que se refiere a la etapa de modelamiento, se probaron varios enfoques, uno directo (base de datos + variable fuga), uno indirecto (base de datos + variable de segmentación + variable fuga), uno agregado (base de datos + variable de segmentación creada a partir de la variable fuga) y uno separado (se separa el conjunto de datos en dos en base a la variable fuga para segmentar cada subconjunto por separado y posteriormente predecir con esta variable incluida. Este último fue el usado en el experimento final y es bosquejado a continuación:

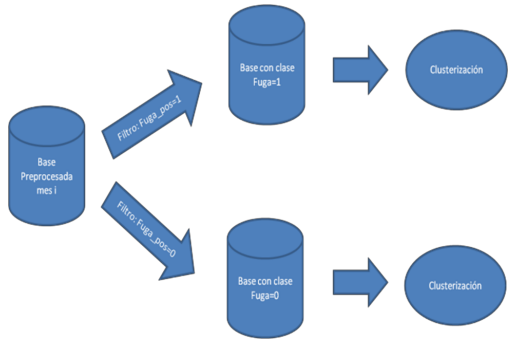

Figura 7: Modelamiento del experimento final

Para la etapa de evaluación, se contemplaron 2 perspectivas, la técnica y la comercial. La primera se orientó primordialmente en validar los resultados la mayor cantidad de veces que fuese posible, es decir, se procedió bajo la siguiente secuencia:

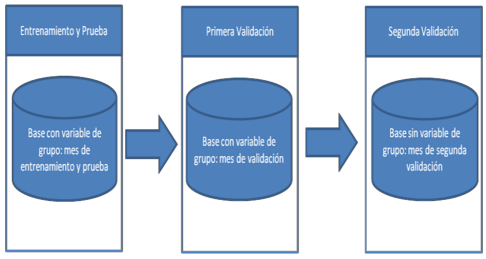

Figura 8: Esquema de evaluaciones

Donde la fase de entrenamiento-prueba es donde el modelo y aprende con datos preexistentes, la fase de primera validación es donde se predice sin conocimiento del valor de la variable objetivo, pero conociendo el grupo (del mes anterior) y al momento en que se posea dicho valor se contrasta para establecer la comparativa correspondiente. La fase de segunda validación, es en el caso en que aparte de no contar con la variable objetivo no se cuenta con la variable grupo del mes anterior. Las medidas técnicas utilizadas fueron todas las descritas en el capítulo anterior. Cabe mencionar que la predicción se midió marginalmente para ver la robustez de los modelos de minería de datos en el corto plazo (un mes).
Para la evaluación comercial se tomaron dos muestras, la primera se realizó posterior al estudio histórico efectuado en pos de obtener la veracidad en el tiempo del modelo escogido, la cual consistió en un muestreo aleatorio simple (MAS). Mientras que la segunda muestra se ejecutó en el experimento final, mediante un muestreo estratificado por grupo.

4. Aplicación experimental del proceso

En esta sección se presentan los resultados de la aplicación del proceso KKD para la predicción de fuga de clientes. Se comienza con la presentación de los resultados de aplicar transformaciones a las variables en estudio. Luego, se presentan una serie de experimentos que sirvieron como base para el experimento final. Por último, se analizan estos resultados en pos de proveer una clasificación que sea consistente en el tiempo.

4.1. Resultados de transformaciones relevantes

Previo a mostrar los resultados finales, se destacan los resultados de las transformaciones más relevantes. Para el caso de la segmentación de planes, se tomaron 5 variables: BA (Banda Ancha), velocidad, cantidad de teléfonos (ANIS), ADSL 2+(expansión del ADSL común) y Tecnología (Wiimax o de cobre). Usando el algoritmo de clusterizacion Two-Step clúster, se obtuvo lo siguiente:

Características por Grupo							
Grupo	Total	Anis [N]	Velocidad Glosa([KB])	BA	Tecnología	ADSL	Nombre
1	6422	(1 a 5)	Media (392 a 1714)	SI	COBRE	NO	Plan Estándar
2	1155	(1 a 5)	Alta (2000)	SI	MIXTO COBRE	SI	Plan ADSL
3	1518	(1 a 6)	Media Baja (587 a 685)	SI	WIMAX	NO	Plan Wimax
4	3833	(1 a 5)	Nula (0 a 0)	NO	COBRE	NO	Plan Sin Banda Ancha
5	830	(1 a 25)	Baja (0 a 179)	NO	MIXTO WIIMAX	BAJO	Plan Personalizado

Tabla 3: Resultado de segmentación de planes

Esta clusterización convertida a segmentación se realizó el mes de Julio, para la clasificación de nuevos planes se implementó un árbol de decisión simple. El número de clústeres se tomó a partir del resultado con las mismas variables de un clúster jerárquico.
La segunda transformación, es decir, el ACP, agrupo 6 variables de facturación y 6 variables de consumo en dos factores que describían la facturación y el consumo respectivamente. Los resultados que validaron el ACP para los meses acordes al último experimento, se muestran a continuación:

KMO y prueba de Bartlett				
Meses		dic-10	ene-11	feb-11
Medida de adecuación muestral de Kaiser - Meyer-Olkin.		0.9	0.9	0.92
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	178316.36	186998.92	202874.63
	gl	66	66	66
	Sig.	0	0	0

Tabla 4: Indicador KMO para el análisis ACP

Esta tabla representa que la transformación por ACP de estas 12 variables fue adecuada.

4.2. Experimentos previos

Los primeros experimentos permitieron establecer una estrategia coherente en las etapas de Integración y preprocesamiento en el KDD, además, fueron el primer acercamiento a una predicción validada. Los resultados que comparan los modelos se muestran a continuación, para el período de Marzo 2010:

	Modelos			
Criterios técnicos	J48	LADTree	Random Tree	Random Forest
Accuracy [%]	89.54 %	89.76 %	68.29 %	88.76 %
Medida F [%]	88.94 %	89.31 %	62.64 %	87.70 %
AUC	0.928	0.948	0.672	0.943
Lift [%]	232.41	231.19	164.47	238.07
TN	1545	1536	1489	1581
TP	852	867	339	795
FN	88	73	601	145
FP	192	201	248	156

Tabla 5: Comparativa entre múltiples modelos

Esta tabla se obtuvo al hacer un muestreo estratificado del total de clientes vigentes. En ella se puede apreciar que el modelo tentativo a escoger es el LADTree. Utilizando este modelo, se llegó al siguiente resultado validado para el período de Abril 2010:

Categorías predichas	Vigencia real	Fuga real	Precisión de la clase
Vigente	5587	12	99.79 %
Fuga	89	42	32.06 %
Recall de la clase	98.43 %	77.78 %	
Medida F clase 1 (Fuga)	45.41 %	F TOTAL	75.42 %
Medida F clase 0 (Vigente)	99.10 %		
Accuracy	98.24 %		
Correctamente Clasificadas	5629	98.24 %	
Incorrectamente Clasificadas	101	1.76 %	

Tabla 6: Tabla de confusión en la validación del mes de Abril de 2010

Este modelo se entrenó y probó con la base de datos del período Marzo 2010. Sin embargo, se deseaba establecer un modelo continuo que se sustentase

en el tiempo como ventana móvil. Para el período de Mayo 2010 se efectuó una segunda validación, cuyos resultados fueron:

Validación entrenando con Marzo			
	Categorías Real		
Categorías Predicción	Vigente	Fuga	Precisión de la clase
Vigente	5627	40	99.29 %
Fuga	127	1	0.78 %
Recall de la clase	97.79 %	2.44 %	
Accuracy	97.12 %		
Medida F Total	50.08 %		
Medida F clase SI (Fuga)	1.18 %		

Tabla 7: Tabla de confusión para el mes de Mayo de 2010

A partir de los cuales se desecha el procedimiento ejecutado anteriormente. Esto permitió deducir que si bien un modelo permite predecir en una ventana de tiempo determinada, no implica necesariamente que ese aprendizaje y forma de predecir se mantenga en el tiempo.

4.3. Estudio histórico

En pos de buscar una solución a la problemática anterior de que el resultado fuese sustentable en el tiempo, se realizó un estudio histórico, para el cual se rescató información del data warehouse en el caso de las bases de datos BNGN, mientras que para las otras bases de datos se solicitó información solo para los meses correspondientes. Debido a errores en el manejo operacional del data warehouse, se optó por añadir a los fugados posteriores a Julio como clientes vigentes de Junio, por ello, las bases de datos anteriores a Octubre presentan un sesgo y una mayor cantidad de clientes catalogados como vigentes. Esta situación se puede observar en el gráfico resultante de este estudio representado a continuación:

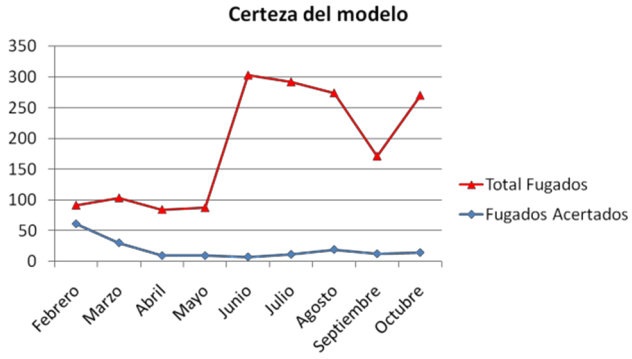

Figura 9: Gráfico histórico de certeza del modelo LADTree para un período de 11 meses

Cabe destacar que el modelo cuyos resultados fueron consistentes en el tiempo fue el LADTree, por lo tanto, se aplicó una predicción para los posibles fugados del mes de Diciembre de 2010, cuyos resultados fueron validados técnicamente con la variable fuga extraída a partir de los términos de contrato de la base de datos RC.

	Validación entrenando con Marzo		
Categorías Predicción	Categorías Real Vigente	Fuga	Precisión de la clase
Vigente	5627	40	99.29 %
Fuga	127	1	0.78 %
Recall de la clase	97.79 %	2.44 %	
Accuracy	97.12 %		
Medida F Total	50.08 %		
Medida F clase SI (Fuga)	1.18 %		

Tabla 8: Tabla de confusión post-estudio histórico

Aparte de esta validación técnica, se realizó un MAS para efectuar una encuesta telefónica para generar una validación comercial. Los resultados de este MAS se representan en la siguiente tabla:

Glosa	Cantidad de respuestas
Sin información	2
No	33
Sí	15
Universo total	50
Efectividad del modelo	30.00 %

Tabla 9: Resultados de la evaluación comercial post-estudio histórico

De esta manera se deduce que la variable fuga resulta no ser confiable para evaluar la certeza del modelo, pero sí es útil para participar en el aprendizaje del modelo, además, en un problema de desbalanceo las medidas de evaluación tradicionales pueden ser no apropiadas para medir la calidad del modelo en cuestión [18].

4.4. Experimento final

El experimento presentado a continuación corresponde a los resultados de la aplicación de un piloto. Por lo tanto, resume en cada una de sus etapas los aprendizajes adquiridos a partir de los experimentos anteriores. Las bases de datos preprocesadas contempladas fueron para los períodos de Diciembre 2010, Enero 2011, Febrero 2011 y Marzo 2011, siendo esta última aquella en la que se buscaba predecir el comportamiento de fuga de servicio. Particularmente en la etapa de modelamiento se agregó la clusterización de clientes en forma separada, tal y como se ejemplifica en la figura 7. Cabe señalar que los problemas de la marginalidad detectados en las fases anteriores fueron solucionados en base a un seguimiento de las bases de datos incrementales (BNGN) para establecer la vigencia de los clientes en cada mes sin duplicar cliente como en el caso histórico, ni desestimar la cantidad de clientes fugados, de esta manera, la distribución de fugados por meses:

Número de registros por sub-bases								
Mes	Base Diciembre 2010		Base Enero 2011		Base Febrero 2011		Base Marzo 2011	
Categoría Variable	Fuga	Vig	Fuga	Vig	Fuga	Vig	Fuga	Vig
Frecuencia	52	5457	57	5707	68	5653	0	5692

Tabla 10: Distribución de fuga por bases preprocesadas para último

Donde en el último mes no se detectan fugados por el hecho de que el experimento final requiere el cierre del mes de Abril 2011 para ver los fugados

en el mes de Marzo 2011, por lo que se asumen como vigentes. En la clusterización también se observó que los grupos establecidos en cada una de las sub-bases de datos (sub-base de datos F para fugados y NF para vigentes) contenían características similares, denominados los grupos "Grandes", "Reactivos" y "Pasivos" para cada sub-base de datos. Para la elección del modelo en la etapa de minería de datos, se optó por un modelo estilo regresivo u SVM, debido a que aquellos de aprendizaje por reglas (Árboles de decisión) y probabilísticos (Bayes) no habían entregado resultados suficientemente buenos y válidos. La comparativa del modelo final con el resto se muestra a continuación:

Modelos	TN	FP	FN	TP	Accuracy	Recall clase 1	Precision clase 1	Medida F clase 1
SVM 6	4949	15	0	67	99.70 %	100.00 %	81.71 %	89.93 %
SVM 10	4948	16	2	65	99.64 %	97.01 %	80.25 %	87.84 %
SVM 3	4872	92	0	67	96.05 %	100.00 %	42.14 %	59.29 %
Naive Bayes	4666	298	1	66	94.06 %	98.51 %	18.13 %	30.63 %
LADTree	4933	31	0	67	99.38 %	100.00 %	68.37 %	81.21 %

Tabla 11: Comparativa de modelos para último experimento

En esta tabla, SVM hace referencia a las Support Vector Machines, para cada ejecución del SVM se asignan distintos parámetros, SVM 3 es aquella configuración donde el kernel es rbf (Radial basis function),$C = 0$, y $\gamma = 0{,}5$, en SVM 6 el kernel es rbf, $C = 0$, y $\gamma = 0{,}0$ y en SVM 10 el kernel es rbf, $C = 10$, y $\gamma = 0{,}95$, donde C es el parámetro que controla la compensación entre errores de entrenamiento y generalización y γ es el error de clasificar erróneamente. De estos modelos se escoge el SVM 3 (por la cantidad de fugados posibles que predice para Marzo 2011 más que por los resultados en la base de datos preprocesada de Febrero 2011). De esta manera se obtiene tras la segunda validación los siguientes resultados:

	Categorías Reales		
Categorías Predichas	Vigente	Fuga	Precisión de la clase
Vigente	5492	38	0.993 %
Fuga	157	5	0.031 %
Recall	0.972	0.116	
Accuracy	96.57 %		
Medida F Total	52.76 %		
Medida F clase 1(Fuga)	4.88 %		

Tabla 12: Tabla de confusión en el último experimento

Adicional a las métricas usuales se agrega la Curva ROC correspondiente cuyo valor de área bajo la curva es 0.975 y su gráfico es:

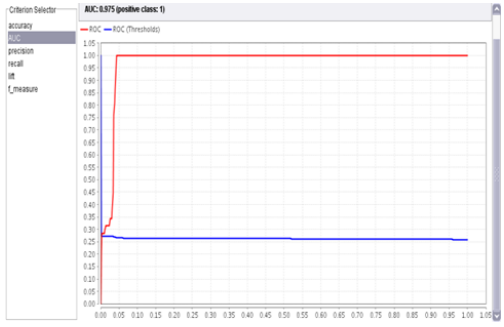

Figura 10: Gráfico de curvas ROC para modelo SVM seleccionado (gamma=0 y C=0)

Esta métrica presenta la calidad del modelo, no obstante, la tabla de confusión señala que no lo es, lo que conlleva a dudar de las métricas anteriores, esto se justifica para un problema de rarezas, en donde Weiss en [18] sugiere utilizar esta última métrica a modo de evaluación técnica, puesto que las otras no contemplan el efecto de la fuga de clientes como clase desbalanceada. Por consiguiente, se ejecuta una evaluación comercial vía telefónica, en donde los clientes objetivos fueron seleccionados a través de un muestreo estratificado por la categorización establecida por la clusterización del mes anterior (en este caso el clúster al que pertenecían en Febrero). Los resultados de esta encuesta se bosquejan posteriormente:

Descripción de ANIS	Cantidad	Porcentaje %
ANIS Vigentes	21	48.837
ANIS Sin tono	14	32.558
ANIS Contactado que declara mala atención	1	2.326
ANIS Con tono pero por sistema se retira	7	16.279
Total de Anis en la muestra	**43**	**100**
ANIS en Vigencia	21	48.837
ANIS en Retiro	22	51.163

Tabla 13: Resultados de evaluación comercial en el último experimento

Donde el concepto ANIS se refiere a número de teléfono. Esta encuesta señala que aquellos clientes que se intentó retener ya habían desconectado sus servicios, esto se debió a que el proceder de la encuesta se vio retrasada por los cambios organizacionales de la empresa en la que se desarrollaba el piloto. No obstante, declara que la certeza del modelo es cercana a un 51 %, distante del 5 % que señala la medida F como medida técnica. Es decir, se comprueba que el

error posiblemente se encuentra en la variable escogida como fuga (en este caso la variable original señala el término de contrato del producto). Con lo anterior, se puede deducir que la metodología aplicada muestra que en caso de tener el problema de rareza y, además, se tenga una alta presencia de valores fuera de rango con información útil, no son aplicables las técnicas de submuestreo, ni sobremuestreo como los demostrados en [6], y las técnicas de aplicación de pesos como las mostradas en [34], requieren de un estudio sobre la asignación para cada instancia (en este caso con alto porcentaje de valores fuera de rango), lo cual se traduce en un procedimiento de alta complejidad. También resulta infactible en un entorno de multiplataforma efectuar un procedimiento riguroso contemplando el costo que posee cada cliente fugado de un servicio particular como el propuesto en [10], puesto que los datos de facturación se encuentran dispersos en las distintas plataformas por las que se manejan los datos del cliente, a lo que se agrega la veracidad completa de dichos datos.

5. Recomendaciones

En este paper se cuenta con una experiencia previa que permitió el conocimiento de las fuentes de información, mas no así de las múltiples plataformas existentes como parte del proceso del producto de telefonía fija. En el caso de un producto cuyos datos se distribuyen entre distintas plataformas, la etapa de integración es la más relevante, por ende, se propone una serie de pasos que permiten una integración correcta:

- Investigar las variables relacionadas con el producto directamente con el área del producto.
- Conciliar dichos datos con otras variables pertenecientes a otras áreas (por ejemplo, facturación, reclamos, órdenes de trabajo, etc.)
- Averiguar si existen las variables fugas, renegociación y migración dentro del servicio, en caso de que no existan efectuar un seguimiento manual de ellas o proponer que se inserte como requerimiento informático dentro de las plataformas.
- Dado que la interpretación también es relevante en un proyecto con el KDD aplicado transversalmente se sugiere originar una descripción de las instancias que permita establecer causales de la fuga voluntaria, de tal manera de generar retenciones con valor agregado y no depender del modelo a utilizar. Una herramienta relevante para estos casos es la segmentación, debido a que bosqueja una exploración rápida del mercado

y permite una familiarización con el producto o servicio. Además, sirve como integrador entre bases de datos cuyas relaciones son de 1 a n (por ejemplo, un cliente muchos planes).

- Establecer la naturaleza de las bases de datos que cada plataforma genera, es decir, si la base de datos es marginal (dinámica) o incremental, en particular, se concluye que usar bases de datos operacionales instantáneas (no sujetas a cambio) son más útiles que usar bases de datos incrementales cuando se trata de datos transaccionales, así como también, la fuga correspondiente. Además, para llevar a cabo un proyecto de minería de datos se debe contar con dos tipos de fuentes de datos como mínimo y estas son un Data Warehouse (o algunas bases de datos incrementales) y las bases de datos marginales, donde las primeras no implican las segundas.

- Si se tiene una base de datos incremental y se desea establecer la marginalidad, este paper señala que es posible, mas el resultado presenta un comportamiento erróneo en pequeña escala agregando valores fuera de rango o con mayor porcentaje de valores perdidos lo que dificulta establecer un modelo no influenciable por la temporalidad de los datos. En otras palabras, si un Data Warehouse se implementa bajo una cultura que prescinde de otras bases de datos para darle prioridad, no tendrá la misma confiabilidad que presenta en la teoría.

En este trabajo, se descubre que la fuga de clientes del servicio, calculado en la empresa, tiene dos características relevantes: primero no es la fuga de clientes (pues un cliente en telefonía fija tiene muchos productos y servicios) y segundo, no se sabe qué clientes se van, es decir, el cálculo es global. No existe una única forma de bosquejar la variable que describe la fuga, no obstante, se entregan distintas perspectivas que pueden resultar útiles:

1. Verificar la base de datos relacionada con el call center de la empresa para observar si posee registro de fugas y analizar la cantidad de clientes que usan ese canal para terminar con el servicio de la compañía.

2. Verificar si existe una base (de datos) con los datos del contrato en toda empresa, en caso de que exista, validar la veracidad de los campos que contiene así como también, la temporalidad del mismo, si está explícita en una fecha de modificación (válida) se sugiere usar esta base de datos como principal.

3. En caso de que no se tenga registro en la empresa del cálculo de la variable fuga voluntaria individual, se aconseja calcularla en base a los registros

que desaparezcan o cambien su estado en una base de datos incremental consolidada en un período establecido para el estudio.

Las perspectivas anteriores aplican netamente a casos en los que las bases de datos trabajadas por el área encargada del producto sean de naturaleza incremental. Así como el cálculo de fuga voluntaria es clave, también lo es el establecimiento de las instancias vigentes, el cual, requiere de una estandarización o variables que declaren la certeza de su vigencia. Dentro de las consecuencias inmediatas de los errores de vigencia se encuentran las apariciones de valores perdidos y fuera de rango. Posterior a la aplicación del Preprocesamiento y la Transformación. El modelamiento plantea un desafío en cuanto a la cantidad de aprendizaje facilitado para el modelo, cada cual tiene su capacidad, pero en el mercado de las telecomunicaciones se suele dar una fuga de clientes mensual baja, lo que conlleva a la aparición del problema de rarezas o clases raras. El aplicar técnicas de muestreo simples (sub o sobre muestreo) así como también, aplicar sin muestreo conllevan a resultados no viables en el tiempo, por lo que en este paper se muestra que la segmentación válida de clientes y su seguimiento permiten mitigar este problema. La etapa de evaluación en el KDD se presenta como solamente de tipo técnica, no obstante, aparte de esta perspectiva se concluye que la vista comercial puede entregar validaciones con mayor confiabilidad que las técnicas en estos casos de sistemas con multiplataforma, puesto que tanto la variable objetivo como las instancias vigentes poseen pérdida de información al provenir de distintas bases de datos y plataformas.

6. Conclusiones

Este trabajo presenta una descripción metodológica de la aplicación de KDD para predecir la fuga de clientes en una empresa de telecomunicaciones. Por lo tanto, busca ser una ayuda a la hora de implementar un proyecto con características similares. Uno de los principales aprendizajes de este trabajo corresponde a la determinación de la integración como una de las etapas más importantes del proceso KDD para ser aplicado en empresas con información repartida en múltiples plataformas, debido a que esta etapa determina el resto de los resultados. Además, con la integración correcta aparece otra serie de problemas relacionados con variables puntuales, el trato de valores perdidos y las transformaciones respectivas. Se concluye que la clave, para llevar a cabo ambas etapas posteriores en el KDD sin error, es la variable a predecir en este caso la fuga del servicio, para lo cual se debe bosquejar la situación actual de cómo se calcula esta. Finalmente, otra de las conclusiones de este paper es que

tanto el preprocesamiento como la transformación se deben ejecutar de manera rigurosa en estas situaciones, porque el hecho de que los datos se encuentren disgregados en múltiples plataformas facilita la pérdida interna de valores para la mayoría de las variables por lo que un valor perdido no debe ser eliminado, sino que reemplazado, estimado o ignorado. Como resultado de nuestra aplicación se logró consolidar información de multiples plataformas y corroborar su calidad para ser utilizada en los modelos predictivos. Utilizando estos datos se realizaron varios experimentos entre Marzo del 2010 y Marzo 2011 para predecir la fuga de clientes, utilizando diversos algoritmos como SVM, LADTree, Naive Bayes, J48, Random Tree, entre otros. Con estos resultados fue posible construir un benchmark amplio, de tal modo de poder descubir que modelo se comporta mejor para predecir la fuga de clientes. Finalmente, se realiza un experimento tomando los RUTs de los clientes predichos como posibles fuga por el mejor modelo; y se realizó una encuesta para validar si efectivamente esto era cierto. Los resultados de la encuesta muestran que el modelo acierta en un 51 % de los RUTs predichos como fuga. Esto es un muy buen resultado pues aunque el 51 % de aciertos parece bajo para un modelo predictivo. El modelo permite que la empresa en lugar de generar acciones al azar de la base de 9000 clientes del producto NGN, simplemente se focalice en un número reducido de clientes empresas (114 en nuestro esperimento final) que el modelo predijo como posibles fugas y por ende, generar estrategias más personalizadas para aumentar su retención. Como trabajo futuro se propone la utilización de esta metodología para el cálculo de fuga de clientes con un periodo de observación mucho mayor con el fin de medir la capacidad predictiva de los modelos propuestos que solo fueron implementados a nivel de plan piloto.

Agradecimientos.

Los autores desean agradecer el continuo soporte del Instituto Sistemas Complejos de Ingeniería (ICM: P-05-004- F, CONICYT: FBO16) (www.isci.cl). Así mismo nos gustaría dar las gracias al Sr. Andrés Chacón, Sr. Eduardo Duran y a la Sra. Sandra Molina por su dispocisión para el buen desarrollo de este trabajo. Finalmente, a Don Ricardo Muñoz por su valiosa ayuda para editar la versión final de este trabajo.

Referencias

[1] J. Álvarez Menéndez. Minería de datos: Aplicaciones en el sector de las telecomunicaciones. Technical report, Universidad Carlos III, 2008.

[2] V. Barnett and T. Lewis. Outliers in Statistical Data. Wiley, 1994. 584 P.

[3] J. Pedroso, N. Murata. Support vector machines for linear programming: motivation and formulations. BSIS Technical Report, August 1999.

[4] Broadband stimulus kickstarts ICT industry growth del sitio: 2010 ict market review & forecast extraído el 29 de octubre del 2010 fuente: http://www.tiaonline.org/market intelligence/mrf/webinar/tia broadband webinar 20100319 final.pdf.

[5] B. Q. Huang, T. M. Kechadi, B. Buckley, G. Kiernan, E. Keogh, and T. Rashid. A new feature set with new window techniques for customer churn prediction in land- line telecommunications. Expert Syst. Appl., 37:3657– 3665, May 2010.

[6] J. Burez and D. Van den Poel. Handling class imbalance in customer churn prediction. Expert Syst. Appl., 36:4626–4636, April 2009.

[7] S. Maldonado, R. Weber. Modelos de Selección de Atributos para Support Vector Machines. Revista Ingeniería de Sistemas, Volumen XXVI, Septiembre 2012.

[8] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering /segmentation algorithms. Technical report, IEEE, 2003.

[9] S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. SIGMOD Rec., 26:65–74, March 1997.

[10] Y. Jie Dong, X. hua Wang, y J. Zhou, Cost BP Algorithm and its Application in Customer Churn Prediction, in 2009 Fifth International Joint Conference on INC, IMS and IDC, Seoul, South Korea, 2009, págs. 794-797.

[11] C. Bravo, S. Maldonado, R. Weber. Experiencias prácticas en la medición de microempresarios utilizando modelos de credit scoring. Revista Ingeniería de Sistemas, Volumen XXIV, Junio 2010.

[12] A. Farhangfar, L. Kurgan, and J. Dy. Impact of imputation of missing values on classification error for discrete data. Pattern Recogn., 41:3692–3705, December 2008.

[13] Estadísticas de inversión y empleo, sitio: Subtel (subsecretaría de telecomunicaciones). Extraído el 20 de octubre del 2010. fuente: http://www.subtel.cl/prontus subtel/site/artic/20100608/asocfile/20100608122246/1 series inversión y empleo dic09 191010 v1.xls.

[14] Estadísticas de reclamos recibidos por el dpto. Gestión de reclamos de la subtel. sitio: Subtel. Extraído el 20 de octubre del 2010. Fuente:http://www.subtel.gob.cl/prontus_oirs/site/artic/20100503/asocfile/20100503154918/estadisticas_reclamos_2010.pdf

[15] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth. From data mining to knowledge discovery in databases. AI Magazine, 17:37–54, 1996.

[16] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. Technical report, HP Laboratories, 2004.

[17] M. Galván and F. Medina. Imputación de datos: teoría y práctica. Technical report, CEPAL Naciones Unidas, 2007.

[18] G. M. Weiss. Mining with rarity: a unifying framework. SIGKDD Explor. Newsl., 6:7–19, June 2004.

[19] D. N. Gujarati. Econometría. McGraw Hill, 2003. 921 P.

[20] D. Hand. Data mining: Statistics and more. The American Statistician, 52:112–118, 1998.

[21] G. Huerta. Balanceo de datos para la clasificación de imágenes de galaxia. Technical report, Universidad Politécnica de Puebla, 2010.

[22] Y. Jiangsheng. Method of k-nearest neighbors. Technical report, Institute of Computational Linguistics, Peking University, 2002.

[23] J.-J. Jonker, N. Piersma, and D. V. d. Poel. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. Working papers of faculty of economics and business administration, ghent university, belgium, Ghent University, Faculty of Economics and Business Administration, 2003.

[24] N. K. Kasabov. Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. MIT Press, Cambridge, MA, USA, 1st edition, 1996. 550 P.

[25] B. Kröse and P. van der Smagt. An introduction to Neural Networks. None, 1996. 135 P.

[26] L. I. Kuncheva. Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, 2004. 350 P.

[27] J. Larsen , J. Rossat, D. Ruta, and M.Wawrzynosek. Customer loyalty, a literature review and analysis. Technical report, UNIPEDE, 1998.E. H.

[28] M. A. P. M. Lejeune. Measuring the impact of data mining on churn management. Internet Research, 11(5):375–387, 2001.

[29] J. Lévy Mangin and J. Varela Mallou. Análisis multivariable para las ciencias sociales. Prentice Hall, 2003. 896 P.

[30] R. J. A. Little and D. B. Rubin. Statistical Analysis With Missing Data. Probability and Statistics. Wiley, New Jersey, second edition, 2002. 381 P.

[31] J. Lu. Predicting customer churn in the telecommunications industry an application of survival analysis modeling using sas. Technical report, Sprint Communications Company, 2001.

[32] C. M. Luque. Clasificadores bayesianos. El algoritmo naive bayes, 2003.

[33] L. Rokach and O. Maimon. Data Mining With Decision Trees:Theory And Applications. World Scientific Publishing, 2008. 244 P.

[34] Geoffrey J. McLachlan, V. Nikulin, «Classification of Imbalanced Marketing Data with Balanced Random Sets», JMLR: Workshop and Conference Proceedings 7: 89-100, KDD Cup 2009

[35] S. Maldonado. Utilización de support vector machines no lineal y selección de atributos para credit scoring. Master's thesis, Universidad de Chile, 2007. 118 P.

[36] T. M. Mitchell. Generative and discriminative classifiers: Naive bayes and logistic regression. In Machine Learning, 2010.

[37] S. Molina. Aplicación de técnicas de minería de datos para predicción del churn de clientes en una empresa de telecomunicaciones. Master's thesis,

Escuela de Ingeniería de la Pontificia Universidad Católica de Chile, 2009. 114 P.

[38] R. Mattison. Data Warehousing and Data Mining for Telecommunications. Artech House, Inc., Norwood, MA, USA, 1st edition, 1997. 282 P.

[39] J. A. Ortega Ramírez. Patrones de comportamiento temporal en modelos semicualitativos con restricciones. PhD thesis, Universidad de Sevilla, 2000.

[40] M. Richeldi and A. Perrucci. Analyzing churn of customers.

[41] P. Ponniah. Data Warehousing Fundamentals. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 2001. 518 P.

[42] D. B. Rubin. Inference and missing data. Biometrika, 63:581–590, 1976.

[43] J. Miranda, P. Rey, R. Weber. Predicción de Fuga de Clientes para una Institución Financiera mediante Support Vector Machines. Revista Ingeniería y Sistemas, Volumen XIX, Octubre 20005.

[44] J. Wang, editor. Data mining: opportunities and challenges. IGI Publishing, Hershey, PA, USA, 2003.

[45] L. Wang. Support Vector Machines: Theory and Applications (Studies in Fuzziness and Soft Computing). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. 431 P.

Programación Matemática para asesorar a un entrenador de fútbol: un juego de fantasía como caso de estudio

F. Bonomo [*]
G. Durán [**]
F. Marenco [***]

Resumen

La película "Moneyball" dejó el mensaje de que la matemática podía ser de gran ayuda para tomar decisiones en el campo del deporte. Un club de béisbol de los Estados Unidos utilizaba herramientas matemáticas para mejorar su rendimiento deportivo. En este trabajo enfocamos este mismo problema y utilizamos como caso de estudio un juego de fantasía organizado por un diario argentino. Presentamos modelos de programación matemática para el desarrollo de un director técnico virtual de fútbol. Diseñamos modelos a priori para tener equipos más robustos para el juego, y a posteriori, para conocer cuál hubiera sido la configuración óptima de equipos durante todo el campeonato, una vez conocidos los

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

[**]Instituto de Cálculo y Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina. CONICET, Argentina. Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile, Santiago, Chile.

[***]Instituto de Ciencias, Universidad Nacional de General Sarmiento, Argentina. Departamento de Computación, FCEN, UBA, Argentina

resultados. Nuestro jugador virtual a priori se posiciona habitualmente entre los mejores de la competencia. La expansión de este tipo de desarrollos debería ser útil para asesorar en la toma de decisiones a entrenadores o gerentes de equipos en deportes reales.

Palabras Clave: Fútbol, Juego de Fantasía, Programación Matemática, Sports Analytics.

1. Introducción

El "Gran DT" es un juego de fantasía creado por un diario argentino en los '90 y que retomó su acción en 2008. El juego ha contado desde su reinicio con más de 1 millón de participantes en cada una de sus ediciones. Consiste en que cada uno de los participantes toma el rol de un director técnico de fútbol a lo largo del campeonato argentino de Primera División, con el objetivo de formar el mejor equipo posible, combinando a jugadores de los diferentes equipos que participan del campeonato. Cada participante suma puntos por la actuación semanal de sus jugadores, existiendo datos objetivos (goles convertidos, valla invicta, tarjetas amarillas, tarjetas rojas) y datos subjetivos aportados por el diario (puntaje sobre la actuación de cada jugador en el partido, figura de la cancha).

El reglamento del juego exige que los equipos virtuales cumplan una serie de restricciones (presupuesto máximo, número de jugadores por puesto, número de jugadores por equipo), y su constitución es dinámica dado que fecha a fecha se pueden hacer una cierta cantidad de modificaciones en el equipo.

El "Gran DT" se inspiró en el "Fantacalcio" [10], juego de fantasía italiano dedicado a su Serie A de fútbol. El "Fantacalcio" fue inventado por Riccardo Albini [12] a fines de los '80 y es también organizado por uno de los principales diarios de Italia. Según relata Albini, se inspiró para la creación del "Fantacalcio" en el juego de fantasía de la Liga de Béisbol de los Estados Unidos [14]. La reglamentación del juego italiano es muy similar a la que terminó adoptando el "Gran DT" en la Argentina. Actualmente existen también otros juegos de fantasía en diferentes lugares del mundo, como el basado en la Premier League inglesa de fútbol [9], o el de la NBA o el fútbol americano en los Estados Unidos de América [15, 16]. Existen también otros juegos similares con mucha aceptación en

el público, pero sobre fútbol virtual [13, 17]. El uso de juegos de fantasía para incentivar la enseñanza de la matemática en la escuela ha sido ampliamente estudiado en los Estados Unidos [21].

En este trabajo presentamos un modelo de programación matemática diseñado a priori, que llamamos prescriptivo, para tener un equipo más robusto para el juego (teniendo en cuenta las actuaciones de los jugadores en torneos anteriores y en el mismo torneo, más ciertas características de las fechas a disputarse); y dos modelos diseñados a posteriori, que llamamos descriptivos, para conocer cuál hubiera sido el equipo óptimo a lo largo del torneo, una vez conocidos los resultados.

Los modelos descriptivos consiguen mostrar el equipo ideal que se debió haber armado fecha a fecha, a lo largo de todo el torneo, para obtener el mayor puntaje posible, cumpliendo con las restricciones del juego.

Testeamos el modelo prescriptivo haciendo competir a nuestro jugador virtual en el juego. Los resultados marcan que el mismo califica habitualmente en el 3 % mejor del juego, llegando en alguna oportunidad a estar en el mejor 1 por 1000 del certamen. Si tomamos los 6 campeonatos en los que el modelo participó del juego como un único torneo, nuestro jugador virtual se posiciona en el mejor 2 por 1000 de la competición.

La literatura orientada a la disciplina conocida como "Sports Analytics" (SA), que engloba desarrollos matemáticos y computacionales orientados a problemas del deporte, ha aumentado notablemente en los últimos años. Importantes revistas de investigación operativa, matemática aplicada, estadística, gestión y economía han publicado numerosos artículos en estos temas [1]. En lo que hace a métodos de programación matemática aplicados a cuestiones deportivas, el mayor desarrollo se ha dado en problemas de confección de fixtures para diferentes competencias. Esta subdisciplina del SA se la conoce como "Sports Scheduling" (SS). Excelentes resúmenes sobre el estado del arte en SS y análisis de distintos problemas abiertos en el tema se encuentran en [2, 7]. Un análisis de las principales instancias para distintos deportes estudiadas en la literatura aparece en [6].

A pesar de que en los últimos años ha surgido software que asiste a entrenadores en diferentes deportes en la tarea de recopilar, almacenar y consultar datos, hasta lo mejor de nuestro conocimiento, no existen aplicaciones de algoritmos de programación matemática para asesoramiento

a directores técnicos, reales o virtuales,del estilo de lo que se presenta en este trabajo. Lo más cercano podría ser el "Fantarobot" [11],una funcionalidad opcional en la página web del "Fantacalcio",que elige el "mejor" equipo dentro del plantel de titulares y suplentes que un determinado jugador seleccionó, pero tenemos entendido que sólo considera la actuación promedio de cada jugador hasta ese momento, a fin de dar una recomendación. En lo que hace a aplicación de técnicas matemáticas para asesorar en la toma de decisiones en deportes reales, el caso más conocido se ha dado en el béisbol de los Estados Unidos [3], donde se aplican fundamentalmente herramientas estadísticas.

Este trabajo está organizado de la siguiente manera. En la Sección 2 se describen en detalle las características del juego. En la Sección 3 se presentan dos modelos descriptivos, modelos de programación lineal entera que encuentran el equipo ideal a lo largo de todo el torneo, si se pudiera prepararlo en cada fecha conociendo los resultados de todo el campeonato. Se muestran allá los resultados alcanzados en distintos torneos. En la Sección 4 se presenta el modelo prescriptivo, que incluye un modelo de programación lineal entera que busca maximizar el puntaje del equipo, una vez que fueron estimados los puntajes de cada jugador en cada fecha. Se exhiben en esta sección resultados obtenidos por el jugador virtual en diferentes torneos. Por último, en la Sección 5 se presentan las conclusiones y el posible trabajo futuro.

2. Descripción del juego

El juego empieza en la fecha 4 o 5 del Torneo Argentino de Primera División (habitualmente en los torneos Clausura empieza en la fecha 5 y en los Apertura, en la 4). De este modo, el juego se desarrolla a través de 15 o 16 fechas (dado que todos los torneos los disputan 20 equipos en un sistema de todos contra todos, a lo largo de 19 fechas).

Cada competidor participa del juego con su número de documento nacional y debe armar un equipo integrado por jugadores de la Primera División del fútbol argentino, compuesto por 11 titulares y 4 suplentes.

Dentro de los 11 titulares, se pueden armar 3 tácticas de juego diferentes: 1 arquero, 4 defensores, 4 volantes y 2 delanteros; 1 arquero, 4 defensores, 3 volantes y 3 delanteros; o 1 arquero, 3 defensores, 4 volantes

y 3 delanteros. Los 4 suplentes son un arquero, un defensor, un volante y un delantero. Los suplentes sólo son considerados para el juego en una determinada fecha si alguno de los titulares de su misma posición no juega, o juega menos de 20 minutos, y por lo tanto no recibe puntuación por su actuación de parte del diario.

Cada jugador tiene un valor monetario simbólico, que va de los $300.000 (los que aún no debutaron en primera), hasta más de $10.000.000 (los jugadores de mejor nivel de los torneos anteriores). Los equipos no pueden superar los 65 millones de pesos (este valor ha variado entre 60 y 70 millones a lo largo de los torneos). No puede haber más que 3 jugadores del mismo club en un equipo.

Cada jugador titular suma o resta puntos en cada fecha por criterios subjetivos (la nota que le pone el diario, o si es catalogado como la figura de la cancha) y objetivos (si convierte goles; si no le convierten, en el caso de arqueros o defensores; si es expulsado; si es amonestado). La nota que el diario le pone a cada jugador es un valor numérico entero entre 1 y 10. Por convertir un gol en jugada de campo, un arquero recibe 10 puntos extras; un defensor, 9 puntos; un volante, 6 puntos y un delantero, 4 puntos. Un gol de penal le aporta 3 puntos al jugador que lo convierte, cualquiera sea su posición en la cancha. La figura de la cancha, determinada por el diario, recibe 4 puntos extra. Un arquero al que no le hacen goles recibe 3 puntos extra, y un defensor que termina el partido con la valla invicta, recibe 2 puntos extra. Un arquero que recibe goles, resta tantos puntos como la cantidad de goles que le hicieron; mientras que un arquero que ataja un penal suma 4 puntos extra (puntos que le son descontados al jugador que falló el penal). Una tarjeta amarilla resta 2 puntos, y una tarjeta roja resta 4 puntos. El jugador que juega menos de 20 minutos se lo considera como que no participó de esa fecha, y si es titular en el equipo del juego, le deja su lugar al suplente de la misma posición. En caso que haya más de un titular de una misma posición que no participa de la fecha, el equipo en cuestión juega con menos de 11 jugadores.

Fecha a fecha se pueden cambiar titulares por suplentes de manera ilimitada y se pueden realizar hasta 4 transferencias, reemplazos donde se incorpora a un jugador que no pertenecía al equipo y se retira a uno que sí pertenecía, manteniendo el equipo las restricciones básicas en cuanto a su conformación y su presupuesto (este parámetro fue modificado, dado

que en las primeras ediciones del juego se admitían hasta 3 transferencias por fecha).

El juego fue presentado en los '90, época en que se realizaron 3 ediciones del mismo. En agosto de 2008 fue reiniciado, y desde entonces fue jugado de manera ininterrumpida en los 9 campeonatos de Primera División que se han disputado hasta la fecha. La máxima cantidad de participantes se dio en el primer semestre de 2009, con casi 2 millones (cerca del 5 % de la población de Argentina), mientras que la más baja se dio en el primer semestre de 2012, con un poco más de 1 millón de participantes.

3. Modelos descriptivos

Los modelos "a posteriori", o descriptivos, se ejecutan al finalizar el campeonato (también podrían ejecutarse después de cada fecha), y permiten encontrar una configuración de equipos óptima una vez que se conocen los resultados (es decir, toma como dato los puntajes obtenidos por cada jugador en cada fecha).

Desarrollamos dos modelos descriptivos. El primero arma un equipo fijo con 11 titulares, cumpliendo con las restricciones del juego, y no lo modifica a lo largo de todo el torneo. No considera a los jugadores suplentes para nada, sólo guarda $1.200.000 para colocar 4 jugadores suplentes de $300.000 cada uno, de modo que el equipo sea válido. Busca maximizar el puntaje a lo largo de todo el torneo considerando sólo a esos 11 jugadores. El segundo modelo arma lo que llamamos el "equipo perfecto". Empieza por un equipo inicial en la primer fecha del juego, y nos dice fecha a fecha que intercambios entre titulares y suplentes debe hacerse, y que jugadores deben incorporarse al equipo, de modo de obtener el mayor puntaje final posible.

Formulamos ambos modelos a través de programación lineal entera en el que la función objetivo maximiza el puntaje total, y las restricciones garantizan que fecha a fecha se cumple con las condiciones del juego, en cuanto a conformación del equipo, número de transferencias permitidas y presupuesto máximo.

3.1. Equipo fijo sólo con titulares: formulación matemática

Este modelo busca los 11 jugadores fijos que maximizan el puntaje total, cumpliendo las restricciones del juego. Llamamos E al conjunto de equipos, J al conjunto de jugadores y $P = \{$arquero, defensor, volante, delantero$\}$ a las posiciones dentro del campo de juego. Para formular el modelo, para cada jugador $j \in J$ introducimos la variable binaria x_j que vale 1 si y sólo si el jugador j se incluye en el equipo. Los parámetros del modelo incluyen los siguientes datos:

- Para cada jugador $j \in J$, el parámetro $\text{equipo}_j \in E$ representa el equipo al que pertenece el jugador j, el parámetro $\text{posicion}_j \in P$ especifica la posición del jugador j dentro de la cancha, el parámetro $\text{precio}_j \in \mathbb{R}_+$ representa el precio del jugador, y finalmente $\text{puntaje}_j \in \mathbb{Z}$ informa el puntaje acumulado total del jugador j a lo largo del campeonato.
- Para cada posición $p \in P$, los parámetros máx_p y mín_p especifican la cantidad mínima y máxima de jugadores en esa posición, de acuerdo con las tres estrategias de juego permitidas.

Con estas definiciones, el modelo se plantea del siguiente modo:

$$\begin{aligned}
\text{máx} \quad & \sum_{j \in J} \text{puntaje}_j x_j & & & \\
\sum_{j \in J} \text{precio}_j x_j &\leq 63{,}800{,}000 & & & (1)\\
\sum_{j \in J:\text{equipo}_j = e} x_j &\leq 3 & \forall e \in E & & (2)\\
\sum_{j \in J} x_j &= 11 & & & (3)\\
\sum_{j \in J:\text{posicion}_j = p} x_j &\geq \text{min}_p & \forall p \in P & & (4)\\
\sum_{j \in J:\text{posicion}_j = p} x_j &\leq \text{max}_p & \forall p \in P & & (5)\\
x_j &\in \{0,1\} & \forall j \in J & & (6)
\end{aligned}$$

La función objetivo busca maximizar el puntaje total de los jugadores. Las restricciones (1) solicitan que el monto total desembolsado por los jugadores no exceda \$63.800.000 (lo cual corresponde al límite superior de \$65.000.000, descontando cuatro suplentes de \$300.000, el menor valor posible). Las restricciones (2) impiden que haya más de tres jugadores de un mismo equipo. La restricción (3) pide que el equipo se componga de 11 jugadores titulares, mientras que las restricciones (4) y (5) piden que se respeten las cantidades mínima y máxima de jugadores por posición (notar que estas restricciones, más el hecho de pedir que el equipo está compuesto por 11 jugadores, garantizan que se elige alguna de las tres estrategias permitidas). Finalmente, las restricciones (6) especifican la naturaleza de las variables.

Es interesante analizar la estructura de este modelo. Las restricciones (1) son restricciones de tipo *knapsack*, una estructura que se encuentra muy estudiada en los paquetes comerciales para resolver modelos de programación entera. El resto de las restricciones puede complicar un poco la resolución –particularmente las restricciones (4) y (5), que imponen un rango de jugadores por posición–, pero globalmente el modelo es una versión levemente complicada del problema de la mochila. Esto sugiere que en la práctica la resolución computacional del modelo puede no ser complicada. Efectivamente comprobamos que esta cuestión estructural, sumado a que la instancia considerada es pequeña, da como resultado tiempos de resolución muy bajos para los casos considerados en este trabajo.

3.2. Equipo Perfecto: formulación matemática

Este segundo modelo busca la combinación óptima del equipo inicial y las modificaciones a realizar en cada fecha, de modo tal de sumar la mayor cantidad posible de puntos. Nuevamente, llamamos E al conjunto de equipos, J al conjunto de jugadores y P a las posiciones dentro del campo de juego. Además, definimos F como el conjunto de fechas, y llamamos $F' = F \backslash \{\text{mín}(F)\}$ al conjunto de todas las fechas excepto la primera (necesitamos definir este subconjunto de fechas dado que las modificaciones al equipo se hacen a partir de la segunda fecha del juego). Los parámetros del modelo son los mismos que en el modelo anterior, a excepción del puntaje: para cada jugador $j \in J$ y cada fecha $k \in F$, el parámetro $\text{puntaje}_{jk} \in \mathbb{Z}$ especifica el puntaje que obtuvo el jugador j

en la fecha k.

Para cada jugador $j \in J$ y cada fecha $k \in F$, introducimos las variables binarias x_{jk}, que informan si en la fecha k el jugador j es titular, e y_{jk}, que informan si en la fecha k el jugador j es suplente. Además, para $j \in J$ y $k \in F'$ introducimos la variable binaria z_{jk}, de modo tal que $z_{jk} = 1$ si y sólo si el jugador j se incorpora al equipo a partir de la fecha k. Con estas definiciones, el modelo es el siguiente:

$$\begin{aligned}
\text{máx} \quad & \sum_{j,k \in J \times F} \text{puntaje}_{jk} x_{jk} & & \\
x_{jk} + y_{jk} &\leq 1 & \forall j,k \in J \times F & \quad (7)\\
\sum_{j \in J} x_{jk} &= 11 & \forall k \in F & \quad (8)\\
\sum_{j \in J : \text{posicion}_j = p} y_{jk} &= 1 & \forall p,k \in P \times F & \quad (9)\\
\sum_{j \in J} \text{precio}_j (x_{jk} + y_{jk}) &\leq 65{,}000{,}000 & \forall k \in F & \quad (10)\\
\sum_{j \in J : \text{equipo}_j = e} x_{jk} + y_{jk} &\leq 3 & \forall e,k \in E \times F & \quad (11)\\
\sum_{j \in J : \text{posicion}_j = p} x_{jk} &\geq \text{min}_p & \forall p,k \in P \times F & \quad (12)\\
\sum_{j \in J : \text{posicion}_j = p} x_{jk} &\leq \text{max}_p & \forall p,k \in P \times F & \quad (13)\\
x_{jk} + y_{jk} - x_{j,k-1} - y_{j,k-1} &\leq z_{jk} & \forall j,k \in J \times F' & \quad (14)\\
x_{jk} + y_{jk} &\geq z_{jk} & \forall j,k \in J \times F' & \quad (15)\\
1 - (x_{j,k-1} + y_{j,k-1}) &\geq z_{jk} & \forall j,k \in J \times F' & \quad (16)\\
\sum_{j \in J} z_{jk} &\leq 4 & \forall k \in F' & \quad (17)\\
x_{jk}, y_{ik} &\in \{0,1\} & \forall j,k \in J \times F & \quad (18)\\
z_{jk} &\in \{0,1\} & \forall j,k \in J \times F' & \quad (19)
\end{aligned}$$

Nuevamente, la función objetivo solicita maximizar el puntaje total obtenido por los jugadores titulares del equipo a lo largo de todo el torneo (dado que podemos intercambiar titulares y suplentes sin límite no se pierde generalidad considerando sólo a los titulares). Las restricciones (7) especifican que cada jugador puede ser titular o suplente, o bien no

estar seleccionado para el equipo en cada fecha, y las restricciones (8) piden que el equipo tenga exactamente 11 jugadores titulares. Además, las restricciones (9) solicitan exactamente un suplente por posición, totalizando así los cuatro suplentes del equipo. Las restricciones (10) y (11) especifican los límites de presupuesto total y cantidad máxima de jugadores por equipo, respectivamente. Por su parte, las restricciones (12) y (13) imponen los límites inferior y superior a la cantidad de jugadores en cada posición dentro del campo de juego.

Las restricciones (14), (15) y (16) relacionan las variables x e y con las variables z, de modo tal que $z_{jk} = 1$ si y sólo si el jugador j se incorpora al equipo (como titular o suplente) en la fecha k. Esta definición les permite a las restricciones (17) limitar un máximo de 4 incorporaciones en cada fecha. Finalmente, las restricciones (18) y (19) especifican la naturaleza de las variables.

Este modelo es sensiblemente más complicado que el anterior, dado que incorpora un número máximo de transferencias entre cada fecha y la siguiente. En cada fecha se sigue teniendo la estructura de *knapsack* del modelo anterior, pero la introducción de las variables z para representar los jugadores transferidos y de las restricciones (17) que imponen un máximo a estas transferencias, hace que la resolución computacional se complique. En sintonía con estas observaciones, más el hecho de que las instancias resueltas ahora son de tamaño mediano, los tiempos de resolución de este modelo son más altos que para el modelo anterior.

3.3. Resultados

En el Cuadro 1 se muestran los resultados obtenidos por los 2 modelos en los cuatro torneos disputados entre los años 2009 y 2010, y los tiempos de corrida de cada uno de ellos. Se exhibe también el puntaje del ganador del juego.

Cabe notar que en los cuatro torneos se da prácticamente un empate entre el ganador del juego y el modelo del equipo fijo con sólo titulares (el modelo gana por muy poco en 3 de los 4 torneos, y pierde por muy poco en el restante). Estos resultados muestran que la actuación del ganador del juego en todos los casos es muy meritoria ya que obviamente compite sin conocer los resultados que se darán a posteriori. Claramente los buenos jugadores aprovechan la naturaleza dinámica del equipo para ir mejorándolo fecha a fecha, y de este modo competir a la par con un

modelo que deja fijo el equipo a lo largo de todo el torneo, pero que juega con los resultados conocidos.

El Equipo Perfecto, que marca una cota superior en puntaje a lo que cualquier jugador puede alcanzar, supera al ganador del juego con puntajes que están entre un 50 % y un 70 % por encima, según el campeonato. Esta gran diferencia se debe fundamentalmente a que el Equipo Perfecto captura a jugadores que tienen grandes actuaciones de manera bien esporádica, lo que habitualmente no es encontrado ni siquiera por jugadores expertos.

La diferencia de puntajes entre los torneos Clausura y los torneos Apertura se explica en que en estos últimos el juego comienza una fecha antes.

Los resultados obtenidos por nuestros modelos fueron publicados en el diario en diversas oportunidades [18, 19, 20].

En lo que respecta a los tiempos de corrida, el primer modelo corre muy rápidamente, mientras que la obtención del Equipo Perfecto puede demorarse hasta 2 horas. El modelo del equipo fijo tiene alrededor de 500 variables (el número de jugadores total del campeonato) y poco más de 20 restricciones. El modelo del Equipo Perfecto tiene alrededor de 2000 variables y 3500 restricciones. En todos los casos los modelos se resuelven a optimalidad en los tiempos reportados en el Cuadro 1. Los experimentos se realizaron con Cplex 12.2 en una PC con 2 GB de memoria RAM y un procesador con una frecuencia de 1.6 GHz.

Torneo	Ganador GDT	Equipo Titulares	Tiempo Corrida	Equipo Perfecto	Tiempo Corrida
Cl. 09	1279	1318	2 seg	1990	10 min
Ap. 09	1375	1336	1 seg	2173	120 min
Cl. 10	1227	1232	1 seg	2027	50 min
Ap. 10	1394	1412	2 seg	2289	116 min

Tabla 1: Resultados de los modelos descriptivos en los torneos de 2009 y 2010

4. Modelo prescriptivo

El modelo "a priori", o prescriptivo, presenta un mayor desafío, dado que ahora debemos encontrar equipos óptimos sin conocer el desempeño futuro de cada jugador. Para ello, confeccionamos un índice para cada jugador, que es una predicción del puntaje que va a alcanzar en la fecha siguiente. Una vez obtenidos dichos índices, corremos un modelo similar al de la sección anterior, donde ahora en vez de tener los puntajes de cada jugador tenemos los índices. La pregunta es cómo armar dichos índices, de modo de que sean una representación razonable de lo que irá a pasar en la realidad.

Tras algunas pruebas iniciales, llegamos a la conclusión de que el promedio de puntos que cada jugador sacó en las últimas fechas no es un buen predictor del puntaje que el mismo jugador sacará el siguiente fin de semana, ya que no tiene en cuenta consideraciones claves como el rival al que va a enfrentar, si es local o visitante, la actualidad de su equipo, etc.

Decidimos entonces construir el índice considerando el puntaje de cada jugador en las fechas que se han jugado hasta ese momento en el torneo en cuestión, pero ponderando dicho promedio por 3 factores: la condición de local o visita (usamos un ponderador de 1,05 para los locales, y de 0,95 para los visitas), la posición en la tabla del rival que va a enfrentar (ponderamos entre 1 y 1,05 si se enfrenta a los últimos de la tabla, y entre 0,95 y 1 si se enfrenta a los primeros), y la actualidad del jugador o de su equipo (hasta un 5 % más a jugadores o equipos que vienen en buenas rachas, y hasta un 5 % menos a jugadores o equipos que vienen en malas rachas, o que llegan cansados por venir por ejemplo de una disputa cercana en un torneo internacional). Los promedios obtenidos por dicho jugador en los 2 últimos torneos se consideran como si fueran una actuación más en una fecha del torneo actual.

Por último, hay un factor más de ponderación en cada índice, que llamamos "juega o no juega", y que es un 1 para aquellos jugadores que se anuncian en la prensa o por parte de los directores técnicos como titulares en la fecha que va a venir, y un 0 para el resto. De esta manera, tratamos de garantizar que nuestros 11 titulares disputen la fecha. Este

es un dato crucial, y que a veces no se conoce a priori. Hay que tener en cuenta que la constitución de los equipos del juego para cada fecha cierra una hora antes del inicio del primer partido. Como las fechas se suelen jugar entre viernes y lunes, hay veces que uno puede no conocer el viernes la formación titular que un equipo tendrá el día domingo o el día lunes. Ahí entran a tener importancia los jugadores suplentes, por ello también es aconsejable tener "buenos" jugadores suplentes, aunque puede ser importante no gastar un gran presupuesto en ellos porque en la mayoría de los casos no serán utilizados. Utilizamos con esta idea un ponderador de 0,1 para el índice de los jugadores suplentes en la función objetivo, valor que fijamos tras realizar previamente algunas pruebas. Este ponderador podría ajustarse haciendo un estudio estadístico de en cuántas oportunidades los jugadores suplentes terminan participando efectivamente del equipo, o incluso corriendo el modelo a posteriori para varios campeonatos con diferentes ponderadores para deducir cuál entrega mejores resultados habitualmente.

Notar que si un jugador jugó menos que k partidos en el torneo actual (y solemos trabajar con $k = 3$), no se lo considera como potencial candidato a formar parte del equipo por más que en la fecha siguiente se lo vuelva a anunciar como titular (para que el modelo no se vea tentado a poner a un jugador que tuvo una gran actuación pero después no se consolidó como titular en su equipo).

El modelo prescriptivo tiene entonces dos etapas: la constitución del equipo inicial y las modificaciones sugeridas fecha a fecha.

4.1. El equipo inicial

Una vez que tenemos definidos los índices de cada jugador, preparamos un equipo inicial que intenta sacar el mayor puntaje posible en la primer fecha del juego. Nuevamente, el conjunto E representa a los equipos, el conjunto J corresponde a los jugadores y P representa las cuatro posiciones en el campo de juego. Los parámetros son los mismos que para los modelos descriptivos, a excepción del parámetro $\text{indice}_j \in \mathbb{R}$ asociado con cada jugador $j \in J$, que representa el índice descrito más arriba. Para cada jugador $j \in J$, introducimos una variable binaria x_j que vale $x_j = 1$ si y sólo si el jugador j es titular en el equipo inicial, y una variable binaria y_j que vale $y_j = 1$ si y sólo si el jugador j es suplente en el equipo inicial. Con estas definiciones, el modelo es el siguiente:

$$
\begin{aligned}
\text{máx} \quad & \sum_{j \in J} \text{indice}_j x_j + 0{,}1\ \text{indice}_j y_j & & & \\
x_j + y_j &\leq 1 & \forall j \in J & & (20)\\
\sum_{j \in J} \text{precio}_j (x_j + y_j) &\leq 65{,}000{,}000 & & & (21)\\
\sum_{j \in J:\text{equipo}_j = e} x_j + y_j &\leq 3 & \forall e \in E & & (22)\\
\sum_{j \in J:\text{posicion}_j = p} x_j &\leq \max_p & \forall p \in P & & (23)\\
\sum_{j \in J:\text{posicion}_j = p} x_j &\geq \min_p & \forall p \in P & & (24)\\
\sum_{j \in J} x_j &= 11 & & & (25)\\
\sum_{j \in J:\text{posicion}_j = p} y_j &= 1 & \forall p \in P & & (26)\\
x_j, y_j &\in \{0,1\} & \forall j \in J & & (27)
\end{aligned}
$$

La función objetivo busca maximizar el índice total del equipo, ponderando con un 10 % a los suplentes. Las restricciones (20) piden que cada jugador sea titular o suplente, o que no esté en el equipo, mientras que las restricciones (21)-(26) establecen las condiciones que el equipo debe cumplir dadas por las reglas del juego. Finalmente, las restricciones (27) especifican la naturaleza de las variables.

4.2. Cambios y transferencias

A partir de la segunda fecha del juego se puede ir actualizando el equipo, intercambiando titulares por suplentes y realizando hasta las 4 transferencias que el juego permite. Obviamente el índice de cada jugador se actualiza con lo ocurrido en la fecha que pasó, sumado a las características de la fecha que viene (local, actualidad, rival, juega o no juega).

El modelo para hacer las modificaciones en el equipo fecha a fecha utiliza los mismos conjuntos, parámetros y variables que el modelo para determinar el equipo inicial. Para cada jugador $j \in J$, se tienen las

variables binarias x_j e y_j, que representan si el jugador j es titular o suplente en el equipo, respectivamente. Además, definimos $A \subseteq J$ como el equipo actual, sobre el cual se realizarán las modificaciones. Con estas definiciones, el modelo es el siguiente:

$$\begin{aligned}
\text{máx} \quad & \sum_{j \in J} \text{indice}_j x_j + 0{,}1\, \text{indice}_j y_j & & & \\
x_j + y_j &\leq 1 & \forall j \in J & & (28)\\
\sum_{j \in J} \text{precio}_j (x_j + y_j) &\leq 65{,}000{,}000 & & & (29)\\
\sum_{j \in J : \text{equipo}_j = e} x_j + y_j &\leq 3 & \forall e \in E & & (30)\\
\sum_{j \in J : \text{posicion}_j = p} x_j &\leq \max_p & \forall p \in P & & (31)\\
\sum_{j \in J : \text{posicion}_j = p} x_j &\geq \min_p & \forall p \in P & & (32)\\
\sum_{j \in J} x_j &= 11 & & & (33)\\
\sum_{j \in J : \text{posicion}_j = p} y_j &= 1 & \forall p \in P & & (34)\\
\sum_{j \in A} x_j + y_j &\geq 11 & & & (35)\\
x_j, y_j &\in \{0,1\} & \forall j \in J & & (36)
\end{aligned}$$

El modelo es similar al presentado en la sección anterior, buscando maximizar el índice del nuevo equipo (otra vez ponderando con un 10 % a los suplentes) y respetando las restricciones impuestas por el juego. Como única diferencia, la restricción (35) solicita que el nuevo equipo tenga al menos 11 jugadores que estaban presentes en el equipo anterior (lo cual corresponde a haber realizado a lo sumo 4 transferencias).

Tanto este modelo como el de la subsección anterior cuentan con alrededor de 1000 variables y 500 restricciones, y se ejecutan a optimalidad en un par de segundos.

4.3. Resultados

Analizamos la actuación de nuestro modelo en los dos torneos del año 2010, Clausura y Apertura (en ese orden, dado que en la Argentina en la primera mitad del año se juega el torneo Clausura y en la segunda mitad, el Apertura). En los Cuadros 2 y 3 se puede ver fecha a fecha la predicción global de puntaje para el equipo elegido por el modelo y el puntaje real que obtuvo, para cada uno de los dos torneos analizados. Como era de esperar, dado que el modelo selecciona a los mejores jugadores hasta el momento y la variabilidad de los rendimientos de los jugadores de fútbol es bastante amplia, la predicción de puntajes suele estar por encima del puntaje real obtenido. Para confirmar esta idea durante el Clausura 2010 también armamos un equipo "random", es decir, un equipo conformado por jugadores elegidos al azar, con la única condición de que fecha a fecha los 11 titulares participaran de los partidos del fin de semana. El equipo random sacó 899 puntos, con una predicción de puntaje de 934 puntos, es decir el puntaje real se acerca mucho más a la predicción: menos de un 4 % de diferencia (contra un 21 % y un 14 % en los equipos generados por el modelo, para el Clausura y el Apertura de 2010, respectivamente).

En el Clausura 2010 participaron del juego 1.442.682 jugadores. El ganador del juego sacó 1227 puntos. Nuestro modelo, con 1070 puntos se ubicó en la posición 13.547, quedando en el mejor 1 % de la competencia. El equipo random, con 899 puntos, terminó en la posición 498.726. La posición del equipo random da también una idea de cuántos equipos "activos" existen (consideramos activo a un equipo que se va actualizando fecha a fecha), alrededor de 1 millón, ya que es esperable que el equipo random finalice por la mitad de la tabla de posiciones, dentro de los equipos activos. Un equipo no activo suele terminar el torneo participando con menos que 11 jugadores en cada fecha, ya que no logra reemplazar a aquellos jugadores que a lo largo del torneo se fueron lesionando o perdiendo la titularidad. La estimación de 1 millón de equipos activos coincide con la estimación de los organizadores del juego, que tienen observado que 2/3 de los equipos se actualizan fecha a fecha. Si comparamos entonces la actuación de nuestro modelo contra el total de los equipos activos, tenemos que el mismo se ubica en el mejor 1,5 % de la competencia.

En el Apertura 2010 participaron del juego 1.445.531 jugadores. El ganador del juego sacó 1394 puntos. Nuestro modelo, con 1322 puntos se

Fecha	Predicción	Puntaje Real
5	97,84	48
6	90,47	76
7	90,06	59
8	86,40	63
9	86,84	67
10	95,64	77
11	87,05	81
12	91,27	81
13	90,12	88
14	94,81	69
15	86,51	69
16	90,31	73
17	91,01	72
18	91,72	59
19	83,68	88
Total	1353,73	1070

Tabla 2: Resultados del modelo prescriptivo en el Clausura 2010

ubicó en la posición 643, quedando en el mejor 1 por mil de la competencia, incluso considerando sólo a los equipos activos. Esta fue claramente la mejor actuación del modelo, considerando los 6 campeonatos en los que participó.

Cabe destacar que en ambos casos, la predicción de puntaje global superó incluso al ganador del juego, pero como ya mencionamos, no era esperable conseguir en nuestro modelo un puntaje similar a dicha predicción.

Una posible explicación de por qué el modelo suele tener mejores actuaciones en el segundo torneo del año que en el primero es la existencia de la Copa Libertadores en el primer semestre de cada año. Los mejores equipos disputan esta Copa, y por lo tanto los mejores jugadores lo hacen. Esto hace que los mejores jugadores jueguen a veces cansados en el torneo local, o directamente no jueguen. También suele pasar que los entrenadores anuncien su formación inicial horas antes de los partidos. Todo esto complica la constitución del equipo virtual, dado que genera una mayor incertidumbre, y puede afectar a los resultados finales.

Fecha	Predicción	Puntaje Real
4	89,85	91
5	101,77	77
6	92,64	105
7	97,07	95
8	93,9	65
9	89,6	83
10	102,44	82
11	95,85	97
12	95,01	84
13	89,25	88
14	100,34	55
15	101,16	90
16	95,58	80
17	96,73	78
18	98,03	67
19	92,56	85
Total	1531,19	1322

Tabla 3: Resultados del modelo prescriptivo en el Apertura 2010

Un dato adicional es que si consideramos a los 6 campeonatos en los que el modelo participó del juego como un único campeonato, hay 343.017 competidores que participaron en todos ellos, estando nuestro modelo en la posición 530. Es decir, se ubica en el mejor 2 por mil de la competencia.

5. Conclusiones y trabajo futuro

El objetivo de este trabajo es analizar el aporte de la programación matemática a la hora de diseñar un entrenador virtual, o asesorar a un entrenador deportivo real. Utilizamos como caso de estudio un juego de fantasía realizado en el marco del torneo argentino de fútbol.

Presentamos modelos de programación matemática diseñados a priori, que llamamos prescriptivos, y diseñados a posteriori, que llamamos descriptivos, en la búsqueda de conseguir equipos óptimos para este juego

organizado por un diario argentino. El juego moviliza a más de un millón de jugadores en cada edición.

Los modelos descriptivos aquí desarrollados consiguen mostrar el equipo ideal que se debió haber armado a lo largo de todo el torneo, para obtener el mayor puntaje posible, cumpliendo con las restricciones del juego.

El modelo prescriptivo utiliza datos históricos y características de la próxima fecha a disputar, a fin de armar un equipo competitivo para el juego. Lo testeamos haciéndolo participar de la competencia. Los resultados marcan que nuestro modelo se posiciona habitualmente en el 3 % mejor del juego, llegando en una oportunidad a estar en el mejor 1 por 1000 del certamen. Tomando los 6 campeonatos en los que el modelo participó del juego como un único torneo, nuestro jugador virtual se posiciona en el mejor 2 por 1000 de la competición. Era esperable que cuánto más largo sea el torneo, más chances hay de que nuestras herramientas de estadística y optimización funcionen mejor.

El modelo prescriptivo desarrollado fue usado siguiendo sus indicaciones al 100 %. Se podría pensar al modelo como asistente de un jugador experto. Por ejemplo, proponiendo los k mejores conjuntos de transferencias para el equipo en cada fecha y que el experto elija uno de ellos. Esto se puede hacer corriendo k veces el modelo que sugiere las modificaciones, prohibiendo en cada caso las soluciones óptimas que se van obteniendo. O decidiendo de manera externa al modelo la inclusión o exclusión de un determinado jugador, y corriendo luego el modelo de cambios y transferencias para determinar el resto de las modificaciones.

En cuanto a qué se podría hacer para intentar mejorar a nuestro jugador virtual, surgen diferentes ideas. Una de ellas es que la optimización sea más global y no tan golosa. Es decir, cuando se van a hacer las modificaciones al equipo en una fecha dada, no sólo se mire dicha fecha sino también una o dos más para adelante. E incluso pensar en una optimización más global para armar el equipo inicial. También se podría armar un equipo de "alto riesgo", con jugadores que presenten altas varianzas en sus puntajes aunque no tengan los mejores índices, tal cual se definieron en nuestro modelo. Esto podría llevar a peores resultados en general, pero podría otorgar cada una cierta cantidad de campeonatos una muy buena performance.

Otra alternativa sería intentar encontrar a los mejores jugadores de cada fecha, para formar al equipo virtual. Dado que los puntajes de los

jugadores suelen tener grandes variabilidades, puede ser interesante, más que predecir su puntaje, utilizar herramientas estadísticas para predecir quienes serán los mejores jugadores del fin de semana que se avecina. Para ello se podrían implementar modelos estadísticos sofisticados, utilizando datos históricos, para intentar encontrar cuáles son las variables que determinan en forma más fidedigna quienes serán los mejores jugadores en un determinado fin de semana (local, rival, estadio, árbitro, etc).

Cabe destacar también que el problema aquí encarado presenta algunas similitudes con el problema de armar un portafolio de acciones empresariales, a fin de maximizar las ganancias de un inversor. Por ello creemos que algunos de los modelos que se usan en finanzas para predecir el comportamiento futuro de determinadas acciones, como los modelos de tipo CAPM [4, 8], basados en la teoría del portafolio de Markowitz [5], también podrían ser útiles a fin de conseguir equipos robustos para nuestro director técnico virtual.

Herramientas como las desarrolladas en este trabajo deberían servir para asesorar a entrenadores en deportes reales. La combinación entre deportes y matemática con este fin es muy bien tratada en la película de Bennett Miller, "Moneyball", basada en [3] e interpretada por Brad Pitt y Jonah Hill. Esta película trata sobre la historia real de Billy Beane y Peter Brand, gerentes deportivos de un club de béisbol modesto de los Estados Unidos, quienes cambiaron radicalmente la forma de actuar, incorporando la matemática en la toma de decisiones deportivas, obteniendo de esta manera excelentes resultados.

Agradecimientos: A Carlos Prieto, que lamentablemente ya no está con nosotros, y acompañó este proyecto con gran entusiasmo. A Diego Javier Romero y Jorge Blanco, responsables del juego en el diario, por su ayuda permanente para la concreción de este trabajo. A Andrés Farall, Leonardo Faigenbom, Leonel Spett y Pablo Groisman, por las interesantes discusiones mantenidas sobre este proyecto. A Mario Guajardo por la lectura detallada del trabajo y sus sugerencias para mejorar la versión final. A Sebastián Ceria y Gustavo Braier, por sus comentarios y observaciones. El presente trabajo fue parcialmente financiado por los proyectos ANPCyT PICT-2012-1324 (Argentina), CONICET PIP 112-200901-00178 (Argentina), UBACyT 20020090300094 (Argentina), y por el Instituto Milenio "Sistemas Complejos de Ingeniería" (Chile). El segundo autor es

parcialmente financiado por el proyecto FONDECyT 1110797 (Chile).

Referencias

[1] Coleman B.J., Identifying the "Players" in Sports Analytics Research, Interfaces 42 (2) (2012) , 109-118.

[2] Kendall G., Knust S., Ribeiro C.C., Urrutia S., Scheduling in sports: An annotated bibliography, Computers & Operations Research 37 (1) (2010), 1-19.

[3] Lewis M., Moneyball: The Art of Winning an Unfair Game, Norton & Company (2003).

[4] Lintner J., The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, Review of Economics and Statistics 47 (1) (1965), 13-37.

[5] Markowitz H.M., Portfolio Selection, The Journal of Finance 7 (1) (1952), 77-91.

[6] Nurmi, K., Goossens, D., Bartsch, T., Bonomo, F., Briskorn, D., Durán, G., Kyngas, J., Marenco, J., Ribeiro, C.C., Spieksma, F., Urrutia, S., Wolf-Yadlin, R., 2010. A framework for scheduling professional sports leagues. In Ao, S-I., Katagir, H., Xu, L., Chan, A.H-S. (eds) IAENG Transactions on Engineering Technologies, American Institute of Physics. Vol. 5, pp. 1428.

[7] Ribeiro C.C., Sports scheduling: Problems and applications, International Transactions in Operational Research 19 (2012), 201-226.

[8] Sharpe W.F., Capital asset prices: A theory of market equilibrium under conditions of risk, The Journal of Finance 19 (3) (1964), 425-442.

[9] http://fantasy.premierleague.com/

[10] http://www.fantacalcio.kataweb.it/

[11] http://fantacalcio.repubblica.it/index.php?page=faq&ck_fantacalcio#16

[12] http://www.fantagazzetta.com/esclusive-fg/come-inventai-il-fantacalcio-intervista-esclusiva-a-riccardo-albini-inventore-del-fantacalcio-165805

[13] http://www.hattrick.org/

[14] http://mlb.mlb.com/mlb/fantasy/

[15] http://www.nba.com/fantasy/

[16] http://msn.foxsports.com/fantasy/football/

[17] http://www.xperteleven.com/

[18] http://edant.clarin.com/diario/2009/07/09/deportes/d-01955551.htm

[19] http://edant.clarin.com/diario/2009/12/20/deportes/d-02104838.htm

[20] http://edant.clarin.com/diario/2010/05/20/deportes/d-02197713.htm

[21] http:www.fantasysportsmath.com/

Programas de Postgrado y Postítulos DII

DOCTORADO

Doctorado
en Sistemas de Ingeniería

Formación de especialistas, con una sólida base tecnológica y un conocimiento profundo de las herramientas que permiten modelar, entender y optimizar sistemas complejos en que interactúan elementos físicos y de comportamiento humano.

El Doctorado en Sistemas de Ingeniería es un programa académico e interdisciplinario.

Los graduados, contarán con una visión rigurosa y transversal de la Ingeniería, con un énfasis en la investigación de alto nivel y estarán preparados para aportar a los crecientes desafíos de desarrollo productivo y social, tanto en instituciones de investigación y educación superior como en aquellas del ámbito empresarial y gubernamental.

Informaciones: www.dsi.uchile.cl
linda.valdes@dii.uchile.cl
562-29784017/ 562-9784073

El Plan de Estudios comprende un ciclo común, un ciclo de especialización en una de cuatro áreas (Economía, Gestión de Operaciones, Redes Eléctricas y Transporte) y cursos electivos. La elaboración de una Tesis y un Examen de Grado. Adicionalmente, los postulantes deberán aprobar un Examen de Calificación.

El claustro del programa de Doctorado en Sistemas de Ingeniería está conformado por académicos de muy alta calidad. Al claustro pertenecen académicos de los siguientes departamentos de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile: Ingeniería Civil, Ingeniería Eléctrica, Ingeniería Industrial e Ingeniería Matemática.

Requisitos de Admisión

Poseer el grado de licenciado en Ciencias de la Ingeniería o su equivalente.

Calendario de Postulaciones

Semestre Otoño
Periodo de postulaciones: de octubre a 15 de diciembre.
Inicio de clases: marzo de cada año.

Semestre Primavera
Período de postulaciones: de abril a 15 de junio.
Inicio de clases: julio de cada año.

fcfm
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

MAGÍSTERES

Magíster en Gestión de Operaciones MGO

El Magíster en Gestión de Operaciones es impartido por el Departamento de Ingeniería Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile.

El Magíster busca formar graduados de excelencia en investigación de operaciones, quienes podrán enfrentar problemas complejos en gestión de operaciones, integrando herramientas matemáticas, económicas y tecnológicas.

El programa prepara a sus graduados para desempeñarse en cargos de primer nivel en empresas de servicios y manufactura, donde la logística y las operaciones son fundamentales en su estrategia y la planificación juega un rol central.

El Magíster en Gestión de Operaciones se inició en el año 2000. En el año 2012, en su tercera acreditación, el MGO obtuvo 7 años.

Requisitos: Poseer el grado de Licenciado en Ciencias de la Ingeniería o su equivalente.
Postulación: Septiembre a Diciembre para ingresar en marzo de 2014.
Postulación en línea en: www.mgo.uchile.cl

INFORMACIONES: 562-29784017/ 562-29784073.
www.mgo.uchile.cl

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

Magíster en Economía Aplicada

El Magíster en Economía Aplicada tiene una orientación académica y busca formar profesionales y académicos de gran capacidad analítica y sólida base en economía.

El programa se ofrece en modalidad presencial y full-time.

El Magíster habilita a sus graduados para desempeñarse en empresas del sector financiero, de servicios, organismos internacionales, e instituciones reguladoras, entre otras, también los prepara para continuar estudios de doctorado y desarrollar una carrera académica.

El Magíster cuenta con un equipo académico de excelencia con diversidad de líneas de investigación. El programa es impartido por el Centro en Economía Aplicada (CEA).

Requisitos de Admisión: Poseer el grado de licenciado con una formación acorde con las exigencias del programa.

Postulación: Septiembre a Diciembre de 2013 para ingresar en marzo de 2014.
Abril a Junio para ingresar en Julio 2014.

Postulación en línea a través de www.magcea-uchile.cl

Informaciones: www.magcea-uchile.cl- obarrera@dii.uchile.cl- **562-29784072 / 562-29784073**

fcfm
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

Magíster en Gestión y Políticas Públicas MGPP

El Magíster en Gestión y Políticas Públicas, tiene como propósito la formación avanzada de profesionales interesados en la formulación y ejecución de políticas públicas.

El MGPP forma líderes y servidores públicos del más alto nivel, capaces de conceptualizar, pensar y discutir sus visiones e ideas sobre el futuro de América Latina.

Se imparte en dos horarios: Diurno y Ejecutivo.

Características Distintivas

* Excelencia Académica
* Cuerpo docente de primer nivel
* Orientado a profesionales de formación diversa
* Alta tasa de graduación
* Reconocido entre los mejores en su área en América Latina.
* Acreditado por la CNA por 7 años, desde octubre de 2011 a octubre de 2018.
* 18 años formando líderes

Requisitos de Admisión
* Poseer el grado de licenciado o título universitario

Horario Diurno:
Inicio: junio de cada año
Duración: 19 meses

Horario Ejecutivo:
Inicio: julio de cada año
Duración: 24 meses

Postulaciones:

Hasta el **15 de noviembre** para personas que postulan a becas de instituciones

Hasta el **15 de abril** para personas que cuentan con fondos propios

Mayor información: mgpp@dii.uchile.cl | www.mgpp.cl | Tel.: (562) 978 4067 | (562) 978 4043


INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE
MBE
Master in Business Engineering
Magíster en Ingeniería de Negocios con Tecnologías de la Información
Los líderes de hoy comprenden cómo la tecnología lleva a las empresas al éxito.
A Quién está Dirigido
Ejecutivos y profesionales que deseen liderar o ejecutar proyectos innovadores de diseño integral y sistémico de los negocios orientados a mejorar su competitividad.
Metodología
Este es un Magíster integrador, conformado por un conjunto de cursos de gestión, modelos analíticos aplicados, diseño de negocios, arquitectura y procesos, tecnologías de información de base y diseño de aplicaciones, y de inducción de habilidades de innovación.
Además de las evaluaciones tradicionales por medio de controles y exámenes, una parte fundamental del trabajo de los alumnos será el desarrollo de un proyecto de innovación en el negocio de la empresa auspiciadora -donde ejecutará su residencia-, el cual se llevará a cabo durante todo el programa, en los cursos obligatorios del mismo.
Duración:
3 semestres académicos más un semestre para dar término al Proyecto de Grado.
Horario:
Martes o jueves vespertino, viernes de 14:30 a 18:00 horas y sábados de 8:30 a 11:45 horas.
Informaciones:
Coordinadora: Ana María Valenzuela.
(56 2) 978 4835 / anamaria@dii.uchile.cl
www.dii.uchile.cl
fcfm
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE


INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE
MBA
Magíster en Gestión y Dirección de Empresas
VERSIÓN INDUSTRIA MINERA
Formato Week End
Cuarta versión: Santiago
Inicio: abril 2014
Información: (+562) 29784020 | mbamin@dii.uchile.cl | www.mbamin.cl
fcfm
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE