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

MEJORA DEL PROCESO DE PRODUCCIÓN DE ESTIMACIONES DE INSCRIPCIONES DE ALUMNOS PARA INSTITUCIONES DE EDUCACIÓN SUPERIOR

PROYECTO DE GRADO PARA OPTAR AL GRADO DE MAGÍSTER EN INGENIERÍA DE NEGOCIOS CON TECNOLOGÍAS DE INFORMACIÓN

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

IMANOL GABRIEL BIDEGAIN RIVERA

PROFESOR GUÍA:
SEBASTIÁN RÍOS PEREZ

MIEMBROS DE LA COMISIÓN:
NICOLAS OZIMIÇA GACITÚA
CONSTANZA CONTRERAS PIÑA
SEBASTIÁN FLORES BENNER

SANTIAGO DE CHILE
2017

RESUMEN EJECUTIVO

En el presente informe se detalla el desarrollo de la metodología que permite optar al grado de magíster en ingeniería de negocios con tecnologías de información y al título de ingeniero civil industrial.

El proyecto se realiza en la empresa U-planner, empresa dedicada al apoyo de la administración y gestión de recursos educacionales. Mediante la combinación de modelos matemáticos y *softwares* busca satisfacer las necesidades de cada cliente para poder convertirse en un apoyo estratégico de éste. La principal actividad económica es la venta e implementación de estos *softwares*, donde actualmente en la empresa se ofrecen nueve distintos productos.

El proyecto busca realizar una actualización en el producto de U-Forecast (Estimación de demanda), mediante el rediseño de los procesos que involucran a la producción de la información este producto. La oportunidad de mejora nace debido a la expansión dentro de Latinoamérica que ha tenido la empresa, en donde se han visto casos en los que el producto no se adapta a las reglas de negocio existentes en los establecimientos educacionales.

La solución presentada incorpora la creación de una nueva lógica que utiliza modelos de inteligencia artificial basados en árboles de decisión CHAID. Los modelos buscan encontrar patrones en las inscripciones que realizan los estudiantes en años anteriores y con esta información entregar una esperanza de los alumnos que inscribirán cada uno de los ramos. Además se presenta una metodología de uso del algoritmo para mejorar las predicciones, procesos nuevos de limpieza de datos e indicadores nuevos para medir el rendimiento de los algoritmos.

Se realizó una prueba piloto para el prototipo la que consistió en una consultoría a una universidad peruana. Con lo que se verificó el rendimiento del algoritmo bajo los indicadores propuestos en este trabajo. Utilizando la metodología propuesta en esta tesis, se sube en un 22.7% el indicador de cantidad de asignaturas con secciones correctamente estimadas.

Desde el punto de vista económico, se realizó una evaluación como un proyecto privado. El mayor impacto del proyecto, es generar la posibilidad de la captación de clientes a los cuales antes no se les podía ofrecer este producto debido a la baja adaptabilidad de éste. Finalmente se estima que el proyecto es rentable para los escenarios moderado y optimista, es decir, mediante la captación de tres o más clientes nuevos, entregando un VAN de 9.800.000 aproximadamente.

Dedicado a mis padres Marcos y Ana María y a mis abuelos,
quienes me han dado siempre un apoyo incondicional,
y han sido ejemplos a seguir para mi vida.

Agradecimientos

Primero que todo, agradecer a mi familia por darme las oportunidades y el apoyo incondicional para llegar hasta este punto de mi vida. Por la formación que me dieron y la confianza que siempre han tenido en mí. Espero poder ser siempre un orgullo para ustedes.

Agradecimientos a los amigos de la universidad especialmente a los de los viajes extremos, PBP y la bandita de Industrias por ayudarme a pasar esta etapa con alegría y más de alguna fiesta. A mis amigos de TPMP2 por haber estado desde siempre y ser un gran apoyo. A todos ellos por haberme brindado algunos de los mejores recuerdos de mi vida.

A mis profes guías, Sebastián y Nicolás, por la confianza puesta en mí y la ayuda brindada que me sirvió para aclarar las ideas cuando lo necesitaba. Asimismo, agradecer a Ana María Valenzuela y Laura Sáez por su apoyo administrativo a lo largo de todo este proceso dentro del MBE.

Agradecer a toda la empresa U-planner por ayudar a hacer este proyecto posible, en especial a León Montero, Felipe Luengo y Sebastián Flores, quienes fueron piezas fundamentales para concretar este trabajo, entregando conocimientos y apoyo durante toda la duración de la tesis.

A Fernanda Martínez por haber sido un gran apoyo emocional durante toda la carrera, por darme ánimo en todo momento y por creer siempre que sería capaz de realizar mis metas.

Tabla de Contenido

Índice de Figuras

Índice de Tablas

CAPÍTULO 1: INTRODUCCIÓN Y CONTEXTO

En este capítulo se detalla el contexto del proyecto tanto de la industria como dentro de la empresa y luego se explica la oportunidad de mejora identificada. Posteriormente se abordan los objetivos y se analizan los resultados esperados, riesgos y alcance del proyecto.

1.1 ANTECEDENTES DE LA INDUSTRIA DE LA ADMINISTRACIÓN Y GESTIÓN DE RECURSOS DE INSTITUCIONES DE EDUCACIÓN SUPERIOR

La administración y gestión de los recursos en las instituciones de educación superior es un tema que se ha tratado de manera particular por los mismos establecimientos educacionales y en donde han tenido que apoyarse principalmente del autoaprendizaje. En el mercado estadounidense y europeo la industria que brinda servicios de ayuda en la administración y gestión de estos recursos tiene ya varios años de experiencia y está más desarrollada. Por el contrario el mercado latinoamericano es una industria relativamente nueva en donde existen pocas empresas o instituciones que se dediquen a apoyar a las universidades en la resolución sus problemas. A esto se suma que no poseen la experiencia ni el nivel de especialización necesaria para realizar mejoras en los procesos internos de los establecimientos educacionales. Por ello la oferta de soluciones es insuficiente, muy diversa entre competidores, resuelve problemas particulares del cliente y quedan muchas necesidades insatisfechas.

Durante los últimos años en Chile y Latinoamérica se ha formado una tendencia que intenta incorporar el concepto de la gestión de las organizaciones educacionales en términos más estandarizados y de estrategia, que son similares a los aplicados en las corporaciones con el fin de aumentar la eficiencia en el manejo de sus recursos. Esto ha ayudado al surgimiento de empresas que apoyan a las universidades en sus procesos de gestión y administración.

El mercado de la industria de servicios para la administración y gestión de recursos de instituciones de educación superior está compuesto por 17.000 instituciones a lo largo del mundo las que otorgan educación a 173 millones de estudiantes aproximadamente. Si se considera el mercado potencial, en particular el latinoamericano se habla de 3.750 instituciones de educación superior que brindan educación a 40 millones de alumnos aproximadamente [21].

La industria que brinda servicios de apoyo se sustenta bajo la venta de productos que generan soluciones a las principales problemáticas de las universidades como por ejemplo: planificación académica, retención de alumnos, acreditación institucional y de carreras, aprendizaje de alumnos, operaciones durante el semestre, comunidad en la universidad e

investigación. La principal problemática que se debe considerar es que para la administración de las instituciones no existe un único estilo de gestión, por lo que a pesar de que cada producto tiene su propio objetivo la forma de adaptarlos es particular a cada institución.

La industria que brinda servicios de apoyo se caracteriza por tener un estilo de competencia muy similar a la industria desarrolladora de *softwares*. En dicha industria los principales atributos para medir la calidad de un *software* son su capacidad para encontrar una solución al problema, su usabilidad y el tiempo que toma en su ejecución.

1.2 DESCRIPCIÓN GENERAL DE LA EMPRESA

Las organizaciones educacionales, como cualquiera otra organización pública o privada necesitan gestionar sus asuntos y recursos, planificar y evaluar sus actividades, desarrollar una visión de su propia identidad y misión y generar liderazgos internos capaces de generar participación y eficiencia. Según M. Alayon y R. Cuicas (2014) en su artículo declara que las organizaciones que se auto gestionan y comparten su conocimiento logran su máximo rendimiento, esto se ve reflejado en alcanzar un estándar de calidad que es definida como "el conjunto de características inherentes a la prestación de un servicio para satisfacer las necesidades y expectativas del cliente y otras partes interesadas" y así mejorar la competitividad de la organización [9]. Este concepto utilizado por organizaciones a nivel mundial para realizar una gestión de calidad es utilizable en los establecimientos educacionales.

Es así como en el 2009 comienza U-planner, una organización que busca mejorar la gestión en la educación poniéndose como objetivo convertirse en un aliado estratégico de las universidades ayudando y brindando información valiosa para mejorar la toma de decisiones, apoyándolos integralmente, prestando soporte técnico constante y asesoría especializadas durante toda la relación.

U-planner se caracteriza por introducir el concepto de "Ingeniería para la Educación", con lo que busca ayudar a las universidades a optimizar sus procesos internos, la utilización de sus recursos y mejorar la experiencia educativa que brindan a sus estudiantes. La empresa apoya a las universidades en los principales problemas que tienen mediante sus productos y asesorías que entregan información específica de un problema de manera sencilla a los clientes para que puedan tomar mejores decisiones.

Actualmente la empresa cuenta con un equipo de desarrollo y de implementación de primer nivel, especialistas en gestión de recursos académicos y en tecnologías de la información al

servicio de la educación. El equipo está compuesto principalmente por matemáticos, diseñadores, ingenieros, programadores y desarrolladores.

La empresa tiene un compromiso con la educación, es por esto que en sus bases declara una misión y una visión, en donde la primera representa la motivación que tiene la empresa en el presente, mientras que la visión es una imagen de la organización que se plantea a largo plazo.

La misión de la empresa es “La manera de avanzar con nuestro compromiso es incorporar ingeniería a la educación. De esta manera queremos ayudar a nuestros clientes a optimizar de forma eficiente sus procesos internos y recursos, aumentar su calidad de operaciones e incrementar sus ahorros con nuestras soluciones para que puedan mejorar su competitividad y prestigio institucional.”.

La visión de la empresa es “Tenemos la visión de un mundo, donde la educación es de alta calidad y al alcance de todos, independiente a factores sociales, culturales o económicos. En consecuencia nuestro compromiso es de disponibilizar a las instituciones educativas soluciones que generen condiciones que permitan a los estudiantes, profesores y staff de experimentar una experiencia educativa de alta calidad.”

1.3 PROBLEMA U OPORTUNIDAD IDENTIFICADA

Dentro de la cartera de productos que ofrece U-planner, uno de los más solicitados y el más antiguo es U-planning. Este producto es el encargado de la planificación académica para las universidades. A su vez, brinda simulaciones de asignaciones horarias en base a una demanda entregada o estimada, asignando infraestructura y profesores a cada sección, permite crear múltiples soluciones en poco tiempo simulando diferentes escenarios y entregando reportes visuales e inteligencia en los análisis.

El producto mencionado se divide en tres sub productos que funcionan como tres módulos independientes, pero que se complementan y se alimentan con información. El primero de éstos es U-Forecast el cual realiza estimaciones de la demanda de alumnos para cada una de las asignaturas y la generación de secciones, en donde existe un algoritmo para cada uno de estos cálculos. El segundo es U-Schedule que se encarga de la asignación de módulos horarios y salas de clases. Finalmente U-Teacher que asigna a profesores a una sección con su respectivo horario y sala de clases.

Los productos necesitan constantemente mejoras y correcciones en la implementación y en las lógicas de los algoritmos para estar a la vanguardia en la tecnología y en las necesidades

de los clientes. En los productos de U-Schedule y U-Teacher se han realizado grandes mejoras, mientras que el producto U-Forecast se ha quedado atrás requiriendo una actualización basada en los nuevos conocimientos que se tiene de los datos de los clientes.

El problema identificado es que el producto U-Forecast no logra adaptarse, ni realizar buenas predicciones a todos los tipos de clientes que actualmente tiene la empresa. Esto salió a la luz luego que U-planner se expandiera a nuevos mercados y se involucrara con clientes distintos a los chilenos, lo que introdujo nuevas problemáticas para el algoritmo debido a diferencias en las mallas curriculares de los planes y reglas en la inscripción de los ramos.

La baja precisión y poca adaptabilidad del algoritmo en algunos clientes se pueden asociar a la flexibilidad y complejidad de las mallas curriculares, como por ejemplo mallas con la posibilidad de tomar muchos electivos, cursos que no siguen un orden de sucesión en particular o en donde los requisitos pueden ser diversos como varios cursos y/o número de créditos cursados, situaciones que no se habían considerado.

El algoritmo también cuenta con algunas limitaciones para poder realizar la predicción tales como requerir una gran cantidad de información de parte del cliente, lo que puede en algunos casos ser complicado de obtener, solo permite realizar predicciones de asignaturas obligatorias dentro de la malla y se deben entregar algunos parámetros como la proporción de estudiantes que tomará cada ramo. Éste es un parámetro que debe calcular uno de los analistas. Debido a esto el equipo de U-planner debe dedicar tiempo y recursos para adaptar el algoritmo (en caso de ser posible), definir los parámetros, cambiar lógicas y realizar transformaciones en los datos para poder desarrollar el producto específico que necesita el cliente.

Cabe destacar que el proceso de producción de la información no ha sido documentado ni estandarizado para ningún producto y las acciones se realizan tácitamente de acuerdo a la experiencia e intuición de los empleados encargados.

1.4 OBJETIVOS Y RESULTADOS ESPERADOS DEL PROYECTO

1.4.1 OBJETIVO GENERAL

El objetivo del presente proyecto es darle a U-planner una alternativa de oferta que se complemente con el actual producto de U-Forecast, para satisfacer clientes que no se pueden satisfacer o en los cuales se podría tener un bajo desempeño al utilizar el actual producto.

1.4.2 OBJETIVOS ESPECÍFICOS

Para lograr el objetivo general se plantearon los siguientes objetivos específicos:

- Levantar la arquitectura y estructura de procesos de la empresa.
- Declarar el proceso de producción de información de la demanda estimada.
- Diseñar y crear una herramienta que ejecute un algoritmo de estimación de demanda que se complemente con el producto que existe actualmente.
- Proponer indicadores de rendimiento del resultado de los algoritmos para el producto U-Forecast.

1.4.3 RESULTADOS ESPERADOS

El proyecto busca obtener los siguientes beneficios:

- Desarrollo y validación de la lógica que utiliza el algoritmo.
- Tener una alternativa efectiva de oferta para clientes en donde el algoritmo actual no pueda realizar una predicción o no entregue una estimación satisfactoria.
- Disminuir los tiempos de integración del producto con el cliente, dada la menor cantidad de datos que el algoritmo necesita.
- Disminuir los errores de estimación, lo que se refleja en una mayor satisfacción para el cliente.
- Indicadores de rendimiento que ayudarán a escoger la mejor herramienta de predicción para cada caso.

1.5 ALCANCES

El alcance del proyecto incluye los aspectos relacionados al diseño del proceso de producción y a la creación e implementación de la nueva herramienta tecnológica para este proceso:

- Diseño de la Arquitectura y Estructura de procesos de la empresa U-planner.
- Rediseño del proceso de producción de información para el producto U-Forecast.
- Creación de los algoritmos de calibración y ejecución de modelos de Estimación de Demanda.
- Creación del proceso de mantención de los modelos.
- Prueba piloto de la solución tecnológica.
- Creación del proceso de limpieza y carga de los datos que alimentan el algoritmo.
- Plan de gestión del cambio para el uso de la herramienta tecnológica.
- Documentación para la utilización del módulo.

1.6 RIESGOS POTENCIALES

Estos son algunos de los riesgos potenciales identificados:

- **Que los modelos utilizados no se ajusten lo suficiente a los cursos lo que entregaría resultado subóptimos.** Esto implicaría tener que buscar una nueva solución de modelos para realizar la actualización del producto.

- **Resistencia al uso de la herramienta.** En general ante un cambio en un proceso existe una resistencia a dicho cambio por lo que puede existir una aversión a aprender a utilizar la herramienta propuesta. Por esto ante cualquier duda que surja por parte de los clientes o del analista, la solución entregada podría convertirse en una caja negra. La forma de solucionar esto es realizar capacitaciones a las personas que posteriormente estarán involucradas y generar documentación que le permita nuevas personas aprender su funcionamiento fácilmente.

- **Un mal rendimiento en los resultados del algoritmo para algunas universidades.** Puede que las universidades tengan demandas en sus cursos demasiado variables y sin ningún patrón identificable. En este caso la solución presentada tampoco se les podrá ofrecer.

- **Que no se cuente con los datos mínimos para la ejecución del algoritmo.** Las universidades podrían no disponer de los datos para realizar la predicción. Para solucionar esto se debe sostener conversaciones previas y llegar a un acuerdo de cuáles son los datos que se pueden recuperar o construir y determinar si son suficientes para la ejecución del algoritmo.

- **Que el área comercial no logre atraer a la cantidad necesaria de nuevos clientes que requiere el proyecto para ser exitoso económicamente.** Esto se puede mitigar realizando un plan estratégico en conjunto con el área de ventas, el cual tenga considerado un plan de marketing que ayude en las ventas de este producto.

CAPÍTULO 2: MARCO TEÓRICO

Este capítulo explica la metodología utilizada para la realización del proyecto, metodología utilizada en el magister de Ingeniería de Negocios con Tecnologías de Información del Departamento de Ingeniería Civil Industrial de la Universidad de Chile (MBE).

También se describe de manera general conceptos utilizados en la notación para el modelamiento de procesos y la extracción de datos.

2.1 METODOLOGÍA DE INGENIERÍA DE NEGOCIOS

La metodología que se utilizó para llevar a cabo este proyecto es la impartida en el MBE, y desarrollada en el libro "Ingeniería de Negocios, Diseño integrado de Negocios, Procesos, y Aplicaciones TI" creado por Oscar Barros (2009) [2].

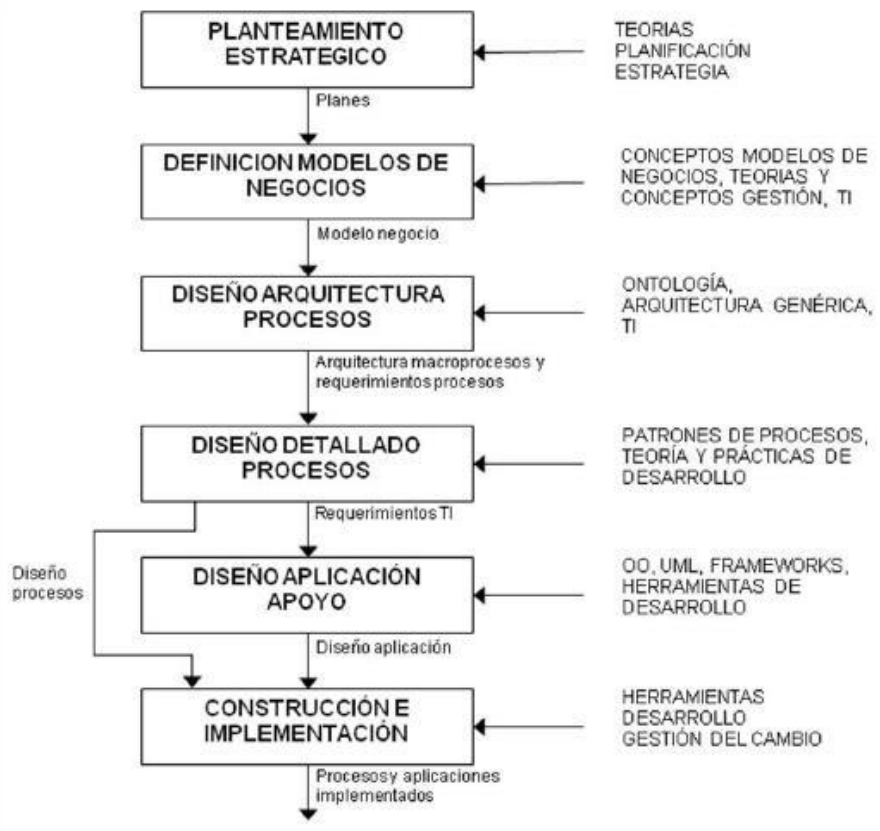

Figura 1: Metodología de Ingeniería de Negocios

Según lo definido por Barros, la Ingeniería de Negocios es una disciplina que busca guiar a las organizaciones en el diseño, construcción e implementación de sus procesos. En la Figura 1 es posible ver el esquema de esta metodología, que incluye los siguientes elementos:

- Planteamiento Estratégico: Este es el punto de inicio, el cual requiere un claro planteamiento respecto al posicionamiento estratégico que busca la organización. A este nivel se aplica la teoría de estrategia de Hax [11]. Asimismo, se busca generar un mapa estratégico donde se relacione la visión y misión de la empresa con las distintas perspectivas del BSC (*Balanced ScoreCard*), observando a la empresa internamente y su interacción con la industria.

- Definición del Modelo de Negocio: Se establece cómo materializar el posicionamiento estratégico en una oferta hacia los clientes que les genere valor y por el cual estén dispuestos a pagar. Esto se diseña en base a la metodología Canvas [14].

- Diseño de la Arquitectura de Procesos: Consiste en diseño de grandes agrupaciones de procesos, que llamaremos macroprocesos, creados a partir del modelo de negocio. Este diseño ayuda a ejecutar de la mejor manera posible tal modelo, el cual utiliza como punto de partida los patrones de arquitectura de procesos propuestos por Barros y Julio [12]. Para la realización de este diseño se utiliza la metodología IDEF0.

- Diseño Detallado de Procesos del Negocio: Se detallan los macroprocesos de la arquitectura, utilizando como base la notación IDEF0.

- Diseño de las Aplicaciones TI: Consiste en detallar los apoyos de tecnologías de información a los procesos definidos en el punto anterior, para lo cual se utiliza la metodología de especificación de requerimientos de *software* UML (*Unified Modeling Language*).

- Construcción e Implementación: Con las herramientas que crean un ambiente de *software* para el tipo de diseño definido anteriormente, se construyen las aplicaciones necesarias y se implementan, llevando a la práctica los diseños de procesos que usan las aplicaciones.

2.2 NOTACIÓN DE MODELAMIENTO DE PROCESOS DE NEGOCIO

En esta sección se explica con mayor detalle las notaciones utilizadas en el modelamiento de procesos, a nivel general (IDEF0).

2.2.1 IDEF0

IDEF0 es un método diseñado para modelar decisiones, acciones, y actividades de una organización o sistema. Esta notación ayuda a organizar el análisis de un sistema para promover una buena comunicación entre el analista y el consumidor. Como una herramienta de comunicación, IDEF0 mejora la participación de expertos en el dominio y apoya la toma de decisiones a través de dispositivos simplificados de gráficas [13].

Esta metodología utiliza gráficas basadas en cajas y flechas, contenidas en diagramas que muestran la función de una actividad como una caja, y las interfaces hacia o desde la actividad como flechas entrando o saliendo de la caja. Para expresar funciones, las cajas operan simultáneamente con otras cajas, con las interfaces de flechas "restringiendo" cuando y como las operaciones son ejecutadas y controladas. La sintaxis básica de un modelo IDEF0 se aprecia en la Figura 2.

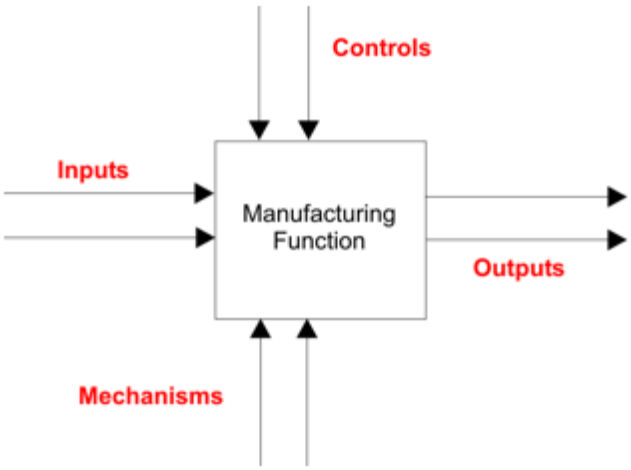

Figura 2: Gráfica de caja y flechas IDEF0

2.2.2 BPMN

Business Process Modeling Notation o BPMN (Notación para el Modelamiento de Procesos de Negocio en español), es una notación estandarizada que provee a las organizaciones la capacidad de entender sus procesos de negocios internos de manera gráfica. Es más brinda a las organizaciones la capacidad de entender procesos colaborativos y transacciones de negocio entre organizaciones, en un formato de flujo de trabajo. BPMN tiene la finalidad de servir como lenguaje común para cerrar la brecha de comunicación que frecuentemente se presenta entre el diseño de los procesos de negocio y su implementación.

El modelado en BPMN se realiza mediante diagramas con un conjunto de elementos gráficos. Existen cuatro categorías básicas de elementos descritos a continuación:

- **Objetos de Flujo**: Son los elementos principales descritos dentro de BPMN y consta de tres elementos principales: Eventos, Actividades y Compuertas (Control de Flujo). Dentro de los eventos podemos destacar los de inicio y fin. Las actividades pueden ser tareas ejecutadas por un actor, o bien un sub-proceso. Las compuertas de control establecen un control del flujo mediante *gateways*.

- **Objetos de Conexión**: Los objetos de conexión permitirán conectar cada uno de los objetos de flujo. Hay tres tipos: secuencias (representados por una línea simple, continua y flechada), mensajes (línea discontinua con un círculo no relleno al inicio y una punta de flecha no rellena al final) y asociaciones (representadas a través de líneas punteadas).

- ***Swimlanes* o Carriles de Piscina**: *Pool* (en español piscina) representa los participantes principales de un proceso, por lo general, separados por las diferentes organizaciones. Una piscina contiene uno o más carriles (en la vida real, como una piscina olímpica). Los carriles son usados para organizar y categorizar las actividades dentro de un pool.

- **Artefactos:** Corresponden a especificaciones del diagrama, que permite a los desarrolladores mostrar más información para hacerlo más legible. Dentro de estas especificaciones podemos encontrar objetos de datos, grupos y anotaciones.

2.3 LÓGICA DE NEGOCIOS

Dada la alta evolución que han tenido las tecnologías y las amplias necesidades que tienen los clientes, se hace necesario contar con nuevas herramientas, que permitan satisfacer en todo ámbito las necesidades de los establecimientos educacionales.

En departamento de ingeniería civil industrial cuenta con un grupo destacado de profesionales que realizan investigación utilizando la minería de datos en diversas áreas, por ejemplo, un estudio para mejorar el contenido de un sitio web utilizando modelos de clasificación [8], un estudio para determinar patrones en comunidades de temas superpuestas en redes sociales [23], estudio que determinan las bases para desarrollar nuevos algoritmos que detecten de forma automática la apnea obstructiva del sueño en niños [26]. Con lo que se ve el reflejo de los múltiples usos de la minería de datos para afrontar distintas problemáticas en variados ámbitos.

La minería de datos, permite a empresas como U-planner generar gestión para la educación de manera inteligente, realizando un trato apropiado de los datos, para transformarlos en información valiosa para las universidades. La aplicación de técnicas de minería de datos, por ejemplo árboles de decisión, permite mirar el problema de la estimación de demanda desde otra perspectiva, en donde se cambia la mirada actual de un modelo determinista a un modelo que utiliza probabilidades del cual se obtiene una esperanza de alumnos.

El cambio de perspectiva permite ver el comportamiento probable de un estudiante comparándolo con las decisiones tomadas por alumnos similares en el pasado, con lo que luego se genera una estimación agregada.

El marco teórico que se presenta a continuación, incluye los conceptos necesarios para la comprensión y aplicación del proceso de ejecución y calibración de los modelos de estimación de demanda, propuesto en este proyecto. Particularmente, se detalla el proceso de construcción de un árbol de decisión, que es el que se desea implementar en la empresa.

2.3.1 Aprendizaje Supervisado

El aprendizaje supervisado busca aprender una función f a partir de datos de entrenamientos de ejemplo. Se mira un set de entrenamiento (normalmente vectores) en donde una componente del par son los datos de entrada y el otro, los resultados esperados llámese *inputs* y *outputs* respectivamente $T = (x_i, y_i),\ i, \dots, N$. Los valores de entrada observados x_i alimentan un sistema artificial, conocido como algoritmo de aprendizaje (generalmente un programa computacional), el que también produce datos de salida $\hat{f}(x_i)$ en respuesta a los datos de entrada. El algoritmo de aprendizaje tiene la propiedad que puede ir cambiando la relación entre los datos de entrada/salida $\hat{f}$ respondiendo a la diferencia $y_i - \hat{f}(x_i)$ entre el original y el dato de salida generado. Este proceso es conocido como aprendizaje por ejemplos. Para terminar el proceso de aprendizaje lo esperable es que los datos de salida artificial y real sean lo suficientemente cercanos para ser utilizado en todos los sets de datos de entrada que puedan presentarse en la práctica.

2.3.2 Árboles de Decisión

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, cuya finalidad es describir a partir de variables explicativas x_i el comportamiento de una variable objetivo Y . La gran ventaja que conlleva la utilización de estos modelos para la

predicción, es su fácil interpretación, en donde se genera una situación que puede ser observable y se puede conocer la razón de las decisiones.

Los métodos basados en árboles dividen el espacio en rectángulos, luego ajustan las relaciones a un modelo simple (por ejemplo a una constante) en cada uno. Consideremos un problema de regresión con una variable de respuesta continua Y y los datos de entrada X_1 y X_2, cada uno tomando valores en el intervalo de la unidad.

En el panel de arriba en la izquierda de la Figura 3 se muestra la partición del espacio por líneas paralelas a los ejes. En cada partición se puede modelar Y con una constante diferente. Sin embargo, existe un problema: a pesar que cada partición tiene una descripción simple como $X_1 = c$, algunas de las regiones resultantes son difíciles de explicar. Para simplificar el ejemplo se restringe la atención a particiones recursivas binarias como la que se muestra en el panel de arriba en la derecha de la Figura 3. Primero se parte el espacio en dos regiones, el modelo entrega la media de Y en cada región. Se elige la variable y el punto de partición para alcanzar el mejor ajuste. Luego una o ambas regiones son particionadas en dos regiones más, este proceso continúa hasta que se cumpla alguna regla de interrupción. Por ejemplo en el panel de arriba en la derecha de la Figura 3, primero se divide en $X_1 = t_1$. Posteriormente la región $X_1 \leq t_1$ es dividida en $X_2 = t_2$ y la región $X_1 > t_3$ es dividida en $X_1 = t_4$. El resultado final de este proceso es una partición de cinco regiones $R_1, R_2, \ldots, R_5$ mostradas en la Figura 3. El modelo de regresión correspondiente predice Y con una constante c_m en la región R_m con la función:

$$\hat{f}(X) = \sum_{m=1}^{5} c_m \, I\{(X_1, X_2) \in R_m\} \qquad (1)$$

Este mismo modelo puede representarse como un árbol binario como muestra el panel de abajo en la izquierda de la Figura 3. El conjunto de datos completo se encuentra en la parte superior del árbol. Las observaciones que satisfagan la condición en cada unión son asignadas a la rama izquierda y las demás a la rama derecha. Los nodos terminales o hojas del árbol corresponden a las regiones $R_1, R_2, \ldots, R_5$. El panel de abajo en la derecha de la Figura 3 muestra un gráfico de la superficie del modelo de regresión. Para la ilustración se escogió para los nodos $c_1 = -5$, $c_2 = -7$, $c_3 = 0$, $c_4 = -2$, $c_5 = -4$ para realizar este gráfico. La venta clave de los árboles binarios recursivos es su interpretabilidad. La partición del espacio es completamente descrita por un árbol.

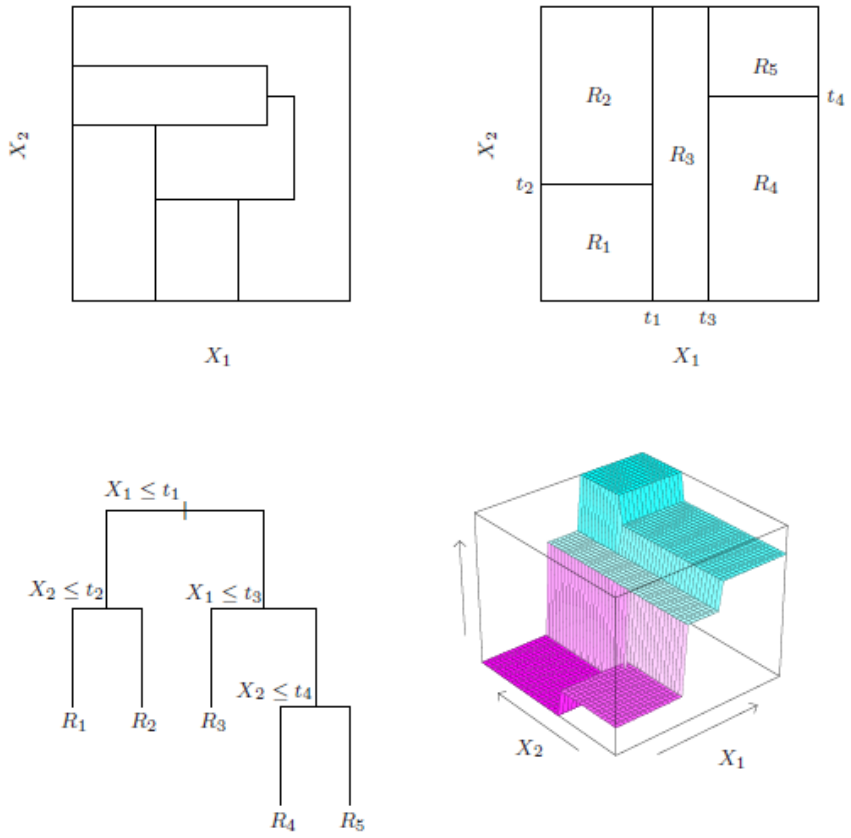

Figura 3: Particiones del espacio

2.3.3 Construcción del árbol de decisión de clasificación

La metodología del árbol de decisión utiliza un algoritmo llamado particionamiento recursivo para la creación del árbol, en donde se fracciona el conjunto inicial de datos de entrada basados en alguna medida de impureza.

Supongamos que nuestra data consiste en p *inputs* y *outputs* para cada N observaciones: esto es (x_i, y_i) para $i = 1,2, \ldots, N$. El algoritmo decidirá en que variables se va a dividir, en qué nivel y también que forma tomará el árbol.

Cuando se encuentra la mejor partición, se dividen los datos en dos regiones y se realiza el proceso nuevamente para cada una de las nuevas regiones. Luego se repite el proceso de división en todas las regiones resultantes.

El algoritmo decidirá qué tan grande será el árbol, debido a que un árbol muy grande puede causar un sobre ajuste, mientras que uno pequeño puede tener una estructura con muchos

casos en cada nodo. El tamaño del árbol es un parámetro que mide la complejidad del modelo y el tamaño óptimo debiese ser adaptable a los datos escogidos.

La estrategia preferible es hacer crecer un árbol T_0 hasta que se genere un nodo con un tamaño mínimo (por ejemplo, 5 datos). Luego este árbol es podado mediante el proceso que se describirá a continuación.

Se define un subárbol $T \subset T_0$ como cualquier árbol que pueda ser obtenido a través de podar a T_0, lo que significa uniendo cualquiera de sus nodos internos, es decir, no terminales. Se define los nodos terminales con m, el nodo m representa la región R_m. Sea $|T|$ el número de nodos terminales de T. Sea

$$N_m = \#\{x_i \in R_m\}$$

Se define el costo de complejidad:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T| \qquad (2)$$

En donde $Q_m(T)$ es la medida de impureza del nodo m para el árbol T.

La idea es encontrar para cada α el subárbol $T_\alpha \subseteq T_0$ que minimiza $C_\alpha(T)$. El parámetro $\alpha \geq 0$ controla el tradeoff entre el tamaño del árbol y la bondad de ajuste de los datos. Para α grandes resultan árboles pequeños, mientras que para α pequeños resultan árboles grandes. Por lo que con $\alpha = 0$ el modelo creado es un árbol completo T_0.

Para cada α se puede demostrar que existe un único subárbol más pequeño T_α que minimiza $C_\alpha(T)$. Para encontrar T_α se usa la poda más débil que consiste en que sucesivamente se unen los nodos internos que producen el incremento más pequeño en $\sum_m N_m Q_m(T)$, y se continua hasta que se produzca un solo nodo (la raíz). Esto entrega una secuencia de subárboles y se puede demostrar que esta secuencia debe contener a T_α. La estimación de α se consigue realizando cinco a diez validaciones cruzadas: se escoge el valor $\hat{\alpha}$ que minimice la validación cruzada que se define posteriormente. El árbol final es el T_α

A continuación se detalla la construcción del índice de impureza del nodo $Q_m(T)$. En un nodo m, representa una región R_m con N_m observaciones se define

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I\,(y_i = k) \qquad (3)$$

Como la proporción de observaciones de la clase k en el nodo m. Se clasifican las observaciones en el nodo m a la clase $k(m) = \arg\max_k(\hat{p}_{mk})$, la clase de la mayoría en el nodo m. Existen diferentes medidas $Q_m(T)$ de la impureza del nodo como por ejemplo:

- Error de clasificación: $\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$.
- Índice de Gini: $\sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$.
- Entropía cruzada o la desviación : $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$.

Para el caso de dos clases, si p es la proporción en la segunda clase, estas tres medidas son $1 - \max(p, 1-p)$, $2p(1-p)$ y $-p\log p - (1-p)\log(1-p)$ respectivamente. Se puede notar que los índices de Gini y de entropía cruzada son más sensibles ante cambios en las probabilidades de los nodos que el error de clasificación. Por ejemplo en un problema de dos clases con 400 observaciones en cada clase (asignamos esto como (400,400)), suponemos una partición crea nodos (300,100) y (100,300), mientras que una segunda partición crea los nodos (200,400) y (200,0). Ambas particiones producen un error de clasificación de 0.25, pero la segunda partición produce un nodo puro y esto es probablemente más preferible. Ambos índices de Gini y de entropía cruzada son menores para la segunda partición. Por esta razón, tanto el índice de Gini y de entropía cruzada deberían usarse al construir un árbol. Para guiar la poda por el costo de complejidad, cualquier medida puede ser utilizada, pero generalmente se utiliza el error de clasificación.

Se puede interpretar el índice de Gini de dos maneras interesantes. En vez de clasificar las observaciones a la clase de la mayoría dentro del nodo, se puede clasificar a la clase k con probabilidad $\hat{p}_{mk}$. Luego el error de entrenamiento en este nodo será $\sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'}$ – el indice de Gini. Del mismo modo, si configuramos cada observación como 1 para la clase k y cero en otro caso, la varianza del nodo es $\hat{p}_{mk}(1 - \hat{p}_{mk})$. Si se suma sobre todas las clases se obtiene nuevamente el índice de Gini.

2.3.4 Validación Cruzada

La eficiencia de predicción es una medida diferente, que establece cuán bien el modelo predice la variable dependiente para observaciones futuras, por lo que es utilizada frecuentemente para la selección de modelos.

La validación cruzada consiste en dividir el set de datos del que se dispone, de forma aleatoria, en dos subconjuntos disjuntos, pero equivalentes: el set de entrenamiento, que se utilizará para calibrar el modelo y el set de prueba, con el que se determinará cuán eficiente es el modelo al momento de predecir nuevos datos.

2.3.5 Sobreajuste

Es posible que un modelo se ajuste bien a los datos de entrenamiento pero al momento de incorporar nuevos datos realice un trabajo deficiente al momento de predecir la variable dependiente. Que un modelo esté sobre ajustado, quiere decir que predice significativamente mejor los datos de entrenamiento que los datos de prueba, por lo que su capacidad de predecir correctamente nuevos casos será baja.

Una forma de verificar que el modelo no se encuentra sobre ajustado, es revisando que el error (o acierto) de predicción es "similar" en el set de entrenamiento y de prueba. Para ello, es necesario construir las matrices de confusión en ambos sets como se muestra en la Figura 4. La matriz de confusión, es una matriz de dos por dos, que contiene en sus filas los valores predichos de la variable dependiente y en sus columnas el valor observado.

		Observado	
		$Y_i = 1$	$Y_i = 0$
Predicho	$\hat{Y}_i = 1$	*A*	*C*
	$\hat{Y}_i = 0$	*B*	*D*

Figura 4: Matriz de confusión.

En la matriz se tiene que:

- A es el número de verdaderos positivos predichos por el modelo
- B el de falsos negativos
- C el de falsos positivos
- D el número de verdaderos negativos.

De aquí se deduce, por ejemplo, que la expresión $\frac{A+D}{A+B+C+D}$ es el porcentaje de acierto global del modelo, y que es el $\frac{A}{A+B}$ porcentaje de acierto de malos, del total de malos observados.

Luego, si $\frac{A}{A+B_{entrenamiento}} \approx \frac{A}{A+B_{test}}$, el modelo no se estará sobre ajustando a los datos, en la predicción de la clase 1. Lo anterior también debe cumplirse para las demás componentes.

La Figura 5 muestra el típico comportamiento de los errores de entrenamiento y testeo, a medida que varía la complejidad del modelo. El error de entrenamiento tiende a decrecer a medida que incrementa la complejidad del modelo, esto es, cuando ajustamos fuertemente los datos. Sin embargo con demasiado ajuste, el modelo se adapta muy cercano a los datos de entrenamiento, por lo que no generalizará de buena manera (por ejemplo, tener un error de testeo muy grande). En este caso las predicciones $\hat{f}(x_i)$ tendrán una alta varianza. En contraste, si el modelo no es suficientemente complejo, se sub ajustará y puede que tenga un alto sesgo, con lo que nuevamente resultará una generalización pobre.

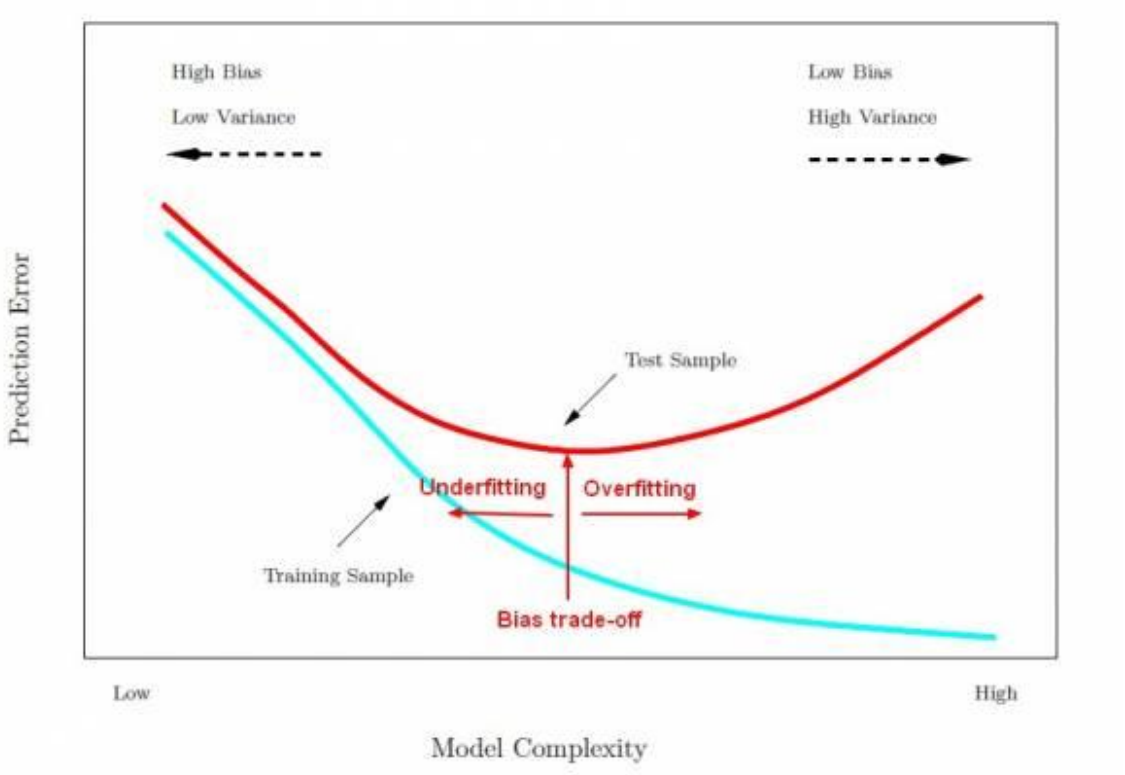

Figura 5: Error de testeo v/s Error de entrenamiento

2.3.6 Procesos de Extracción, Transformación y Carga

Los procesos de extracción, transformación y carga o ETL (por sus siglas en inglés) son procesos que tienen como objetivo facilitar el movimiento de datos y la transformación de los mismos, integrando distintos sistemas y fuentes de la organización y cargarlos en otras bases

de datos o sistemas operacionales para apoyar un proceso de negocio, aunque lo que normalmente se realiza es alimentar bodegas de datos (*data warehouse* o *data mart* en inglés), las que son modelos de datos orientados al análisis donde los datos representan indicadores (medidas) que pueden ser observados de acuerdo a ejes de análisis (dimensiones), en la Figura 6 se muestra diagramado un proceso ETL genérico.

Las fases del proceso se dividen en:

- Extracción: Recolección de los datos desde uno o varios sistemas fuente.
- Transformación: Consiste en aplicarle cambios a dichos datos, es decir, posibilidad de reformatear y limpiar estos datos cuando sea necesario.
- Carga de datos: Es el proceso de cargar en otro lugar o base de datos, un *data mart* o un *data warehouse*, con el objeto de analizarlos o apoyar un proceso de negocio.

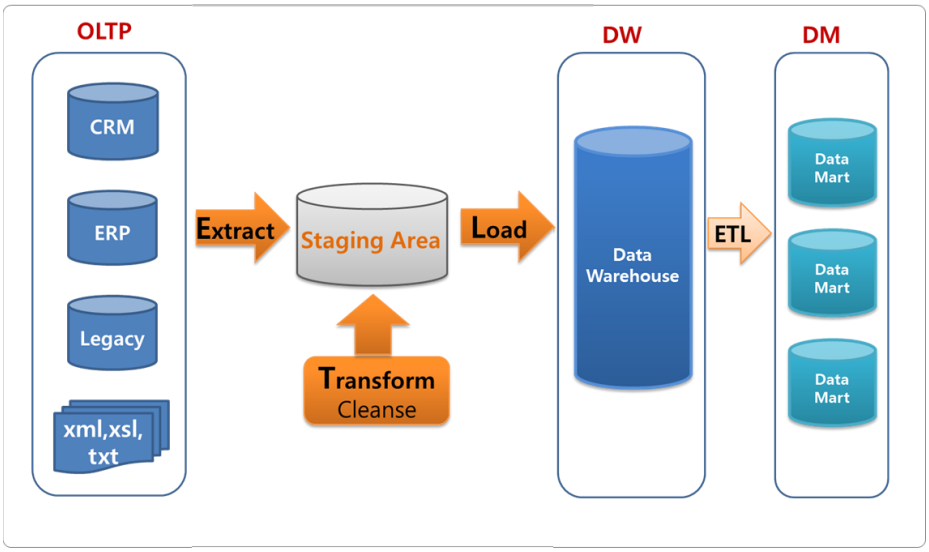

Figura 6: Proceso ETL

La limpieza se podría entender como una fase integrada dentro de la transformación de datos, pero en la actualidad se considera a la limpieza de datos como una fase separada del proceso ETL.

Es de alta importancia tener la información consolidada, que todos los datos sean correctos y con una visión única para todos los usuarios. Solo así se pueden lograr un buen ambiente de trabajo y poder realizar un buen análisis de los datos realmente óptimos y efectivos.

CAPÍTULO 3: PLANTEAMIENTO ESTRATÉGICO Y MODELO DE NEGOCIOS

3.1 POSICIONAMIENTO ESTRATÉGICO

3.1.1 Análisis Externo

Para estudiar la intensidad de la competencia y el potencial de atractivo de una industria se utiliza el modelo de las cinco fuerzas de Michael Porter [15]. A continuación se muestra el análisis para cada una de las fuerzas de Porter:

1.- Amenaza de nuevos participantes y Barreras de Entrada

- Economías de escala bajas debido a la alta caracterización de productos y servicios. ✖ [1]
- Bajas barreras para evitar la copia o imitación de la idea de los productos. ✖
- Baja identificación de marca. ✖
- Bajos costos de iniciación, capital mayoritariamente intelectual. ✖
- Conocimientos necesarios de constante última tecnología. ✔

Barreras de Salida

- Costo de salida bajo, se limita básicamente a recurso humano. ✔
- Restricciones Gubernamentales inexistentes. ✖

2.- Amenazas de Sustitutos

- Productos y servicios de difícil diferenciación para el cliente. ✖
- Actualmente no abarcan el conjunto de necesidades del cliente. ✔
- Sin soluciones inteligentes. ✔

3.- Poder de Negociación de Clientes

- Cliente constituye un nuevo Nicho. ✔
- Bajo número de competidores considerados importantes, por lo que el cliente tiene poco poder de negociación. ✔

[1] La cruz simboliza que lo declarado hace menos atractiva a la industria para la empresa, el ticket simboliza que lo declarado hace más atractivo a la industria y si está en blanco simboliza que tiene un efecto neutro.

- Clientes aislados, las universidades no conversan entre ellas. ✔
- Baja amenaza de integración hacia atrás, es decir, es poco probable que las universidades generen un área de inteligencia que produzcan algoritmos sofisticados que ayuden a la gestión. ✔
- Plazos extensos en el cierre de requerimientos y ejecución de nuevos proyectos. ✖

4.- Poder de Negociación de proveedores

Los proveedores son básicamente empresas proveedoras de software a través de licenciamiento y consultores que ayudan con conocimiento. En este caso la empresa no tiene proveedores, pero en un futuro podría llegar a necesitarlos.

- Bajas amenaza de los proveedores de integración hacia adelante. ✔
- Alta contribución de los proveedores a la calidad o servicio. ✖

5.- Intensidad de la rivalidad de la competencia:

- Industria con crecimiento lento pero sostenido.
- Mercado poco saturado. ✔
- Bajo costos fijos. ✔

Resumen del Atractivo de la Industria

Barreras de Entrada	Poco Atractivo
Barreras de Salida	Neutro
Rivalidad entre los competidores	Atractivo
Poder de los compradores	Atractivo
Poder de los proveedores	Neutro
Disponibilidad de sustitutos	Atractivo

Tabla 1: Resumen del atractivo de la industria

Evaluación General

Al mirar todas las aristas de las fuerzas de la industria se concluye que en aspectos generales la industria de la administración y gestión de recursos de instituciones de educación superior posee un alto atractivo para la empresa.

3.1.2 Modelo Delta

Para poder analizar el posicionamiento estratégico se utiliza el modelo Delta creado por el profesor chileno Arnoldo Hax [11] el cual busca cambiar la mirada de estrategia en las empresas desde una centrada en lo que la organización desea ofrecer, a otra cuyo foco es la necesidad de los clientes, por supuesto sin olvidar la competencia y a partir de allí construir un plan de desarrollo.

Hax plantea que la estrategia de las empresas se concentraba simplemente en derrotar a la competencia, lo que generaba solo pérdidas de valor ya que la competencia se transformaba en un juego de suma cero, es decir, lo que gana un participante es a expensas de lo que pierden los otros debido a que se trata sólo de aprovechar y optimizar la demanda existente. Lo que termina en casi siempre en una guerra de precio o un constate esfuerzo por mantener la diferenciación.

El autor postula que el cambio de visión a una estrategia centrada en el cliente, entrega beneficios que se ven reflejado en relaciones más duraderas y de mayor confianza, que finalmente entregarán una mayor ganancia de valor para la empresa. Para esto se debe tener un gran conocimiento de sus propios clientes que permita ajustar la propuesta u oferta de valor a sus necesidades particulares.

El modelo plantea tres opciones de posicionamiento que se representan como vértices en la pirámide de Hax:

- Mejor producto: El cliente se siente atraído por las características inherentes del producto ofrecido. Lo que puede reflejarse en un bajo costo o mediante la diferenciación del producto o servicio.

- Solución total: La relación con el cliente resulta de mejorar las capacidades del cliente al ofrecerle una solución integrada que aborda sus necesidades críticas. Se logra por una proximidad con el cliente, transfiriendo capacidades y conocimientos centrales y entregando un espectro completo de productos y servicios que satisfacen la mayoría, si no todas, sus necesidades.

- Dominio total: La firma logra una posición dominante en el mercado que le garantiza un liderazgo sin par, en donde los costos de cambio para un cliente son muy altos.

Figura 7: Pirámide de Hax

En la Figura 7 se muestra planteamiento estratégico bajo la mirada de Hax en el que se posiciona la empresa es el de mejor-producto, esto debido a que la empresa concentra sus esfuerzos principalmente en entregar productos y servicios de alta calidad, con el fin de que el cliente valore esta oferta y generar una relación con éstos basado en la confianza de la calidad y experiencia de la empresa en el rubro.

La calidad que entrega U-planner se ve reflejado principalmente en tres aspectos: la constante actualización de sus tecnologías, el talento de las personas que trabajan dentro de la empresa y el conocimiento empresarial que se ha obtenido para resolver las necesidades de las universidades.

Cabe destacar que la empresa está realizando esfuerzos por migrar a un nuevo posicionamiento estratégico hacia servicio integral con el cliente, esto se puede ver en las iniciativas que se han ido implementando dentro de la empresa, como por ejemplo incluir nuevos productos que están enfocados en resolver problemas que las universidades han tenido a lo largo de su historia más allá de la asignación y la optimización. Como algunos productos que ayudan a conseguir la acreditación de las universidades y de las carreras, productos que ayudan a generar mayor comunidad en la universidad y productos que ayudan en la gestión de proyectos de investigación.

Por otro lado se están enfocando esfuerzos en transformar la aplicación principal de U-planner en una basada en módulos en donde cada producto está construido modularmente, lo que significa que el cliente podrá solicitar los productos que resuelvan sus problemas específicos, sin tener que solicitar todo el conjunto de productos de la línea y en caso que requiera un nuevo producto pueda solicitar fácilmente su activación.

3.2 BALANCED SCORECARD

El cuadro de mando integral o balaced scorecard (BSC) propuesto por Robert S. Kaplan y David P. Norton [17] surge debido a la necesidad de suplir la insuficiencia que generaban los indicadores financieros por si solos para llevar a las compañías al éxito, es por esto que se hace necesario analizar nuevas aristas que entreguen mayor información de desempeño real de toda la compañía.

Así nace el Balanced Scorecard que entrega un conjunto de indicadores que reflejan el valor total del crecimiento que las empresas experimentan en un determinado periodo. Esto ayuda a complementar la visión financiera con aspectos importantes en la empresa tales como la relación con el cliente, mejorar los procesos e innovación y en la capacidad de empleo y la motivación.

Por lo que se puede mirar el balanced scorecard como un conjunto de mediciones más robustas que va más allá de las finanzas para crear un beneficio en el futuro, utilizando una mirada en las finanzas, en los clientes, operaciones e innovaciones.

La empresa no cuenta con un BSC creado, pero a partir de conversación con los directivos y gerentes se pudo obtener los principales objetivos en cada una de las perspectivas para este momento de la empresa.

Lo que se busca principalmente son dos objetivos: vender una mayor cantidad y aumentar el crecimiento de la empresa. Esto se pretende lograr mediante el cumplimento de objetivos específicos o secundarios que finalmente ayude en el cumplimento de los objetivos financieros. Estos objetivos se muestran a continuación y están creados bajo el concepto SMART.

Objetivos Financieros

- Vender 4 millones de dólares el 2016
- Subir a 30 millones de dólares anuales al 2020.

Clientes

- Crecer la cantidad de clientes a 50 el 2016, y alcanzar 200 clientes al 2020.
- Consolidar el mercado escolar para el año 2017.
- Ser reconocidos por el buen servicio brindado al cliente.

Operacionales

- Ser líderes en LATAM durante el 2016.

- Expandir a USA y Europa durante el 2017.
- Lograr atender por internet a todos los clientes globales el 2017.

Innovación

- Realizar un cambio en la arquitectura tecnológica para permitir conectar los módulos entre sí, lo que permitirá al usuario escoger diferentes algoritmos para cada producto para el año 2017.
- Adaptar los productos para que puedan ser vendidos a través de internet para el año 2017.

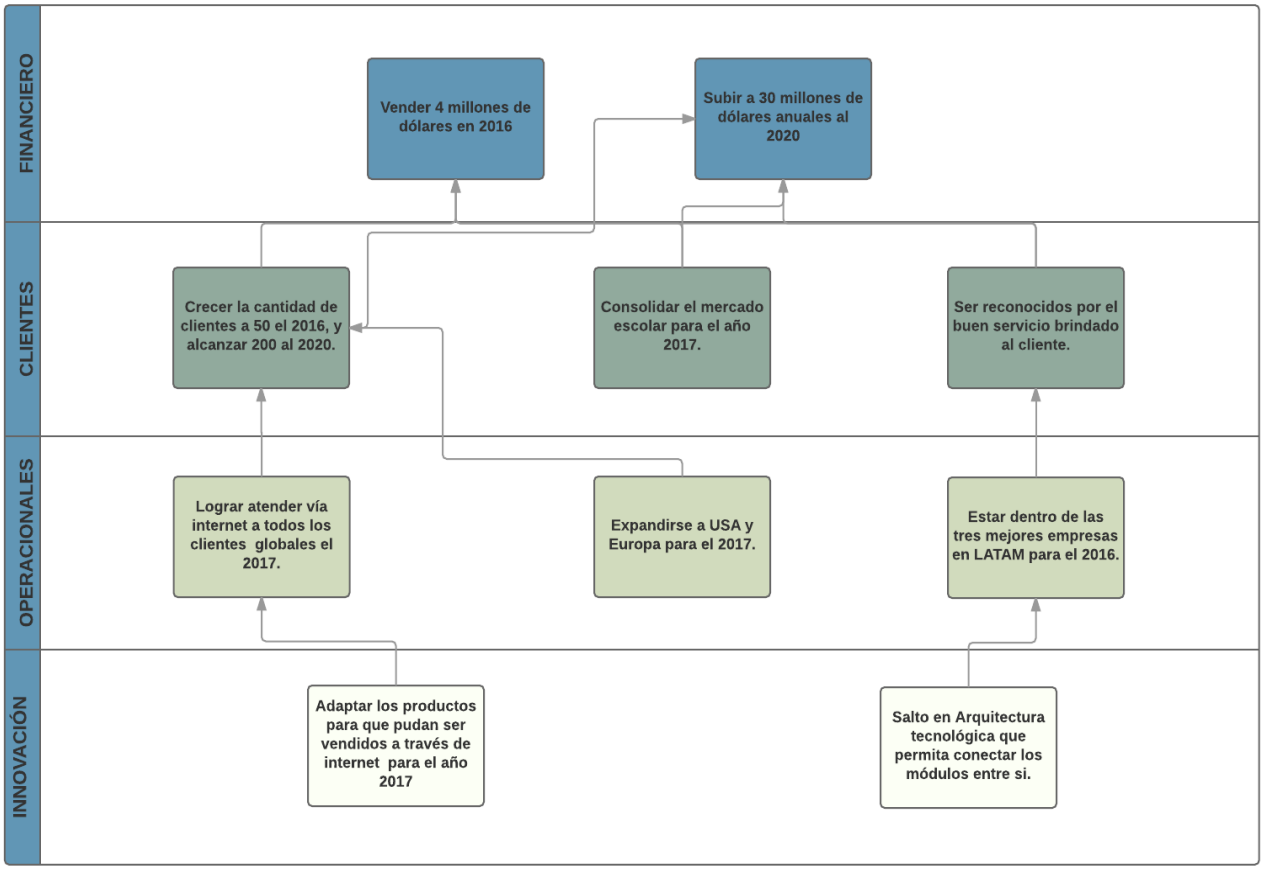

Figura 8: Balanced Scorecard

Se puede observar en la Figura 8 las relaciones de causa-efecto que se producen en los objetivos estratégicos. En la mayoría de los objetivos estratégicos, sobre todo los operacionales y de innovación contribuyen con el plan estratégico de mejor producto, éstos ayudan a la diferenciación, tanto en entregar un mejor de servicio como mejorar calidad de los productos. Cumpliendo estos objetivos se busca llegar a cumplir los objetivos financieros.

El proyecto abordado en esta tesis ayuda a generar una alternativa para clientes a los que son difíciles o no se puede atender por su estructura de mallas lo que contribuye al objetivo de "Crecer la cantidad de clientes a 50 el 2016, y alcanzar 200 clientes al 2020", para finalmente cumplir con los objetivos financieros.

3.5 MODELO DE NEGOCIOS

El modelo de negocio es el mecanismo a través del cual una empresa crea valor y como consecuencia de ello genera ingresos y beneficios. El modelo de negocio describe lo que un negocio ofrece a sus clientes, cómo se relaciona, cuáles son los canales para entregar el producto o servicio y además define cuáles son los procesos y recursos críticos del negocio. Para esto se utilizará la metodología Canvas [14].

A continuación se describen las componentes más importantes para el negocio de U-planner:

- Propuesta de valor: Lo que ofrece la organización es una solución completa a los problemas de planificación de los horarios, asignación de recursos humanos e infraestructura, retención, acreditación e investigación lo que permite el mejoramiento continuo en competitividad y prestigio de las instituciones. El producto se caracteriza por soluciones estables, calidad en el servicio brindado, comunicación oportuna y efectiva con el cliente. Para esto existen tres modalidades de ventas: la primera a través de licencias, la segunda por medio de *Software as a service* (SAAS)[2] y la tercera es mediante consultorías.

- Clientes: Los clientes para U-planner son las universidades, institutos y otros establecimientos educacionales, que tengan dificultades y requieran apoyo en planificaciones, optimización de procesos internos y/o mejorar la experiencia educativa de los alumnos, profesores y funcionarios.

- Relación con el Cliente: Es a través de una asistencia personal en donde se tiene una constante comunicación con el cliente para obtener sus datos y para luego realizar una muestra de las capacidades que tiene la empresa para ayudar a la universidad. También se le muestra al cliente las virtudes y espacios de mejora. Y finalmente el cliente decide si realizar la compra.

- Canales: La comunicación con los clientes es a través del área comercial y servicios, vía teléfono y web.

- Actividades claves: Las actividades principales que requiere la propuesta de valor son la actualización de los productos, integración con el cliente, actividades de soporte y mantenimiento, venta de los productos y atención de nuevos requerimientos de los clientes.

- Partners Claves: Agentes comerciales que ayudan a la venta de los productos y al networking y universidades que funcionan como partners estratégicos y ayudan con investigación.

[2] Software as a service (SAAS) es un modelo de distribución de software en el que un proveedor aloja las aplicaciones y los pone a disposición de los clientes a través de Internet.

Recursos claves: Para poder cumplir con la propuesta de valor los recursos principales que se necesitan son los recursos humanos como técnicos, diseñadores, ingenieros y consultores de alto nivel.

- Flujo de Ingresos: Estos se generan a partir de la venta de los principales productos y de los productos complementarios. Además existe un cobro diferenciado por cada producto y por cada cliente ajustándolo a la cantidad de alumnos que tiene la institución.

- Estructura de Costos: Los costos más importantes en relación al modelo de negocio es el recurso humano. Otro costo considerable son los servidores y licencias de softwares que se necesitan para el día de trabajo.

CAPÍTULO 4: ANÁLISIS DE LA SITUACIÓN ACTUAL

4.1 ARQUITECTURA DE PROCESOS

Para la empresa U-planner se ha diseñado una representación de sus procesos, basado en los patrones de diseño definidos por Cristian Julio y Óscar Barros [12], así es como se forma la arquitectura de procesos que está conformada por cuatro macro-procesos: planificación del negocio, desarrollo de nuevas capacidades, la cadena de valor y gestión de recursos habilitadores.

Cabe destacar que la principal actividad de la empresa es la venta de soluciones para las universidades lo que significa que cuenta con una única cadena de valor. Aterrizando al contexto de las actividades de la empresa se le ha cambiado el nombre por "desarrollo de soluciones integrales para la gestión académica de instituciones educativas".

A continuación en la Figura 9 se muestra la arquitectura de la empresa comprimida en una sola actividad (Macro 0):

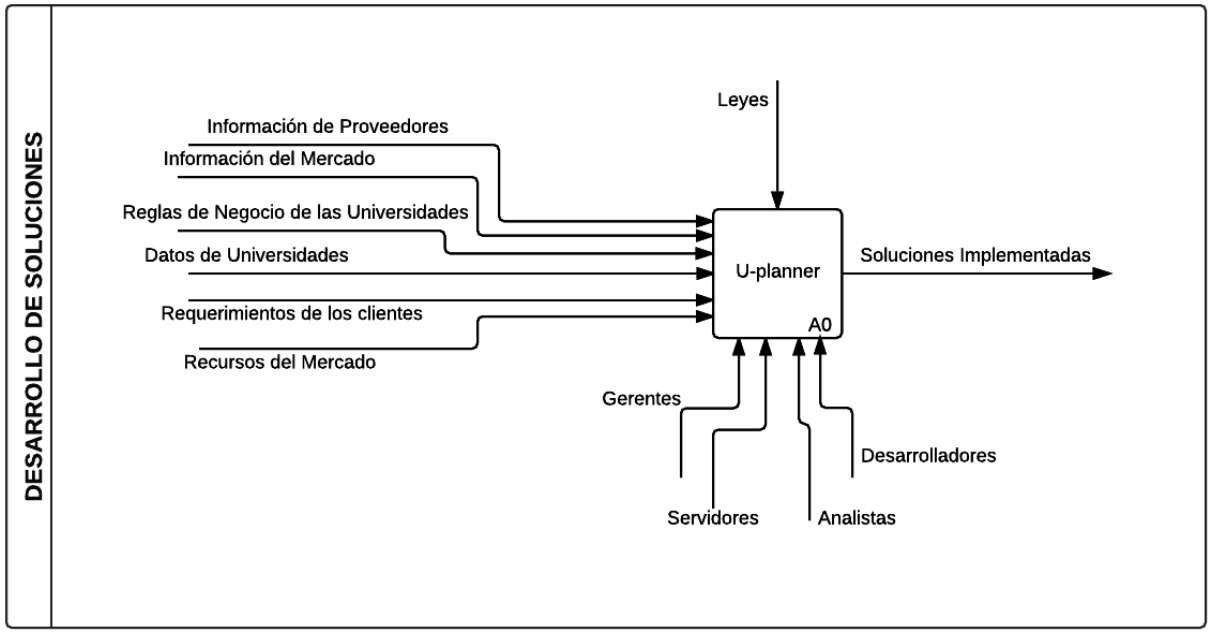

Figura 9: Arquitectura comprimida de U-planner

En la Figura 10 se muestra el diseño de la arquitectura de la empresa en donde se esquematizan los cuatro macro-procesos principales que a continuación se detallan.

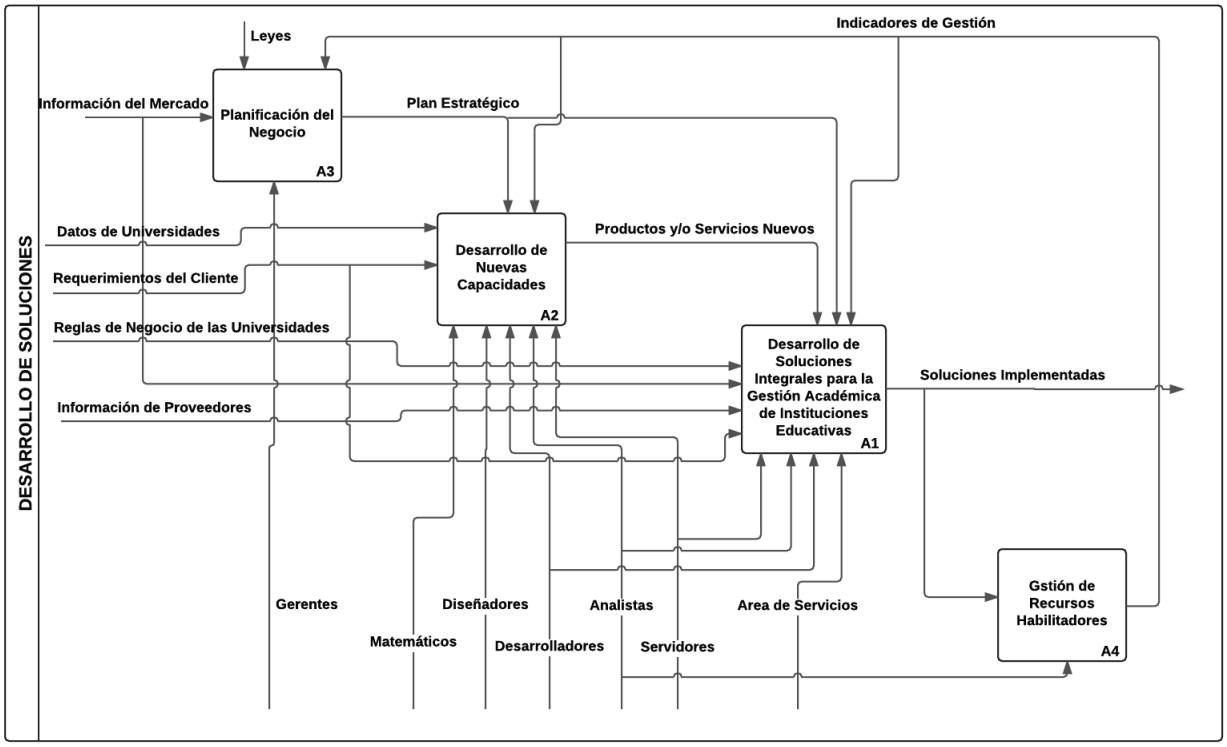

Figura 10: Macros-proceso de la Arquitectura de U-planner.

La planificación del negocio (Macro 3): representa todas las actividades en donde se decide el rumbo que seguirá la empresa en el futuro. Se realizan reuniones semanales y diarias para distintos niveles jerárquicos para guiar los esfuerzos operacionales y alinearlos con los objetivos estratégicos de la empresa.

Recibe como *input* información del negocio, como *output* entrega un Plan Estratégico, como recursos necesita gerentes y como condiciones está restringido por las leyes y por los indicadores de gestión de la Macro 4.

El desarrollo de nuevas capacidades (Macro 2): corresponde a todas las actividades que generan nuevos productos para la empresa. U-planner es una empresa que actualmente está en constante crecimiento, en un mercado poco trabajado por lo que una de sus actividades principales es la identificación de nuevas necesidades en los clientes y generar un producto que las satisfaga, para esto se destina a un grupo de personas y esfuerzos para poder realizarlo.

En estas actividades se recibe como *input* datos e información de las necesidades o problemas con los que lidian las universidades, se entrega como *output* nuevos productos o servicios, se necesitan como recursos desarrolladores, diseñadores, gerentes, analistas, matemáticos y servidores.

La cadena de valor (Macro 1): permite describir el desarrollo de las actividades de una organización empresarial que generan valor al cliente final, es nombrada como "desarrollo de soluciones integrales para la gestión académica de instituciones educativas". Se le brinda este nombre porque se produce finalmente información que agrega valor para la toma de decisiones para los principales problemas de las universidades, esto se puede realizar mediante tres maneras: la venta de licencias, SAAS o consultorías por lo que se necesitan distintos requerimientos para cada tipo de venta. La información se genera mediante la ejecución de algoritmos de alta complejidad y posteriormente se recibe un pago que varía de acuerdo a la forma de venta, tamaño de la universidad y productos que se solicite. También existen actividades de post-venta en donde se realiza mantención y soporte de las aplicaciones.

En el desarrollo de soluciones integrales para la gestión académica de instituciones educativas se toma como *inputs* los requerimientos del cliente, los datos y reglas de negocio de las universidades. Son necesarios como recursos desarrolladores, analistas, área de servicios y los servidores. Como condiciones se tiene a los productos creados entregados por nuevas capacidades y los indicadores de gestión que debe cumplir y el plan estratégico a seguir.

Gestión de Recursos (Macro 4): Conjunto de procesos de apoyo que maneja los recursos necesarios para que los anteriores procesos puedan operar.

Recibe como *input* las soluciones implementadas y recursos del mercado, entrega como *outputs* parámetros de gestión, como recursos necesita analistas y como condiciones se ajusta al plan estratégico.

El proyecto está enfocado dentro de la Macro 1 y como se describirá más adelante afectará el proceso de producción y gestión para uno de los productos.

4.2 MODELAMIENTO DETALLADO DE LA ESTRUCTURA DE PROCESOS

El foco del proyecto de tesis consiste en bridarle una alterativa al proceso de producción de información del producto U-Forecast de manera de poder satisfacer a nuevos tipos de clientes, lo que tendrá una repercusión directamente en la cadena de valor, es decir, en el desarrollo de soluciones integrales. Es por esto que a continuación se detalla la apertura de la cadena de valor:

Administración de la relación con instituciones educativas: esta actividad corresponde a todas las actividades del seguimiento del cliente y la venta, la post venta y el soporte de las aplicaciones y actividades para definir como se planea satisfacer al cliente.

Esta actividad, como se muestra en la Figura 11, recibe como *inputs* la información del mercado, como *output* entrega información acerca del cliente y los procesos de venta, como recursos se necesitan analistas comerciales, gente de soporte y servicios y la información del estado, como restricciones se tiene a los prototipos y el plan estratégico.

Administración de relación con consultores: la empresa no tiene proveedores que afecten directamente a la cadena de valor, pero se puede considerar a consultores que entregan recursos que ayudan a los procesos de la empresa. Esta actividad comprende todos los procesos que tienen relación con obtener una retro alimentación de los consultores que son una gran fuente de información para la empresa.

Esta actividad recibe como *input* información sobre los proveedores, como *output* entrega *feedbacks* de los proveedores, como condición tiene el plan estratégico y como recursos necesitan analistas e información de estado como se muestra en la Figura 11.

Gestión del desarrollo de soluciones para la gestión académica: esta actividad es la encargada de gestionar los procesos de producción y entrega de la información a los clientes. En este paso se realizan planificaciones de trabajo, integraciones con clientes, coordinaciones de equipo y productos y se miden los resultados y plazos de tiempo. Este proyecto afecta esta actividad al crear un nuevo proceso para mantener los modelos de predicción y al entregar indicadores de rendimiento para el producto de U-Forecast.

Detallando la diagramación en IDEF0 de la Figura 11 recibe como *input* información y data del cliente y entrega la programación y estándares de la producción e implementación e ideas de cambio. Para esto tiene como restricciones los prototipos, el plan estratégico, los mensajes de requerimientos entregado por Administración de establecimientos educacionales e indicadores de Gestión. Finalmente como recursos se necesitan desarrolladores y servidores.

Generación y entrega de soluciones para la gestión académica: abarca las actividades de producir la información con las características particulares para cada universidad y entregar dicha información en los sistemas que se estimen convenientes. El proyecto de tesis entrega una nueva lógica de estimación para el producto U-Forecast lo que implica cambios en cómo se producirá esta información, lo que afectará el proceso de producción. Este proceso engloba desde el levantamiento de datos de la universidad hasta la entrega de la información de los resultados.

Como se muestra en la Figura 11 el concepto de la generación y entrega que recibe los datos de las universidades y entrega las soluciones implementadas, está restringida a la

programación y estándares de la producción e implementación e ideas de cambio y necesita desarrolladores y servidores.

Mantención del estado: que comprende todas las actividades que dan algún apoyo computacional para guardar información útil de los procesos de la empresa. Está actividad recibe como *inputs* los cambios de estado de los procesos y entrega como *output* la información de estado actualizada.

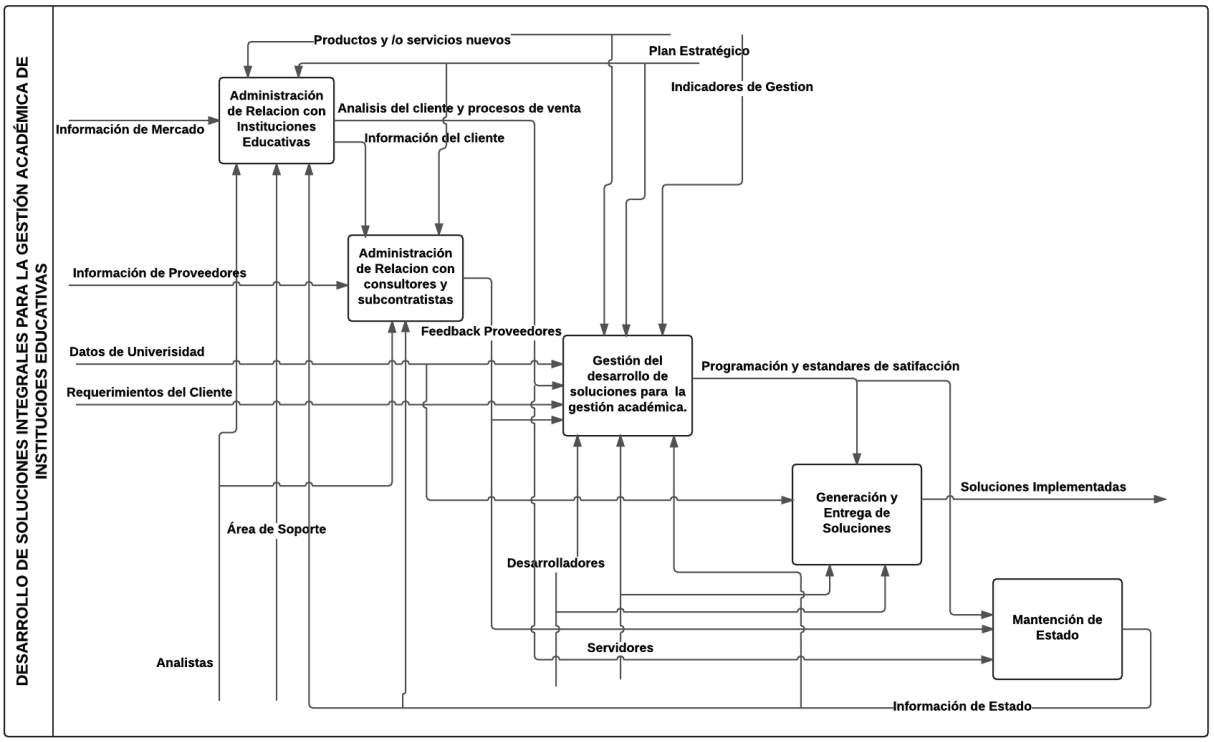

Figura 11: Desarrollo de Soluciones Integrales para la Gestión Académica de Instituciones Educativas.

4.2.1 Gestión del desarrollo de aplicaciones para la gestión académica

La gestión del desarrollo de aplicaciones para la gestión académica se separa en los siguientes grupos de procesos:

Implementación de nuevos productos y sus características: son las actividades que tienen relación con la integración de los productos en los clientes. Para poder realizar la entrega de la información los productos deben tener una buena comunicación con las fuentes de datos de las universidades para poder recibir y entregar la información.

Se muestra en la Figura 12 esta actividad, la que recibe como *inputs* los requerimientos de los clientes, datos de las universidades, información acerca del cliente y *feedbacks* de los consultores, como *output* entrega el diseño de los productos y sus características, necesita como recurso servidores, desarrolladores e información del estado y tiene como condiciones los productos y/o servicios nuevos de nuevas capacidades, el plan estratégico de la empresa e indicadores de gestión que deben cumplir.

Monitoreo y planificación de la generación de información y la entrega: corresponden a las actividades que planifican las integraciones y ejecuciones de los distintos productos y servicios. El equipo prioriza a los clientes y a los productos para realizar una planificación de actividades, determinar los recursos humanos e insumos necesarios para dichas actividades y desarrolla una forma de abordar las características específicas de cada integración. Por otro lado se miden y analizan las métricas de gestión de la producción y entrega de la información.

Esta actividad recibe como *input* el diseño de los productos y sus características, los datos de las universidades, información del cliente y *feedbacks* de los consultores, entrega como *output* la planificación e instrucciones de monitoreo de los procesos de entrega y generación de información, como recurso necesita la información del estado, servidores y desarrolladores y como condiciones se debe ajustar al plan estratégico y a los indicadores de gestión que se hayan generado.

La decisión de entrega de información: Son las actividades en donde se programa el detalle de la entrega de información a los clientes.

Esta actividad recibe como *input* la planificación e instrucciones de monitoreo de los procesos de entrega y generación, entrega como *output* la programación y estándares de la producción e implementación e ideas de cambio y necesita como recursos servidores y desarrolladores.

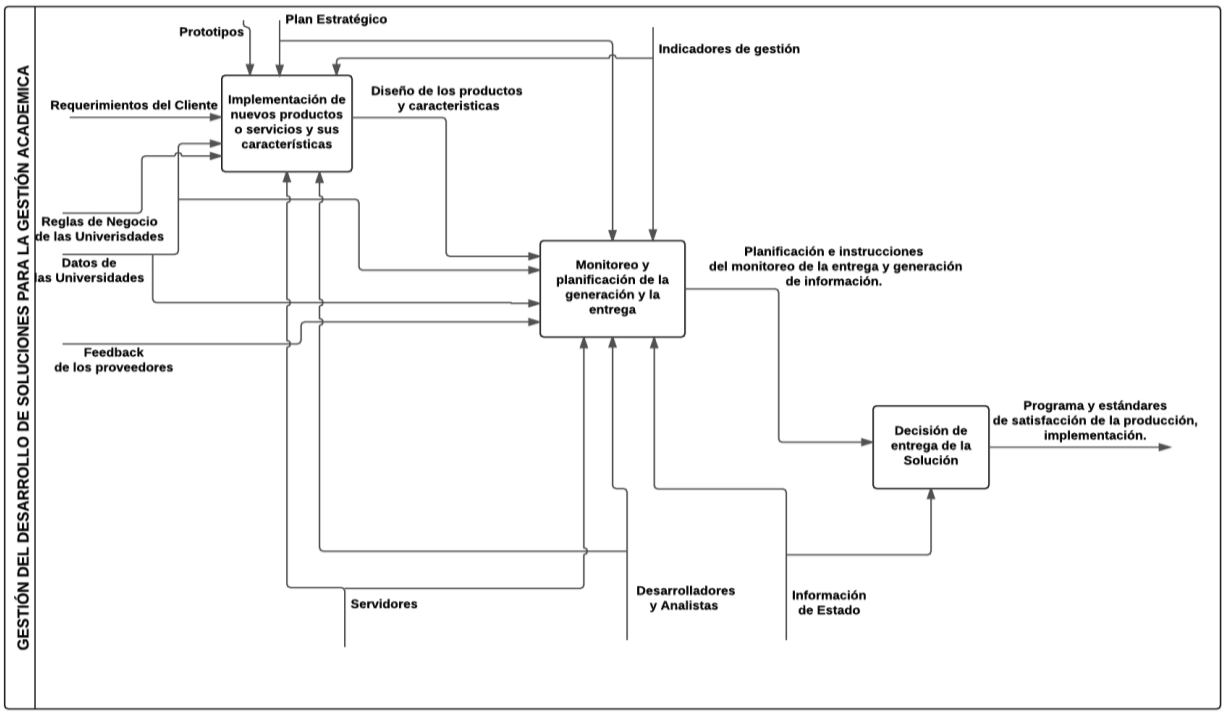

Figura 12: Gestión del Desarrollo de Soluciones para la gestión académica.

4.2.1.1 Generación y Entrega

Generación y Entrega de las soluciones se separa en las siguientes actividades

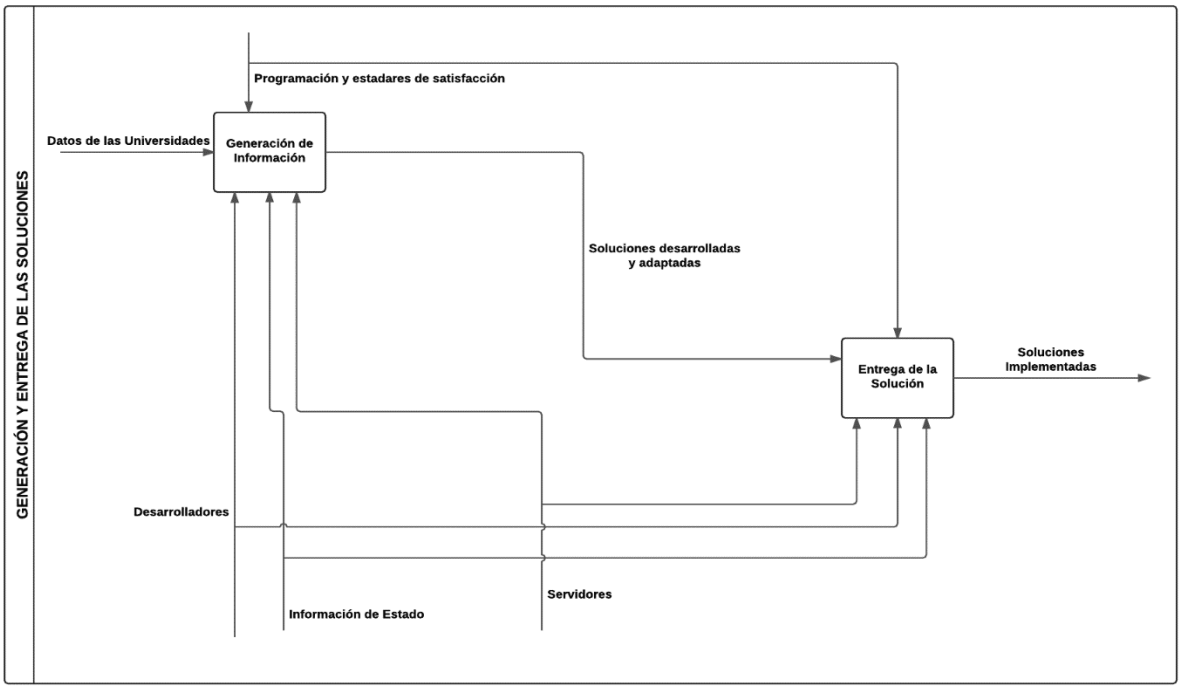

Figura 13: Generación y Entrega de Soluciones

Generación de Información: es el proceso que contiene la creación de la información relevante para el cliente a través de los algoritmos. El presente proyecto de tesis generará el mayor impacto en este proceso al diseñar e implementar un nuevo algoritmo para la estimación de la demanda del producto U-Forecast.

Recibe como *input* los datos de las universidades e insumos necesarios para la producción y entrega como *output* las soluciones desarrolladas y adaptadas en el formato que quiere el cliente, necesita desarrolladores, información del estado y servidores como recursos y se ajusta a las condiciones de programación y estándares de satisfacción, generadas en gestión del desarrollo de aplicaciones.

Entrega: Esta actividad representa un proceso simple en donde se entrega la información final a los clientes junto a los reportes necesarios y/o reuniones explicativas.

Para esta actividad se recibe como *input* las aplicaciones adaptadas, se entrega como *output* las soluciones implementadas, se necesitan como recursos desarrolladores, información de estado y servidores, y se tiene como condición la programación y estándares de satisfacción.

Luego en la apertura del proceso de generación de información se tienen los siguientes procesos:

Creación del Prototipo: Este proceso se encarga de crear un prototipo para los clientes que por primera vez prueban un producto.

Para esta actividad, que se muestra en la Figura 14, se recibe como *input* la información de las universidades, se entrega como *output* el prototipo finalizado, se necesitan como recursos desarrolladores, analistas, información de estado y servidores, y se tiene como condición la programación y estándares de satisfacción.

Producción de Información: Este proceso es el encargado de generar la información relevante para el cliente a través del uso de algoritmos y modelos matemáticos de optimización y predicción. Usualmente se basan en los prototipos realizados, pero suelen necesitarse ajustes, además que la cantidad de información que se debe procesar aumenta.

Esta actividad, como se muestra en la Figura 14, es similar a la anterior en su diagramación, el *input* que se recibe son la información de las universidades y el prototipo finalizado, el *output* entregado son las soluciones desarrolladas y adaptadas, se necesitan como recursos

desarrolladores, analistas, información de estado y servidores, y se tiene como condición la programación y estándares de satisfacción.

Para ambos procesos este proyecto genera un aporte debido a que el algoritmo generado podrá ser usado tanto en la creación del prototipo como de la producción de información.

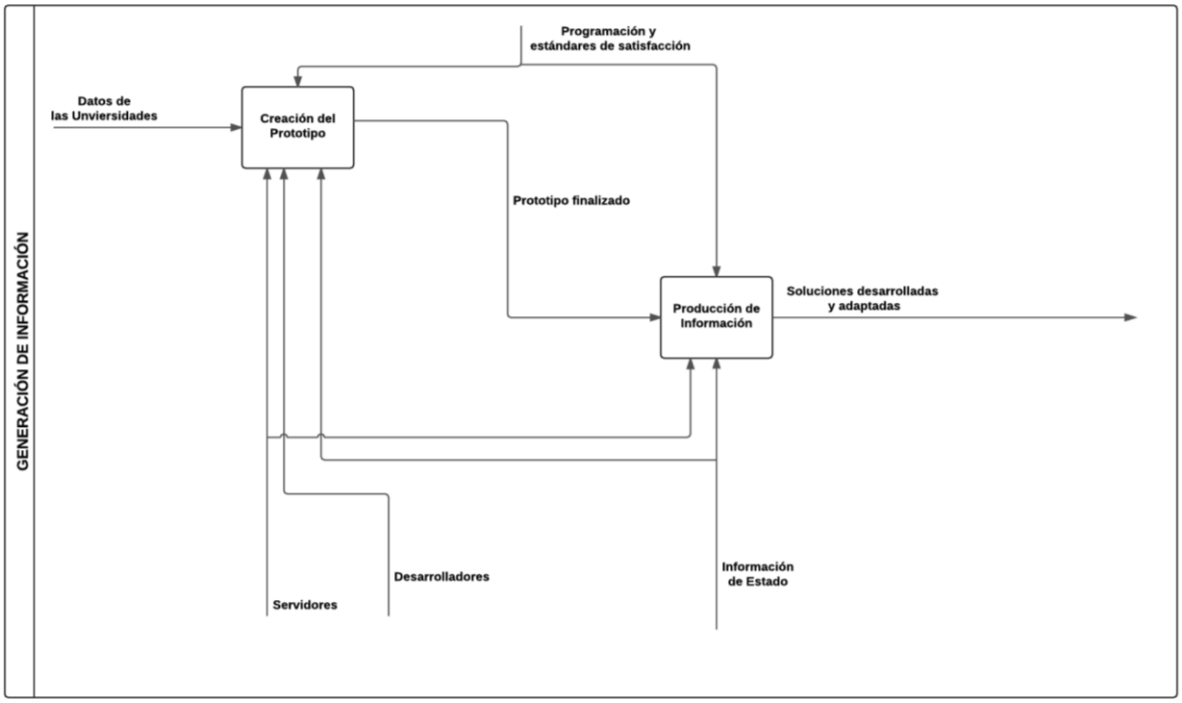

Figura 14: Generación de Información

Cabe mencionar que la forma de generar la información en U-planner consta de, en primera instancia, generar un piloto de prueba en donde se toman los requisitos y datos que tiene la universidad, luego tras varias conversaciones se llega a un acuerdo de los datos mínimos que debe entregar el cliente, niveles de consistencia y limpieza de estos mismos, se le entregan los layouts[3] finales, posteriormente se ejecuta una muestra del producto y si satisface las expectativas del cliente se da la posibilidad de seguir como una consultoría o integrarse al sistema.

[3] Layouts: Plantilla que esquematiza la forma de entrega de los datos para poder ser utilizados por los productos de U-planner.

4.3 DIAGNÓSTICO DE LA SITUACIÓN ACTUAL

El escenario que se presenta actualmente en la empresa es que en la producción de información de demanda de alumnos que inscribirán asignaturas en un periodo futuro (U-Forecast) queda limitado por las capacidades que actualmente posee el algoritmo.

Para poder realizar una venta de un producto se tiene una constante conversación con el cliente en donde se le realiza una prueba piloto proceso que se muestra en la Figura 16 para mostrarle las capacidades y utilidades de los productos y después el cliente decide si realiza la compra.

Luego de este proceso si es que se concreta la venta, posterior a la integración con la universidad, se ejecuta el proceso de producción de información que es muy similar al de la generación de la prueba piloto, este se muestra en la Figura 15.

En la situación actual se podrían llegar a generar ineficiencias dentro de ambos procesos en los casos en que el algoritmo actual no es capaz de entregar una buena estimación, lo que desencadenaría en que no se produzca la venta u optar por la alternativa de realizar una consultoría, en el caso de que sea factible, en donde se genere una solución particular para el cliente. En este último caso se tendrían que realizar esfuerzos especiales por parte de los analistas, desarrolladores y matemáticos para generar layouts, transformaciones y cambios en el código para poder adaptarse a la complejidad del cliente.

Las ineficiencias adicionales dentro de la generación de información con el algoritmo actual son el cálculo de los parámetros que identifican a las universidades, el cual actualmente se realiza de forma manual y se le entrega al modelo como un *input*. Este cálculo debe ser realizado por un analista matemático lo que toma alrededor de una semana poder estimarlos, lo que puede variar dependiendo de la experiencia del analista.

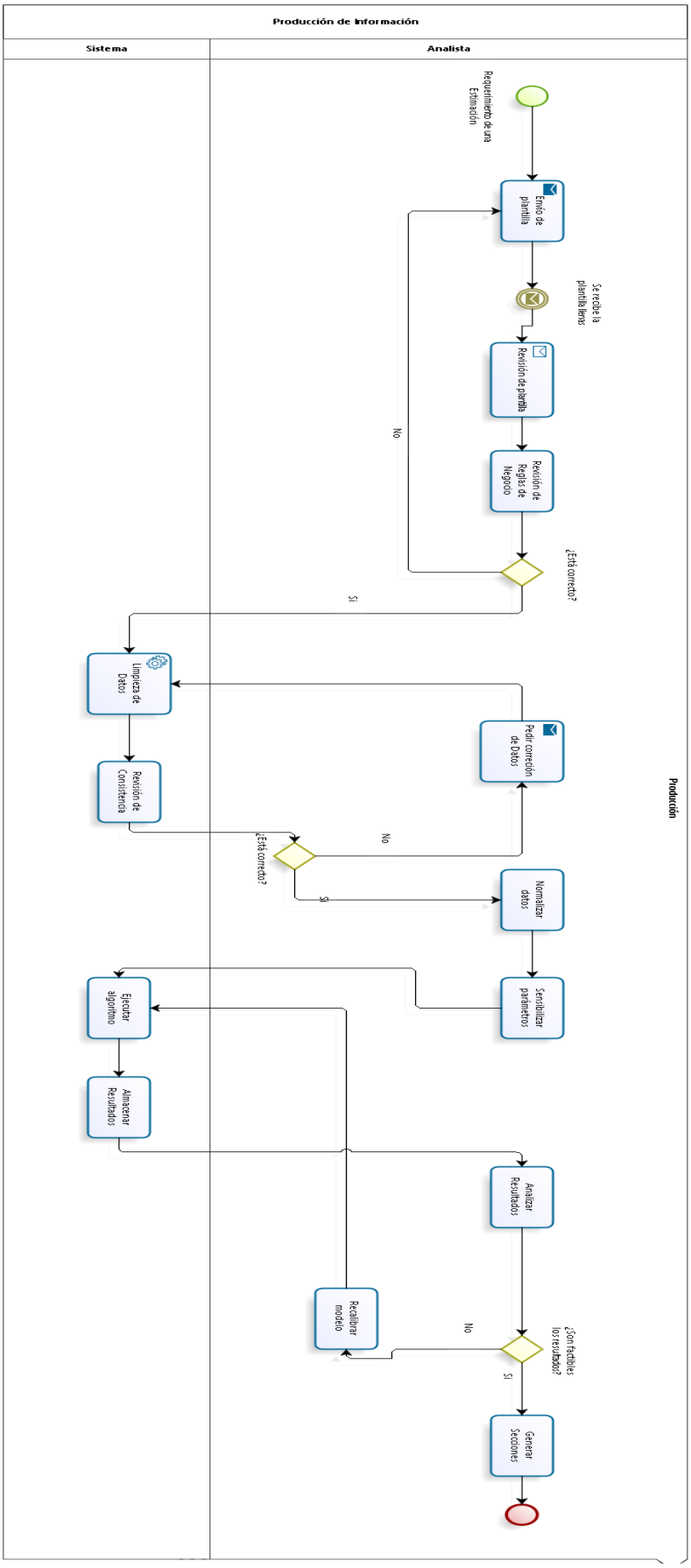

Figura 15: Proceso de Producción de Información

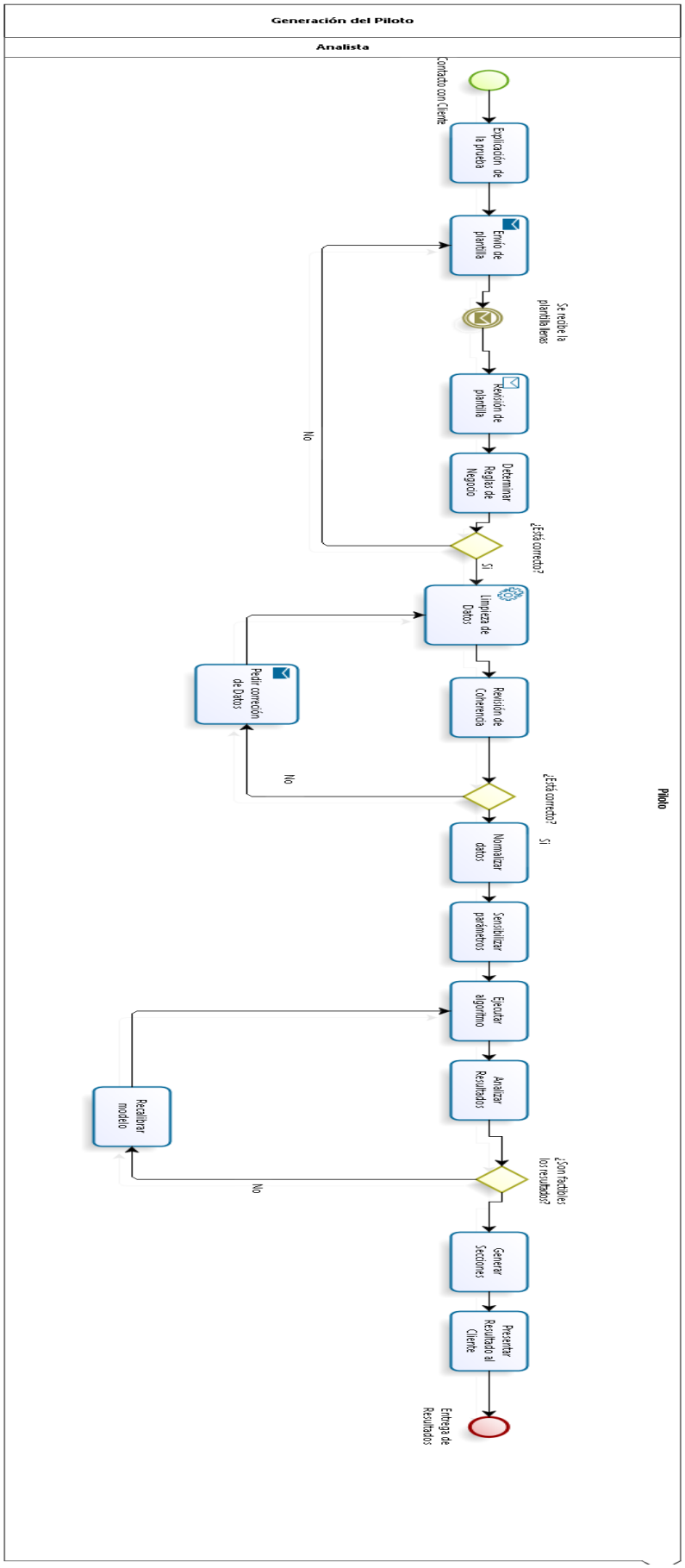

Figura 16: Proceso de Generación del Piloto

4.4 CUANTIFICACIÓN DEL PROBLEMA U OPORTUNIDAD

A partir de las ineficiencias del algoritmo de estimación de demanda anteriormente mencionadas es que surgen oportunidades de mejora.

El problema que generan estas ineficiencias es principalmente la pérdida de posibles ventas del producto U-Forecast.

Recientemente en dos oportunidades no se pudo realizar la venta de este producto, pero a medida que U-planner sigue en plan de expansión es probable que se encuentre con más casos complejos como éstos.

El mercado potencial dentro de Latinoamérica, según Webometrics Ranking of World Universities (2012) [21], es de 3.750 instituciones. Como no se conoce el número de universidades que tienen una estructura de mallas compleja es difícil hacer un cálculo exacto de lo que la empresa deja de percibir, por lo que para cuantificar la magnitud estos ingresos, basta con hacer el supuesto que tan solo el 1% de las instituciones sean del tipo complejas, lo que conllevaría tener problemas en el uso del producto U-Forecast, para que 2.096 millones aproximadamente de pesos anuales en ingresos no se pudiesen obtener jamás. Lo anterior se basa también en que lo experimentado por la empresa es que se han tenido pérdidas de oportunidades con 10 clientes debido a este problema, lo que equivale a 56 millones de pesos anuales aproximadamente en ingresos que no se pudieron aprovechar y considerando que actualmente la empresa cuenta con 20 clientes que utilizan el producto, este es un porcentaje bastante alto de clientes que se han perdido.

Por otro lado están los problemas operacionales que generan las ineficiencias del algoritmo, es decir, tiempos mayores en la integración con el cliente y generación de pilotos específicos. Además de tiempos mayores en la generación de información en los casos de consultoría.

Por esto surge la oportunidad de crear un nuevo algoritmo para estimar la demanda de cursos que pueda predecir mejor con universidades cuyos planes son muy flexibles y/o complejos, con el fin de manejar un plan alternativo al algoritmo actual que pueda brindar una solución más simple y rápida que pueda complementarse con el actual algoritmo de estimación de demanda.

CAPÍTULO 5: PROPUESTA DE DISEÑO DE PROCESOS

5.1 DIRECCIONES DE CAMBIO

Para mostrar la propuesta de rediseño se comenzará presentando las direcciones de cambio que tendrá el rediseño que busca principalmente ayudar a la empresa en el plan de ofrecer un servicio integral con el cliente específicamente ayudándola a ampliar su oferta en uno de sus productos. Para esto se creó un marco de referencia a partir de las variables de diseño que muestran la generación del cambio.

5.1.1 Anticipación

Esta variable detalla la planificación de actividades relevantes para los procesos productivos, con lo que busca anticiparse a eventos futuros.

Lo que se pretende para este proyecto es realizar una planificación de los modelos necesarios para las estimaciones que se realicen durante ese periodo. Esto es necesario debido a que en el proceso de mantención de modelos se genera más de un modelo por cada curso, por lo que se debe realizar una elección y correcta calibración de los modelos que se utilizarán para realizar las predicciones durante dicho semestre.

Anticipación	Actual	Propuesto
Planificación mantención de modelos.	No.	Planificación simple una vez cada semestre.

Tabla 2: Dirección de Cambio: Variable Anticipación.

5.1.2 Coordinación

Esta variable analiza la participación y colaboración de las distintas áreas de la empresa en los procesos internos. Son una alternativa más económica a la planificación del punto anterior, permitiendo una coordinación con más recursos de holgura.

La empresa cuenta con una aplicación para la gestión operativa de proyectos y una aplicación de versionamiento con la cual se pretende realizar una coordinación colaborativa en la que se entregan tareas y requerimientos necesarios para éstas y donde los empleados pueden subir los grados de avance que llevan en cada una. Se usarán las herramientas que ya se encuentran dentro de la empresa para ayudar a realizar los proyectos de producción de información de demanda.

Coordinación	Actual	Propuesto
Reglas	Informales	Reglas formales y estandarizadas para los procesos re-diseñados.
Jerarquía	Se utiliza en casos excepcionales de falla.	Mantener situación actual.
Colaboración	Colaboración intensiva mediante la utilización de software de gestión de proyectos.	Generar proyectos dentro del software con los responsables necesarios.
Partición	No.	Mantener situación actual.

Tabla 3: Dirección de Cambio: Variable Coordinación

5.1.3 Prácticas de Trabajo

Las prácticas de trabajo materializan y detallan las opciones de diseño expresadas en los puntos anteriores. Ellas deben permitir ejecutar las tareas del proceso de manera que se cumpla con tales diseños.

El proyecto utilizará la misma arquitectura tecnológica que existe en la empresa por lo que las actividades que realizan los empleados para los procesos de integración con el cliente y ejecución de los algoritmos cambiarían relativamente poco, pero los procedimientos de comunicación e integración entre los trabajadores dejarían de realizarse tácitamente y se cambiarán a un proceso formal.

Actualmente dentro de la empresa la lógica de estimación de demanda es automática y eficiente en términos algorítmicos, es por esto que al cambiar la lógica de estimación se mantienen las características de automaticidad y eficiencia.

Se apoyará a las actividades tácitas de sensibilización de parámetros para los modelos la que se realizará de forma semi-automática a cargo de un analista una vez por semestre.

Se entregará indicadores de rendimiento de las simulaciones, detallados en el capítulo 6.4.1, por lo que los analistas y matemáticos podrán medir las diferentes simulaciones que se vayan generando y entregar a los clientes mayor seguridad sobre las predicciones.

Junto al rediseño se necesita que las personas involucradas en las prácticas de trabajo internalicen las lógicas de negocio para poder ejecutar los procesos de transformación de datos, mantención de modelos y análisis de la información obtenida.

Práctica de Trabajo	Actual	Propuesto
Lógica de negocios automatizada o semi-automatizada		
▪ Predicción de demanda.	Utilización de lógica automática de predicción de demanda.	Nueva lógica automática de predicción de demanda.
▪ Lógica de Limpieza de datos.	Lógica simple, pero automatizada	Se agregan lógicas más complejas y se mantiene la automatización
Lógica de apoyo a actividades tácitas.	No.	Lógica semiautomática de sensibilización de modelos predictivos.
Procedimientos de comunicación e integración.	Existe una idea de cómo debiese ser el proceso.	Definición de flujos de trabajo y comunicación.
Lógica y procedimientos de medición y control.	No se cuenta con indicadores para medir el rendimiento del producto.	Se proponen indicadores para medir el rendimiento del algoritmo bajo varios ámbitos.

Tabla 4: Dirección de Cambio: Variable Prácticas de Trabajo

5.1.4 Integración de Procesos

La variable de integración define el grado de interacción entre los procesos dentro de un macro-proceso o entre diferentes macro-procesos.

Este proyecto de tesis busca re-diseñar el macro-proceso asociado a la cadena de valor de gestión del desarrollo de soluciones para la gestión académica, específicamente la forma en que se realizan las estimaciones de demanda de las asignaturas, por lo que se abordarán procesos asociados a la producción de información, gestión y planificación de modelos.

Esto significa que se crearán los nuevos procesos de producción, gestión y planificación para el nuevo algoritmo basándose en los procesos antiguos, para que el cambio sea lo más sutil posible. Por esta razón se deberán integrar los procesos mencionados para que ejecuten una comunicación entre ellos y que se genere una correcta planificación e instrucciones que deberán seguir las actividades de producción y también que se entreguen los mensajes de estado a las actividades correspondientes.

Integración de Procesos	Actual	Propuesto
Proceso aislado.	No.	El proyecto abarca los procesos de producción y sus interacciones.
Todos o la mayor parte de los procesos.	Sí.	Se abordan varios procesos del desarrollo de soluciones- Cadena de Valor
Dos o más macros que interactúan.	Sí.	El proyecto no abordará varias macros.

Tabla 5: Dirección de Cambio: Variable Integración de Procesos

5.1.5 Mantención del Estado

La mantención de estado existe para proveer todos los datos necesarios para ejecutar las prácticas de trabajo y comunicar las actividades y procesos.

La empresa cuenta con un sistema de bases de datos en donde se tiene un resguardo de los datos tanto ingresados, procesados y de resultados finales, así como también se tendrá la información de la historia de los resultados.

A este actual sistema de bases de datos se le realizará un cambio en el modelo de almacenamiento para que pueda soportar los datos tanto de entrada como los procesados. Se realizará también un almacenamiento de los modelos que se irán actualizando para cada cliente.

Mantención del Estado	Actual	Propuesto
Datos Propios	Sí.	Agregar al sistema actual nuevos modelos de base de datos que soporte los datos necesarios para el modelo.
Integración con datos de otros sistemas de la empresa.	Sí.	Se mantiene.
Integración con datos de sistemas de otras empresas	Sí.	Se mantiene.

Tabla 6: Dirección de Cambio Variable Mantención de Estado

5.1.6 Utilización de TI

Esta variable es el resultado de las opciones que se han tomado respecto de las variables anteriores.

Actualmente la empresa cuenta con una arquitectura tecnológica que contiene varias capas, esto está diseñado para que los componentes que conforman estas capas realicen tareas específicas y encapsuladas.

Dado esto la arquitectura tecnológica permite realizar los procesos de ejecución, mantención de los modelos y procesos de limpieza de datos de forma externa a la arquitectura. Luego realizar pruebas de rendimiento y finalmente anexarlos a la arquitectura.

Para lograr este objetivo se debe realizar un trabajo en conjunto con los administradores de sistemas para poder anexar los procesos a la arquitectura, generando llamados automáticos a éstos.

Utilización de TI	Actual	Propuesta
Lógica de Limpieza de datos.	Lógica Simple	Revisión de consistencias más complejas.

Tabla 7: Dirección de Cambio: Variable Utilización de TI

5.2 DISEÑO DETALLADO DE PROCESOS "TO BE"

5.2.1 Rediseño de Procesos

El proyecto de tesis se concentra en brindar una alternativa al proceso de producción de información de la demanda de los clientes, mediante la realización de modelos probabilísticos que permitan brindar una solución a clientes con mallas curriculares muy complejas. Actualmente a este tipo de clientes no se les pude ofrecer este producto y éste está teniendo una caída en las ventas debido a la necesidad de una actualización.

Es por esto que este proyecto busca rediseñar dentro de la arquitectura de macro procesos el desarrollo de soluciones integrales en donde se busca impactar específicamente el proceso de generación de información y por ende para modificar dicho proceso también se deberá crear un proceso dentro de las actividades de monitoreo y planificación de la generación y entrega. A continuación se detalla los cambios realizados dentro de los procesos:

5.2.1.1 Generación de Información

En este proceso se cambiará la lógica que se aplicará para generar las predicciones, lo que conlleva la creación de nuevas transformaciones para los datos y la aplicación de un nuevo modelo de predicción. En la Figura 18 se muestra los cambios realizados en el proceso de producción, en donde se eliminan las actividades de sensibilizar los datos, normalizar y recalibrar modelo. Estos cambios ayudarán a automatizar el proceso, disminuir los tiempos de entrega de la información y quitarán responsabilidad y carga a los analistas por lo que se habilitarán recursos humanos.

De la misma forma se cambiará el proceso de generación del piloto que se muestra en la Figura 19 eliminando las actividades de sensibilización de los datos, normalización y recalibración del modelo.

5.2.1.2 Monitoreo y planificación de la generación y la entrega

Dentro de este macro proceso el principal cambio será la mantención y construcción de los modelos que consiste en la creación de un proceso, que se muestra en la Figura 17, que se ejecutará a principios de cada semestre bajo la responsabilidad de un analista. Este proceso se encargará de construir los modelos de predicción, almacenarlos y calcular el rendimiento que obtiene el algoritmo basado en los datos que se conocen y utilizando los indicadores propuestos.

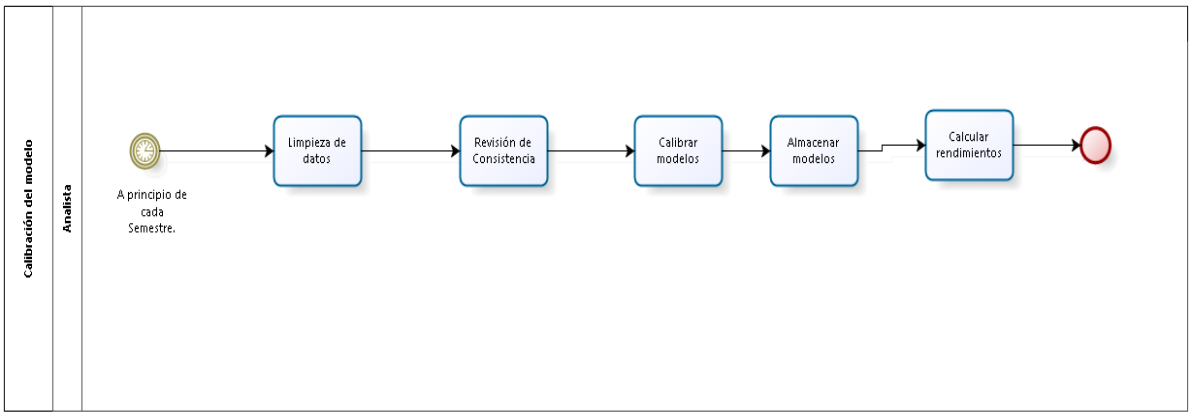

Figura 17: Proceso de Calibración de Modelos

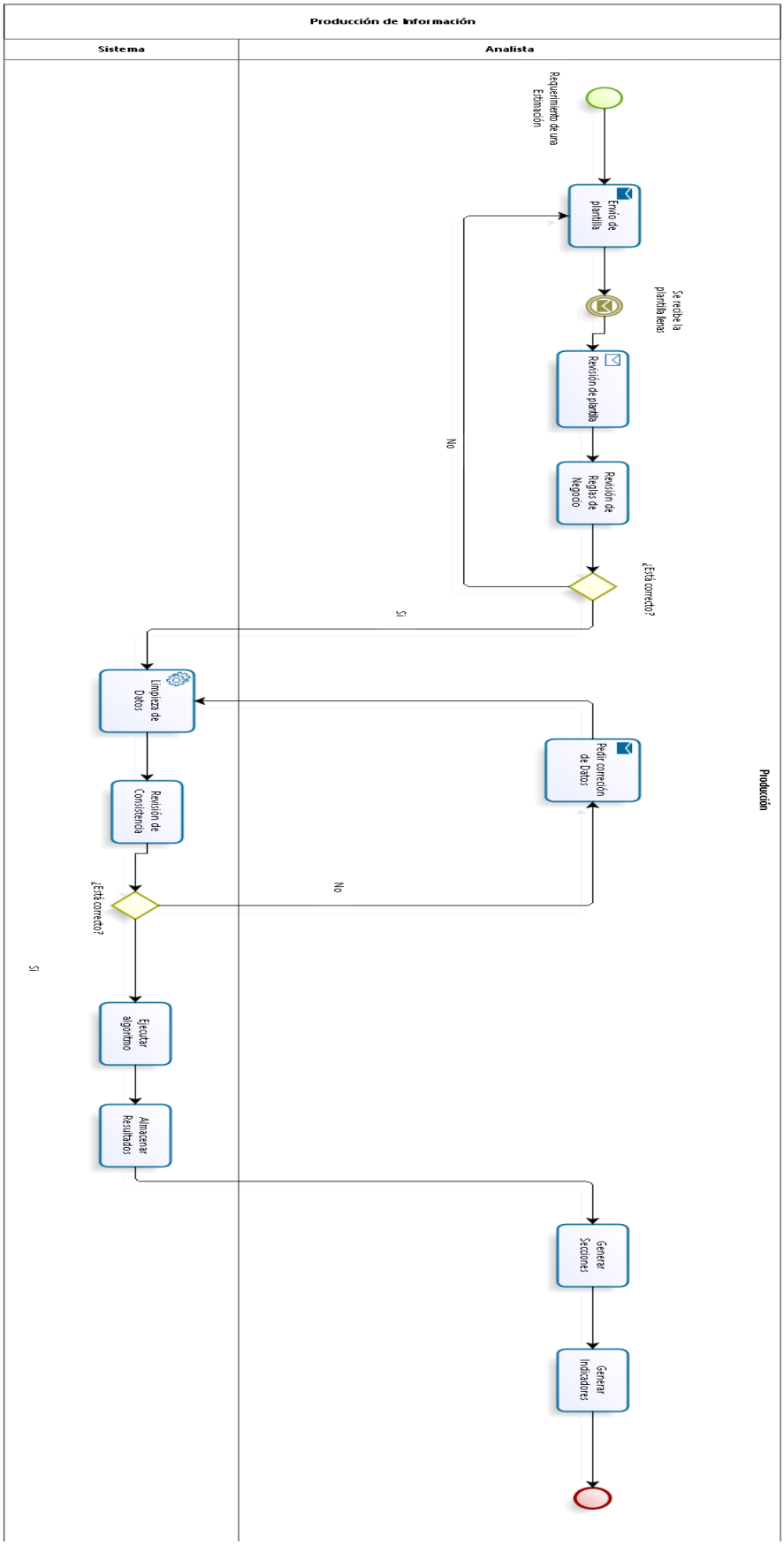

Figura 18: Rediseño del Proceso de Producción

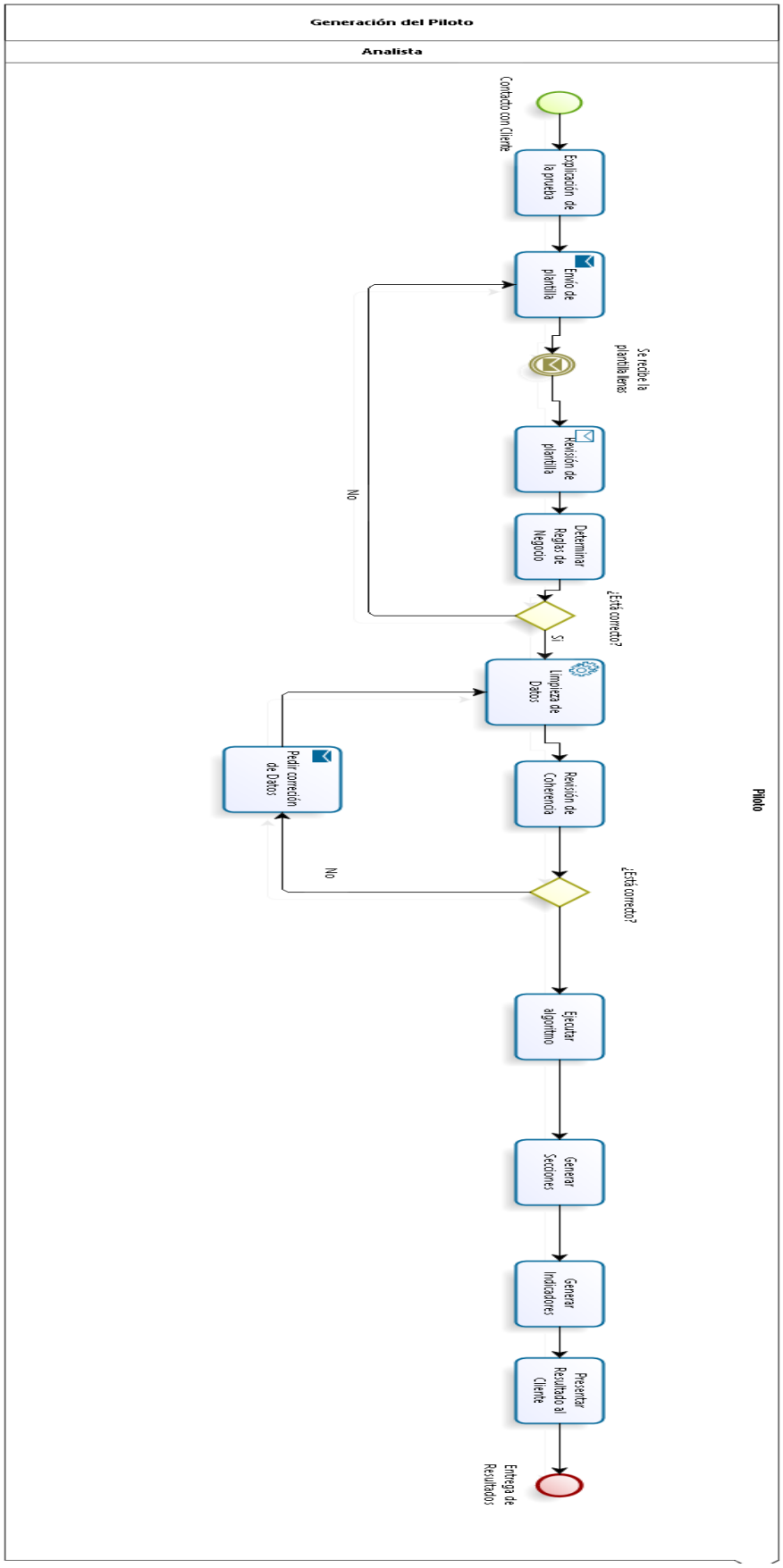

Figura 19: Rediseño del Proceso de Generación de Piloto

5.3 DISEÑO DE LÓGICA DE NEGOCIOS

En este capítulo se explican las lógicas utilizadas en cada una de las partes que se rediseñaron dentro de los procesos afectados.

5.3.1 Extracción, Transformación y Carga de datos

Previo a la ejecución o calibración de los modelos se realiza una serie de procesos de extracción, transformación y carga de datos en los sistemas de U-planner (ETL). En donde sus objetivos son realizar una limpieza de los datos que ingresarán al modelo, un análisis de consistencia de los datos, generar un reporte de errores y cargar los datos en una base de datos que posteriormente alimentará al algoritmo. Este proceso es de gran importancia para la realización de la calibración de los modelos, debido a que es común que las universidades entreguen datos incorrectos, incompletos o inexactos. Dada la cultura que tienen las universidades al momento de mantener sus datos, este problema es siempre un riesgo latente.

La limpieza consiste en eliminar registros duplicados y filtrar registros con campos obligatorios nulos. Luego se realiza el análisis de consistencia entre los layouts entregados y se genera un reporte de las inconsistencias y los registros duplicados que no pudieron ser cargados y finalmente se cargan los que pasaron los filtros.

5.3.2 Mantención y calibración de los modelos

Como se mencionó anteriormente para poder ejecutar un modelo, éste debe previamente ser construido. Para esto se ejecuta el proceso de calibración de modelos que se realiza una vez en cada semestre. La lógica al momento de crear los modelos consiste en crear perfiles de alumnos para cada asignatura y asociarles una probabilidad de tomarla, basándose en los ramos que ha aprobado, reprobado o esté cursando en el actual periodo y posteriormente obtener una esperanza de alumnos que inscribirá cada ramo.

El algoritmo comienza con una serie de transformaciones para crear una matriz de entrada con la que el modelo pueda obtener las probabilidades de inscribir la asignatura. Las transformaciones consisten en integrar los datos de la universidad tales como las inscripciones históricas de asignaturas, el estado histórico de alumnos, las mallas curriculares de los planes, las convalidaciones realizadas por los alumnos y las equivalencias de ramos que las universidades tengan consideradas.

Para crear la matriz de entrada se recorre un plan de estudio a la vez, por lo que toma a todos los alumnos que pertenezcan a dicho plan, las asignaturas de ese plan y asignaturas que hayan inscrito esos alumnos. Posteriormente realiza la siguiente división en las inscripciones:

- Inscripciones del periodo actual
- Inscripciones del periodo anterior
- Inscripciones de dos periodos anteriores
- Inscripciones de tres periodos hacia atrás.

Con lo que se genera una matriz que nos dice que asignaturas ha aprobado, reprobado, está cursando o no ha cursado cada uno de los alumnos de ese plan en el periodo anterior, en dos periodos anteriores, hace más de dos periodos anteriores y se coloca un uno o un cero si inscribió o no la asignatura objetivo (la que se desea predecir) en el periodo actual. En la Figura 20 se muestra una gráfica explicativa con la estructura de la matriz de entrada para construir los modelos:

$$M = \begin{bmatrix} & \begin{bmatrix} asign_1 & \cdots & asign_k & \cdots & asign_n \end{bmatrix} & \begin{bmatrix} asign_1 & \cdots & asign_k & \cdots & asign_n \end{bmatrix} & \begin{bmatrix} asign_1 & \cdots & asign_k & \cdots & asign_n \end{bmatrix} & \begin{bmatrix} asign_{obj} \end{bmatrix} \\ alum_1 & \begin{bmatrix} A & \cdots & 0 & \cdots & 0 \end{bmatrix} & \begin{bmatrix} A & \cdots & A & \cdots & A \end{bmatrix} & \begin{bmatrix} A & \cdots & 0 & \cdots & 0 \end{bmatrix} & \begin{bmatrix} 1 \end{bmatrix} \\ \vdots & \begin{bmatrix} \vdots & & \ddots & & \vdots \end{bmatrix} & \begin{bmatrix} \vdots & & \ddots & & \vdots \end{bmatrix} & \begin{bmatrix} \vdots & & \ddots & & \vdots \end{bmatrix} & \begin{bmatrix} \vdots \end{bmatrix} \\ alum_k & \begin{bmatrix} A & \cdots & A & \cdots & C \end{bmatrix} & \begin{bmatrix} R & \cdots & 0 & \cdots & 0 \end{bmatrix} & \begin{bmatrix} A & \cdots & A & \cdots & C \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} \\ \vdots & \begin{bmatrix} \vdots & & \ddots & & \vdots \end{bmatrix} & \begin{bmatrix} \vdots & & \ddots & & \vdots \end{bmatrix} & \begin{bmatrix} \vdots & & \ddots & & \vdots \end{bmatrix} & \begin{bmatrix} \vdots \end{bmatrix} \\ alum_n & \begin{bmatrix} A & \cdots & A & \cdots & A \end{bmatrix} & \begin{bmatrix} A & \cdots & A & \cdots & R \end{bmatrix} & \begin{bmatrix} A & \cdots & A & \cdots & A \end{bmatrix} & \begin{bmatrix} 1 \end{bmatrix} \end{bmatrix}$$

Estudiantes — Periodo Anterior — Dos periodos Antes — Más de dos periodos Antes — Periodo Actual

Figura 20: Matriz de entrada al modelo

El siguiente paso es la construcción de un modelo para cada asignatura, para esto se escogió construir modelos de árboles de decisión CHAID[4] el cual utiliza la prueba de chi-cuadrado al momento de construir el árbol la cual elige una de las asignatura como la más relevante para determinar si un alumno inscribe la asignatura objetivo, luego la segunda más relevante y así sucesivamente hasta que la siguiente asignatura no es lo suficientemente relevante para la predicción.

En la Figura 21 se muestra un ejemplo de construcción real de un árbol. En el árbol los nodos representan cursos y periodos en el que puede haber tomado un estudiante un curso. Las ramas de los árboles representan la decisión que tomará el modelo para clasificar a los alumnos, en donde estas pueden ser aprobó, reprobó o no ha tomado el curso que se representan por A, R y 0 respectivamente. Las hojas o nodos finales representan la probabilidad de tomar el curso, para el caso en que haya un "1" esto significa que hay altas probabilidades de tomar el curso objetivo, en caso contrario si aparece un "0" esto significa que hay bajas probabilidades que se tome este curso.

[4] Detección Automática de Interacción Chi-Cuadrado.

Para este ejemplo el curso a estimar es el 120006. Aquí se puede ver que el árbol acierta en la consideración como prerrequisito al curso 120001 para el periodo anterior y se puede ver que si se reprueba este curso las probabilidades de no tomar el curso 120006 son muy altas, también se observa que si se aprueba este curso las probabilidades de inscribir el curso 120006 son muy altas. La intensidad de los colores simboliza que existen más casos para esa alternativa. Por último existen casos en donde se reprueba la asignatura 120006 en el semestre anterior y el modelo entrega una probabilidad alta de inscripción el siguiente semestre.

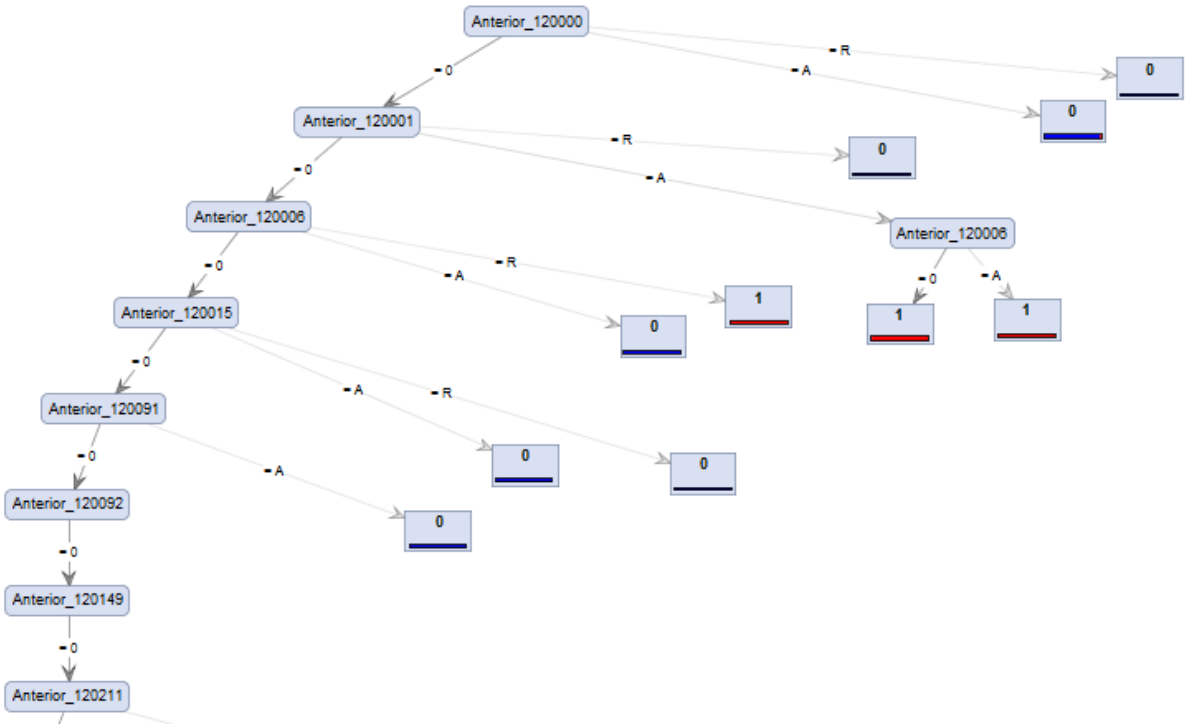

Figura 21: Ejemplo de árbol de decisión CHAID

Como la construcción de los árboles puede ser muy distinta entre asignaturas, se realiza una optimización automática del parámetro de los árboles llamado la mínima cantidad de registros que debe tener un nodo para ser separado. Este parámetro, como su nombre lo indica, determina cuántos registros debe tener un nodo para poder ser separado en dos nodos hijos, lo que ayuda a tener modelos ajustados de acuerdo a la cantidad de registros de éxito y fracaso que la asignatura tenga.

Este proceso se repite para todos los cursos y para todos los planes, posteriormente se almacenan los modelos para poder ser utilizados. Al finalizar el algoritmo realiza una prueba de rendimiento en donde estima la demanda para un periodo conocido utilizando los modelos, para luego calcular los indicadores de rendimiento propuestos.

5.3.3 Lógica de Ejecución de los Modelos

Cuando se tienen los modelos entrenados para todos los cursos que se quiere estimar se puede dar comienzo al proceso de producción de información. El proceso comienza cuando el cliente presiona la opción de "Ejecutar Modelo", con lo que el sistema genera un ambiente en donde los datos limpiados y en el formato necesario son alojados para que el algoritmo inicie su ejecución.

Al igual que en la construcción de los modelos se realizan las transformaciones para generar la matriz de entrada mostrada anteriormente, sin las inscripciones del periodo actual, para cada una de las asignaturas. Luego se utiliza el modelo entrenado de árboles de decisión CHAID con lo que se obtiene una probabilidad para cada alumno de tomar la asignatura. Finalmente se calcula la esperanza de alumnos que tomarán el ramo el próximo semestre y con esto se obtiene la estimación.

El algoritmo repite estos pasos tantas veces como asignaturas y planes se hayan ingresado para estimar para ese semestre. Para lidiar con las estimaciones que están fuera de lo común se realiza una etapa que consiste en realizar una corrección de las estimaciones en base a la tendencia de la demanda histórica, la lógica que se utiliza es acotar las predicciones que estén muy lejanas a la demanda histórica de semestres equivalentes. Luego se guardan los resultados y se dejan en el ambiente generado para que el sistema los pueda tomar y desplegar al cliente.

CAPÍTULO 6: PROPUESTA DE APOYO TECNOLÓGICO

En este capítulo se explica el funcionamiento del apoyo tecnológico del prototipo que se implementó. En primer lugar se describe la estructura tecnológica que posee la empresa, en donde alojará la herramienta propuesta y posteriormente se utiliza la metodología UML (*Unified Modeling Language*) para describir las interacciones entre el sistema y actores y las interacciones internas de la aplicación computacional.

6.1 ESPECIFICACIÓN DE REQUERIMIENTOS

La solución diseñada pretende brindar una nueva alterativa de oferta para el producto de U-Forecast con el fin de aumentar la cantidad de clientes en los que se puede aplicar la estimación de demanda. Para esto se realizó un nuevo algoritmo con el cual se busca abarcar más perfiles de clientes utilizando menos datos obligatorios y que pueda conectarse al actual sistema que se utiliza en la empresa.

6.1.1 Requisitos Funcionales

Los requisitos funcionales son:

- Estimar la cantidad de alumnos de los cursos que se especifique para un periodo futuro basándose en la información de periodos anteriores.
- Al sistema se le debe entregar los datos de la historia de los alumnos, las inscripciones históricas y las mallas de las carreras.
- El sistema debe poder configurar los parámetros del periodo que se quiere estimar.
- Configurar los datos de *inputs* que se quieren utilizar para una ejecución.
- Se deben entregar los resultados de las estimaciones de cada una de las ejecuciones de la aplicación.
- El cliente debe poder cambiar los resultados obtenidos si lo desea.
- Debe presentar un menú en donde se pueda elegir la acción que se quiere realizar.
- El sistema debe ser capaz de guardar los resultados y modelos generados en bases de datos.

6.1.2 Requisitos No Funcionales

Los requisitos no funcionales son:

- Debe existir un cuadro explicativo de cómo operar las distintas opciones.
- Los modelos y resultados deben poder mantenerse ante una caída de los servidores.
- Las consultas de resultados y cambios de parámetros deben tomar un tiempo del orden de minutos en su ejecución.
- La ejecución de los modelos debe tomar un tiempo mucho menor al de la calibración de éstos.

6.2 ARQUITECTURA TECNOLÓGICA

La arquitectura tecnológica de U-planner está basada en un modelo de cinco capas, el que se muestra en la Figura 22, en donde dos de ellas se encargan de la gestión y flujos de las vistas, otra que se encarga de gestionar y ejecutar los algoritmos que representa la capa de lógica y las últimas dos capas que se encargan de la gestión y almacenamiento de los datos en las bases de datos.

La capa de datos llamada capa de sistema de gestión de BD (SGBD) almacena la información procesada y no procesada, ésta se comunica a través de una capa de framework que ejecuta servicios de extracción e inserción de datos.

La capa de lógica es la que se encarga de gestionar las peticiones que se van generando por parte de los clientes y de los funcionarios, que quieren realizar consultas a las bases de datos o ejecutar simulaciones, por lo que debe priorizar las peticiones. Esta capa es la que se encarga de llamar a los proceso de limpieza de datos, ejecutar los llamados a los distintos algoritmos y brindar ambientes de trabajo para realizar las simulaciones aportando con los datos de *inputs* limpios, modelos calibrados y parámetros. Luego ésta encapsula los resultados y reportes generados por el algoritmo y se le entrega a las capas de datos para que sean almacenadas en la bases de datos correspondientes.

La capa de servidores de presentación se encarga ejecutar el despliegue de las páginas que mostrarán la información al usuario, esto se realiza a través de las técnicas y lenguajes XHTML, AJAX y JSF y se complementan con softwares como Tableau para ayudar a mostrar análisis de datos. Al momento en que el usuario ejecuta una petición de consulta de resultados esta capa ejecuta los servicios correspondientes para que el orquestador consiga los datos y se pueda comenzar con el despliegue.

La capa de servidores de presentación se comunica con la capa de presentación por el lado del usuario, que cuenta con una aplicación Java creada para desplegar las páginas, también cuenta con aplicaciones para mostrar información procesada en caso de consultorías en donde los clientes no necesariamente estén integrados con el sistema.

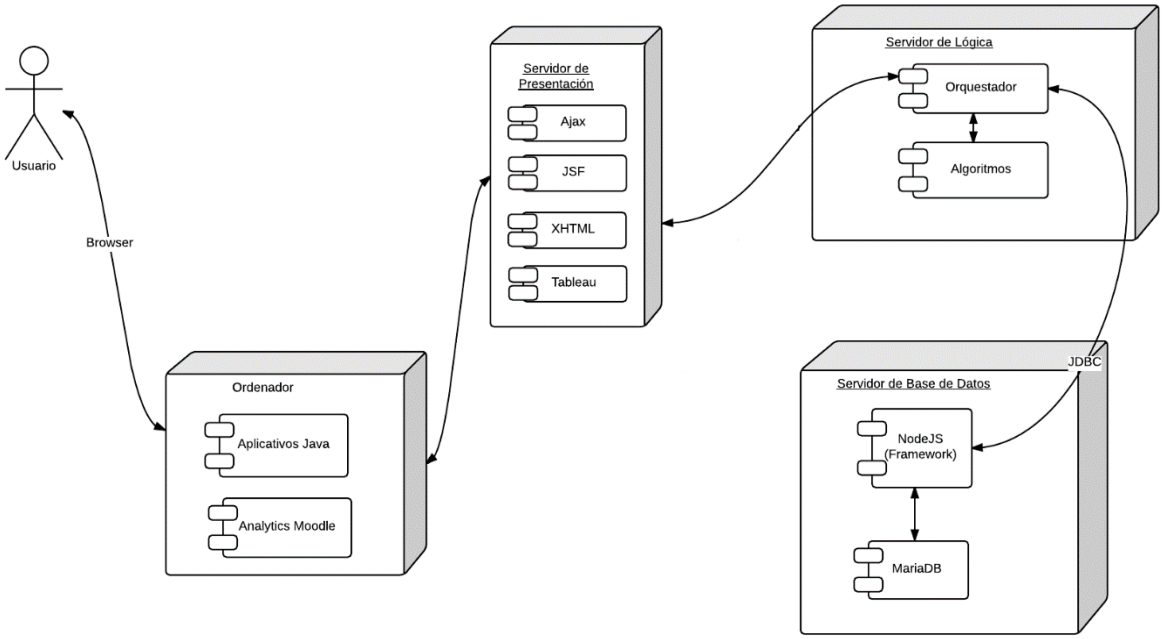

Figura 22: Arquitectura Tecnológica

El rediseño contempla la creación de un algoritmo nuevo que se alojará en el servidor de la lógica para ser llamado por el orquestador. También se crean los procesos de transformación y carga de datos que alimentan al algoritmo.

6.3 DISEÑO DE LA APLICACIÓN

Lo visto hasta aquí es el modelo realizado de la arquitectura de procesos de la empresa. Sustentado en el modelo de negocio se ha trabajado en una lógica nueva para el proceso de negocio de la empresa. A continuación se detalla la construcción de la solución tecnológica que se derivó del modelamiento del proceso.

6.3.1 Casos de Uso

Los diagramas de casos de uso describen las principales interacciones que existen entre las personas y el sistema. Se muestra en la Figura 23 los casos de uso de los procesos que requieren un apoyo computacional. Se modela el apoyo computacional para el diagrama de actividades correspondiente a “Ejecutar predicción de demanda” y “Calibración de modelos”. Para los casos de uso existirán dos actores que pueden interactuar con el sistema: el usuario y un analista.

- Ejecutar predicción de demanda

  Propósito: Permite al usuario aplicar el modelo entrenado en los datos que tiene cargados en el sistema.

- Calibración de modelos

  Propósito: Permite al analista actualizar los modelos y obtener una predicción de prueba.

Figura 23: Casos de Uso

6.3.2 Diagramas de Secuencia

El diagrama de secuencia de sistema es un tipo de diagrama de comportamiento, que al igual que el de casos de uso es parte del lenguaje UML, en donde muestra las interacciones entre los actores y el sistema y el orden de ocurrencia de éstas.

Para el diagrama de secuencia se asume que el cliente ya se identificó en la aplicación y se encuentra en la ventana de estimación de demanda.

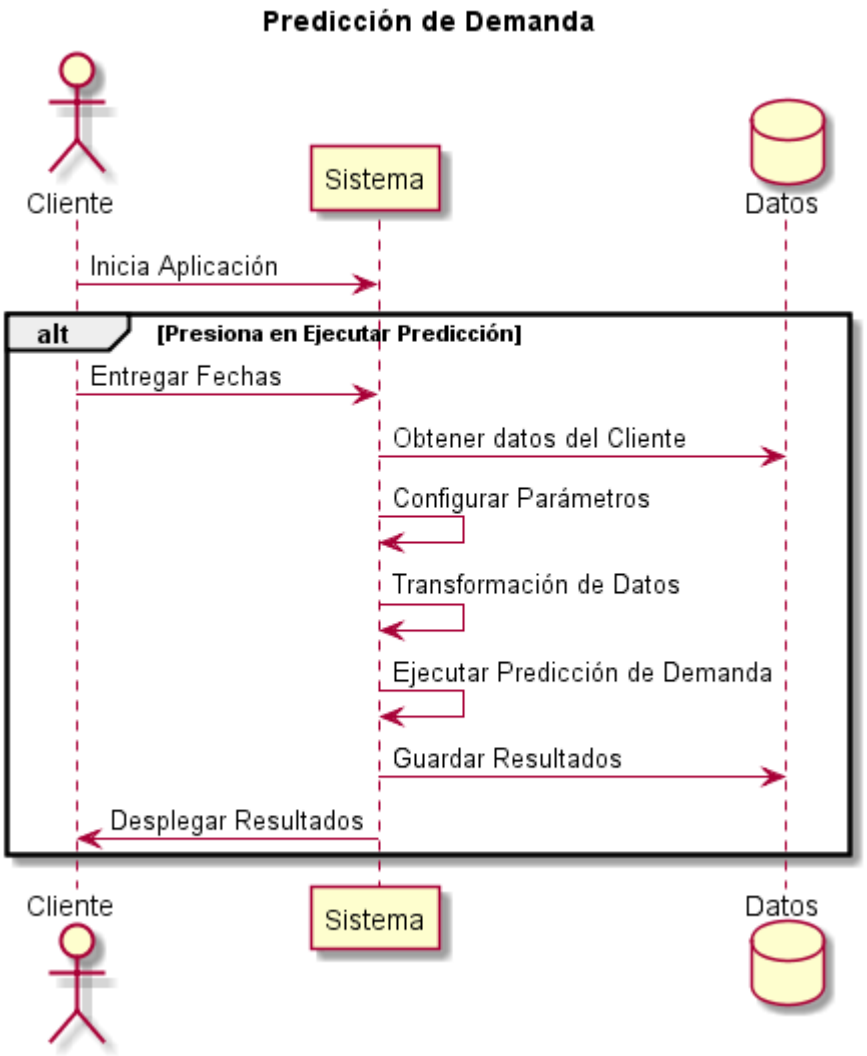

Figura 24: Diagrama de Secuencia por parte del Cliente

Como se muestra en la Figura 24 luego de iniciar la aplicación para realizar la ejecución de predicciones el sistema pide los datos necesarios, que deben ser previamente cargados, a la capa de datos. Configura los parámetros específicos de la universidad y de la simulación, luego realiza la transformación de los datos para poder entrégaselos al algoritmo y se ejecuta la estimación de la demanda. Se guardan los resultados obtenidos en las bases de datos y se muestran al cliente.

A continuación se muestra en la Figura 25 el diagrama de secuencia para la mantención de los modelos.

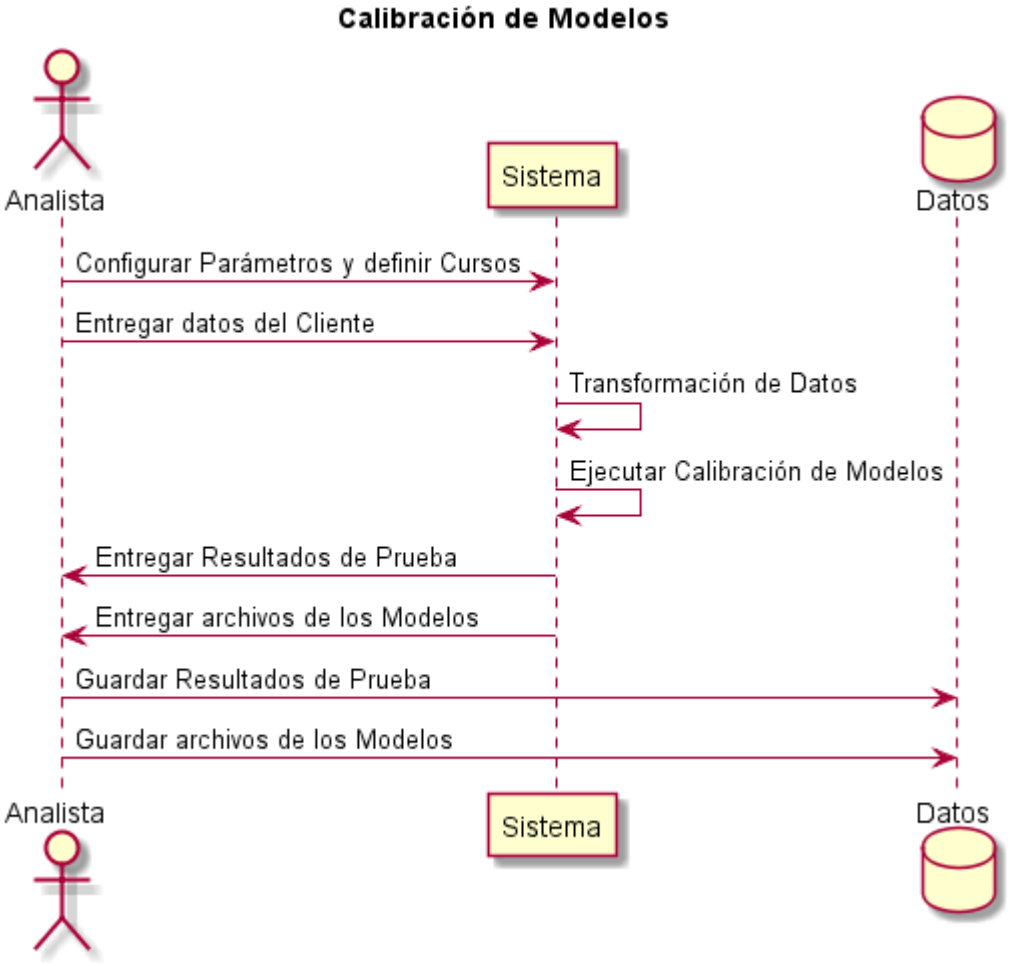

Figura 25: Diagrama de Secuencia por parte del Analista.

El analista al querer o necesitar actualizar los modelos que se utilizan para realizar las estimaciones, ejecutará el proceso de calibración de los modelos. El analista comienza con la configuración de los parámetros de la calibración, luego debe definir para qué cursos se realizará la calibración y entregar los datos del cliente al sistema. El sistema realiza la trasformación de estos datos y posteriormente completar la actualización de los modelos.

El proceso continúa con la entrega de resultados de una predicción utilizando los modelos calibrados para que el analista pueda realizar una prueba del rendimiento de los modelos. Finalmente el analista debe guardar los resultados y los modelos.

Para el proceso de calibración de modelos el alcance de este proyecto de tesis implicó que no se pudiera dejar automatizado, y por lo tanto el analista deberá interactuar manualmente con el sistema para obtener los datos y luego para almacenarlos. Es decir, con el único componente que se relacionará el analista será con el algoritmo. Esto deja para un trabajo futuro la automatización de este proceso, para que el analista pueda tener una vista que despliegue los datos y un controlador que ejecute el algoritmo y almacene los datos de forma automática.

A continuación se presenta la Figura 26 en donde se muestra el diagrama de despliegue anteriormente explicado en el cual se separa el sistema en las distintas capas y se muestra las tareas de cada una.

6.3.2.1 Diagrama de Secuencia Extendido Caso de Uso Ejecutar Modelo.

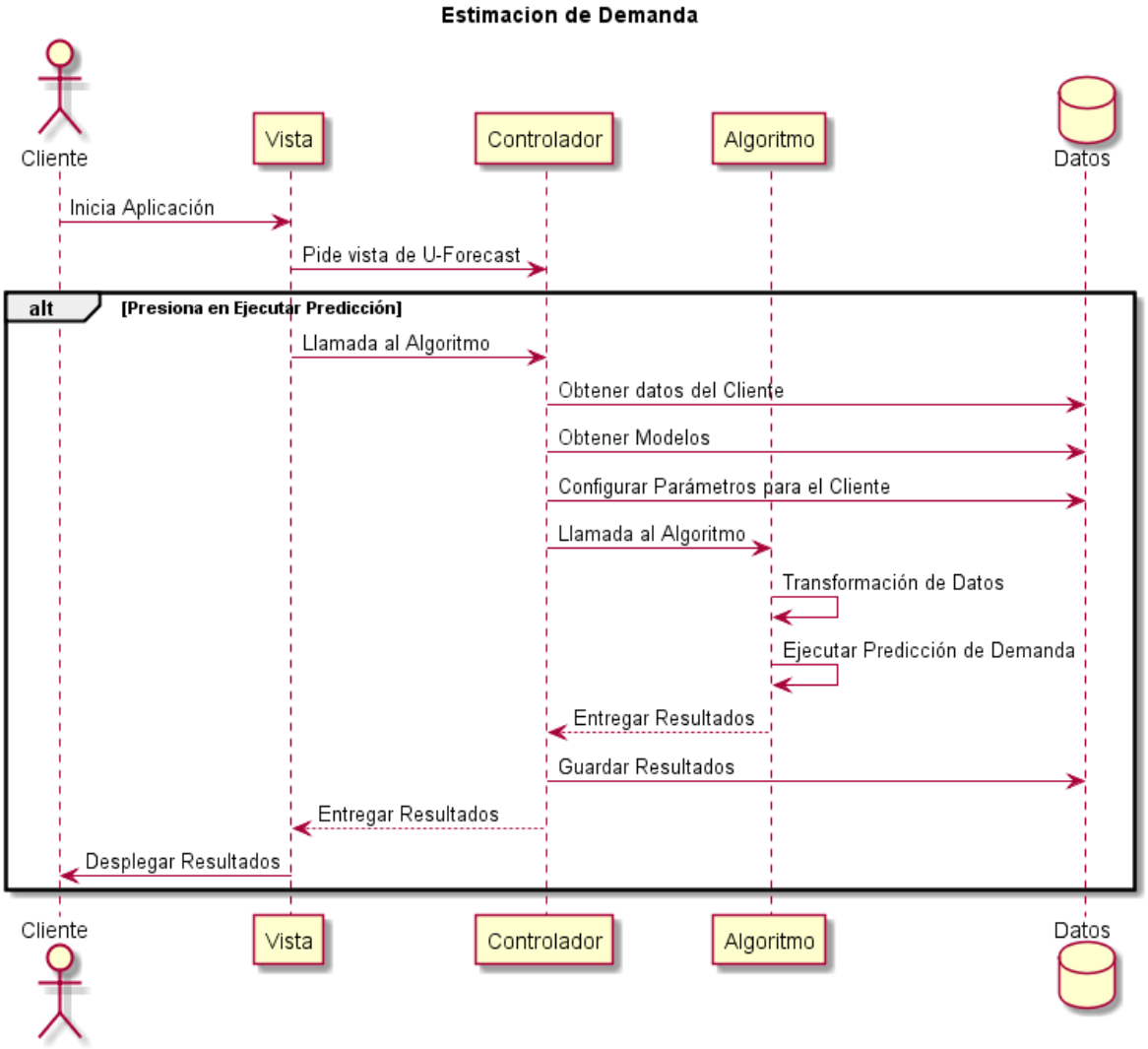

Figura 26: Diagrama de Secuencia Extendido: Ejecutar predicción

6.4 PROTOTIPO FUNCIONAL DESARROLLADO

6.4.1 Métricas Propuestas

Para poder medir el rendimiento que tuvo el prototipo realizado se analizó cuáles eran los mejores indicadores que reflejarían el comportamiento del algoritmo. Tras el análisis se optó por que la gran mayoría de los indicadores se midieran en base al error entre secciones y no en base al error entre alumnos. Estos indicadores resultan ser mucho más significativos para el cliente y los analistas, debido a que la información final del producto U-Forecast es la estimación de secciones. Mostrar los indicadores a nivel de secciones permite que los errores sean más fáciles de comprender, además de que existe una relación directa asociada a costos de sobrestimar y subestimar secciones.

Para los cálculos de indicadores se propuso una regla para la creación de secciones. Esta regla busca crear secciones en donde a lo más una sección pueda tener el máximo de holgura para esa asignatura. La lógica de esta regla se ve reflejada en la relación (*4*).

Como se mencionó en capítulos anteriores actualmente, existe un algoritmo que calcula el número de secciones con una lógica compleja, por lo que para poder tener resultados para el cliente se depende de la ejecución de este algoritmo. Con esta nueva lógica propuesta en este trabajo, el algoritmo de predicción de demanda puedrá ser independiente del algoritmo de creación de secciones, con el fin de entregar un número rápido y certero de las secciones mínimas que se necesitan para albergar la demanda pronosticada, dejando la utilización del algoritmo de secciones solo para cuando sea estrictamente necesario.

Por lo que se definen los siguientes conceptos:

$A = \{a_1, a_2, \dots, a_N\}$, a_j: asignatura j, con N el número de total de asignaturas.

$I = \{ i_1, i_2, \dots, i_N\}$, i_j: inscripciones en la asignatura a_j.

$S = \{ s_1, s_2, \dots, s_N\}$, s_j: secciones en asignatura a_j.

$C = \{ c_1, c_2, \dots, c_N\}$, c_j: cupo académico de las secciones de la asignatura a_j.

$H = \{ h_1, h_2, \dots, h_N\}$, h_j: holgura de las secciones de la asignatura a_j.

$P(I) = \{ p(i_1), p(i_2), \dots, p(i_N)\}$, $p(i_j)$: predicción de inscripciones para asignatura a_j.

$P(S) = \{ p(s_1), p(s_2), \dots, p(s_N)\}$, $p(s_j)$: predicción de secciones para asignatura a_j.

El cupo académico y la holgura son entregados por el cliente junto con los datos para la predicción de demanda, en caso que éste no los entregue se definen un cupo académico y holgura por defecto.

La regla de construcción de secciones relaciona las variables de la siguiente manera:

$$p(s_j) = \left\lceil \frac{p(i_j) - 1}{c_j + h_j - 1} \right\rceil \qquad (4)$$

Por ejemplo, sea una asignatura a_j con una predicción de demanda $p(i_j) = 49$, cupo $c_j = 20$ y holgura $h_j = 5$ por lo que la predicción de secciones será $p(s_j) = 2$, teniendo una de 25 y otra de 24. En cambio otra asignatura a_{j+1} con las mismas características, pero con demanda $p(i_{j+1}) = 50$ tendrá una predicción se secciones de $p(s_{j+1}) = 3$ en donde las dos primeras secciones tendrán 17 alumnos y la última 16. Con esto se cumple la regla que una asignatura puede tener solo una sección con un número de alumnos igual al cupo académico más la holgura.

Si se tiene las siguientes asignaturas de ejemplo:

	Cupo c_j	Holgura h_j
Asignatura 1	20	5
Asignatura 2	10	4
Asignatura 3	25	15

Tabla 8: Ejemplo de asignaturas ficticias

En la Figura 27 se muestra cómo varía el número en la creación de secciones para las asignaturas ficticias de la Tabla 8 a medida que aumenta el número de inscripciones pronosticadas. En la Figura 27 se pueden ver los puntos de cambio del número de secciones, por ejemplo, la asignatura 1 si tiene una predicción de demanda de 73 alumnos se le pronosticarán tres secciones, en cambio si tiene una predicción de demanda de 74 alumnos se le pronosticarán cuatro secciones. Lo anteriormente descrito permite que se generen secciones estrictamente necesarias sin abusar de la holgura en cada sección.

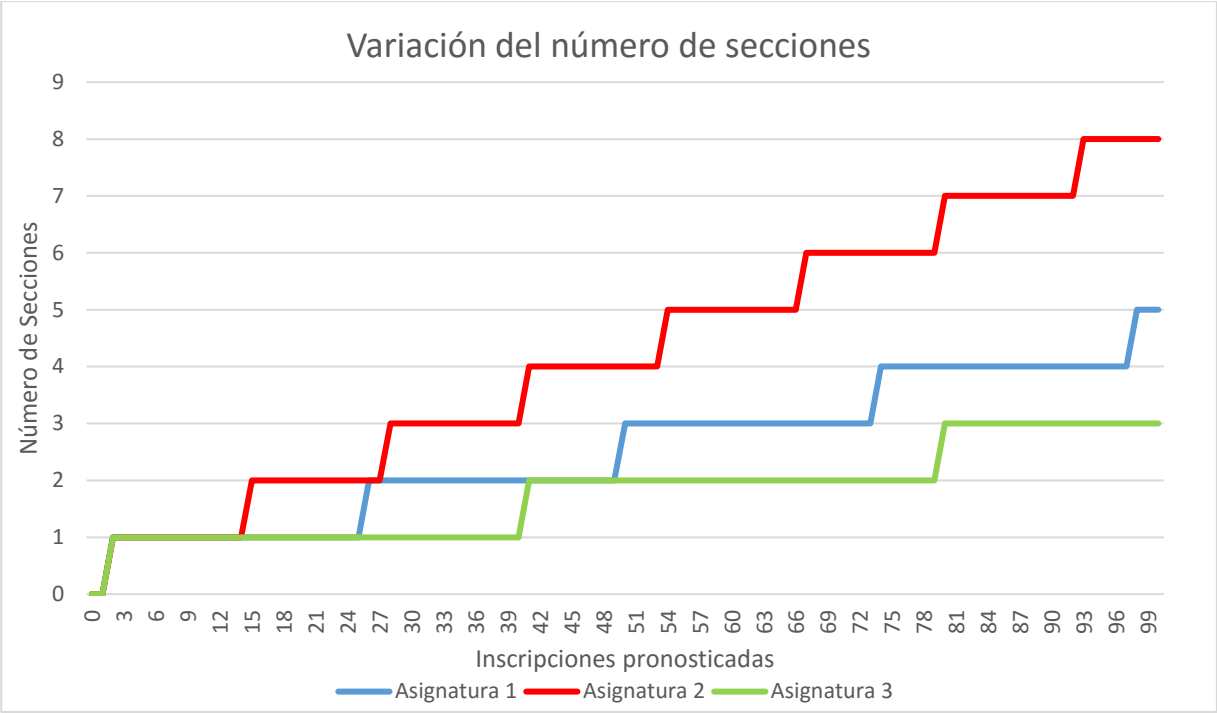

Figura 27: Variación del número de secciones

6.4.1.1 Métricas e Indicadores

Dado lo anterior se definen las siguientes variables:

- N , número de asignaturas
- $N_0 = \sum_1^N \delta(s_j = p(s_j))$, número de asignaturas con predicción de sección igual a lo real.
- $N_1 = \sum_1^N \delta(|s_j - p(s_j)| = 1)$, número de asignaturas con una diferencia en una sección de la predicción y lo real.
- $N_{2+} = \sum_1^N \delta(|s_j - p(s_j)| \geq 2)$, número de asignaturas con una diferencia de dos o más secciones de la predicción y lo real.

En donde δ es la función Delta de Kronecker que se define:

$$\delta(x) = \begin{cases} 0 & x \text{ es falso} \\ 1 & x \text{ es verdadero} \end{cases}$$

Por lo tanto, se tiene que $N_{2+} + N_1 + N_0 = N$.

Los indicadores clave son:

1) Porcentaje de asignaturas con secciones correctas (SC): Este indicador representa el porcentaje de asignaturas que obtuvo la misma sección en lo predicho y en lo real.

$$SC = \frac{\sum_1^N \delta(s_j = p(s_j))}{N} = \frac{N_0}{N} \qquad (5)$$

2) Porcentaje de asignaturas con secciones erróneas (SE): indicador que representa el porcentaje de asignatura que obtuvieron diferencia con lo predicho. Se dividió este indicador en los siguientes dos indicadores para reflejar de mejor manera en dónde se generan los errores:

2.1) Porcentaje de asignaturas con una sección errónea (SE_1):

$$SE_1 = \frac{\sum_1^N \delta(s_j - p(s_j) = 1)}{N} = \frac{N_1}{N} \qquad (6)$$

2.2) Porcentaje de asignaturas con dos o más sección errónea (SE_2):

$$SE_{2+} = \frac{\sum_1^N \delta(s_j - p(s_j) \geq 2)}{N} = \frac{N_{2+}}{N} \qquad (7)$$

Por lo tanto esto implica que $SC + SE_1 + SE_{2+} = 1$.

3) Número de secciones sub estimadas (SuE): Refleja el número de secciones que se sub estiman en la predicción.

$$SuE = \sum_1^N \delta(p(s_j) < s_j) \qquad (8)$$

4) Número de secciones sobre estimadas (SoE): Refleja el número de secciones que se sobre estiman del total de secciones.

$$\mathrm{SoE} = \sum_{1}^{N} \delta(p(s_j) > s_j) \qquad (9)$$

Los indicadores 3 y 4 permiten identificar cuántas secciones hubo de error en la simulación, estos indicadores son relevantes debido a que existe un costo directo asociado a estos errores, que la misma universidad conoce.

5) Matriz de Cursos dictados y no dictados: Matriz que muestra la predicción de los cursos que se dictarían (cursos con demanda mayor a cero) y no se dictarían (cursos con demanda igual a cero) versus los cursos que en la realidad se dictaron y no se dictaron. Esto es importante debido a que tener errores en este ámbito representa un costo alto para la universidad. Aquí no se considera el número de inscripciones, sino solo que se dicte la asignatura.

Se muestra en la Tabla 9 el esquema:

		Reales		
		Dictados (+R)	No dictados (-R)	
Predichos	Dictados (+P)	A	B (Error Tipo I)	Precision
	No Dictados (-P)	C (Error Tipo II)	D	NPV
		Recall	Specificity	Accuracy

Tabla 9: Matriz de cursos dictados y no dictados

Considerando el anterior esquema se propuso los indicadores clásicos de la evaluación binaria [16] para identificar el comportamiento de modelo en este ámbito.

5.1) Accuracy: mide la fracción de todas las instancias que han sido categorizadas correctamente. Es la tasa del número de clasificaciones correctas con el número correcto e incorrecto de clasificaciones.

$$\mathrm{Accuracy} = \frac{A + D}{N} \qquad (10)$$

5.2) Precisión o Confianza: Es la proporción de los casos predichos positivos que son correctamente positivos reales. Siendo una medida de la exactitud de positivos predichos en contraste con la tasa de descubrimiento de positivos reales.

$$\text{Precisión} = \frac{A}{A + B} \qquad (11)$$

5.3) Sensibilidad o Recall: Es la proporción de los casos Positivos Predichos que son correctamente Positivos Reales. Este indicador refleja cuantos casos relevantes toma la regla de los dictados (+P).

$$\text{Recall} = \frac{A}{A + C} \qquad (12)$$

5.4) Specificity o Inverse Recall: Es la proporción de los casos Negativos Predichos que son correctamente Negativos Reales. Este indicador refleja cuantos casos relevantes toma la regla de los no dictados (-P).

$$\text{Specificity} = \frac{D}{B + D} \qquad (13)$$

5.5) Valor Predictivo Negativo o Inverse Precision (NPV): Es la proporción de los casos Negativos Predichos que son correctamente Negativos Reales. Este indicador refleja cuantos casos relevantes toma la regla de los no dictados (-P).

$$\text{NPV} = \frac{D}{C + D} \qquad (14)$$

6) Razón entre de inscripciones totales e inscripciones reales (RI): Las inscripciones totales que se predicen que habrá en un próximo periodo.

$$RI = \frac{\sum_1^N p(i_j)}{\sum_1^N i_j} \qquad (15)$$

6.4.1.2 Gráficos

En esta parte se mostrarán los gráficos propuestos para visualizar el comportamiento de las predicciones del algoritmo. Para esto se utiliza un set de datos ficticios que ejemplifican las funciones de los gráficos, que consiste en 20 asignaturas con distinta características mostradas en la Tabla 10 .

Asignatura	Inscripciones reales	Inscripciones pronosticadas	Cupo	Holgura	Secciones reales	Secciones pronosticadas
a1	0	0	5	5	0	0
a2	0	40	5	5	0	5
a3	5	4	5	5	1	1
a4	5	15	5	5	1	2
a5	10	15	5	5	1	2
a6	10	20	5	5	1	3
a7	30	36	10	10	2	2
a8	50	0	10	10	3	0
a9	50	45	10	15	3	2
a10	50	65	10	15	3	3
a11	55	23	15	5	3	2
a12	60	40	20	5	3	2
a13	80	88	30	10	3	3
a14	100	100	50	10	2	2
a15	100	40	50	10	2	1
a16	100	250	50	10	2	5
a17	100	120	50	10	2	3
a18	100	130	50	10	2	3
a19	200	30	100	20	2	1

a20	200	239	100	20	2	2

Tabla 10: Tabla de asignaturas de ejemplo

1) Histograma de errores en secciones: Muestra la frecuencia con que las asignaturas tienen distintos niveles de error en las secciones. Mediante este gráfico se puede vislumbrar la distribución de los errores. En el ejemplo de la Figura 28 se muestra el resultado para las 20 asignaturas en donde se puede ver que 7 asignaturas fueron correctamente estimadas, mientras que 5 asignaturas están sobre estimadas por solo una sección y 4 asignaturas están sub estimadas por solo una sección.

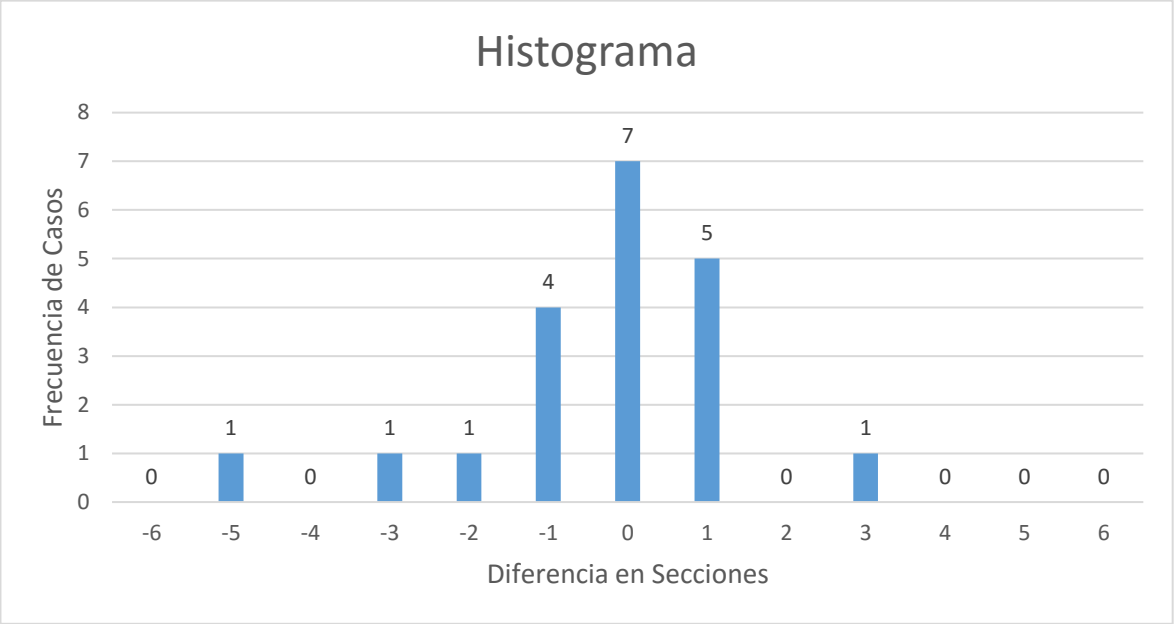

Figura 28: Histograma de ejemplo

2) Gráfico de error de alumnos vs Tamaño de Asignatura: este gráfico muestra el error en el eje de las ordenadas, en cantidad de alumnos, de cada estimación y en el eje de las abscisas el tamaño que tiene esa asignaturas, los colores representan como el error en la predicción de alumnos se refleja en el error de la predicción de secciones. Por ejemplo en la Figura 29 la asignatura a20 tiene un error de 39 alumnos, pero este error no refleja ningún error en la generación de secciones. Por otro lado la asignatura a19 tiene un error de 170 alumnos lo que se refleja en solo un error en las secciones, mientras que la asignatura a16 tiene un error de 150 alumnos lo que genera una diferencia de tres secciones con lo pronosticado.

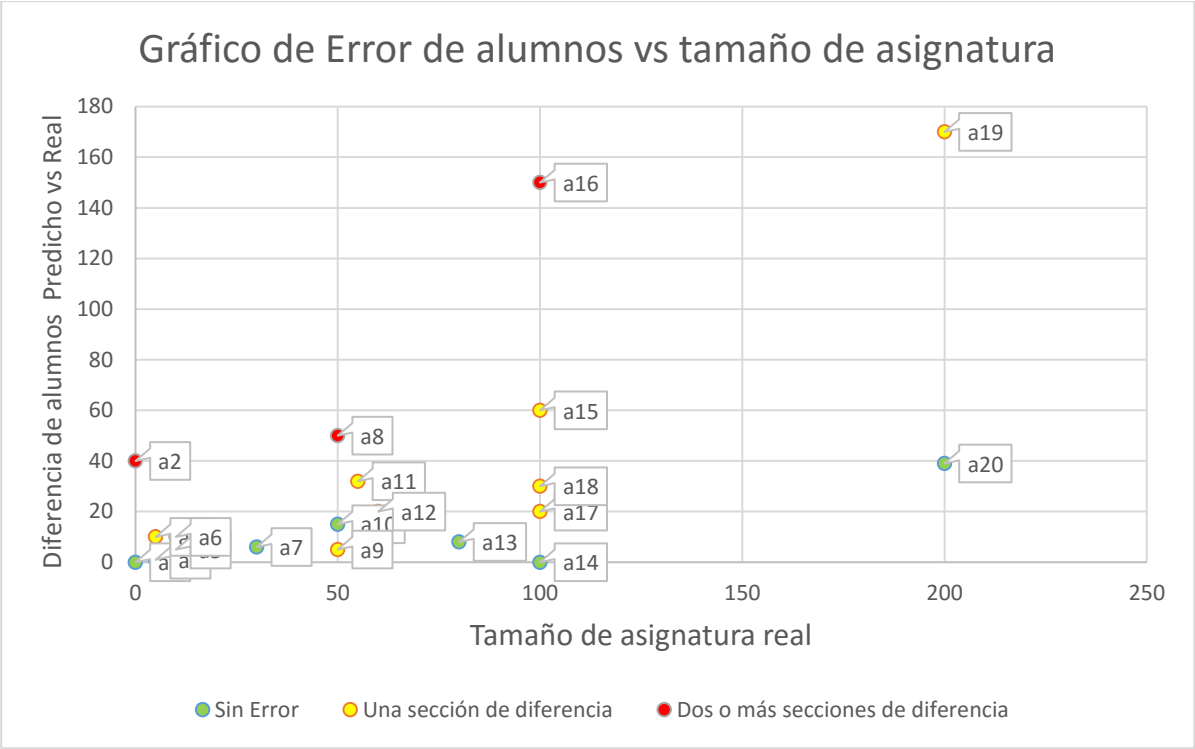

Figura 29: Gráfico Error de alumnos vs tamaño, ejemplo

6.4.2 Lógica del Prototipo Funcional

Para poder validar la solución tecnológica se realizó la construcción de un prototipo funcional que diera una alternativa a la estimación de demanda actual en la empresa. Se llevó a cabo el desarrollo del prototipo a partir del diseño presentado de la herramienta tecnológica y de los requisitos identificados. Se desarrolló para que siguiese las bases del planteamiento estratégico y el modelo de negocio de la empresa. Éste fue desarrollado utilizando el software RapidMiner para la creación y utilización de los modelos, el lenguaje Java que realiza cambios de parámetros y llamadas a los procesos en RapidMiner y procesos ETL creados en el *software* Pentaho para la limpieza y carga de los datos.

El prototipo consiste en un algoritmo que estima la demanda de alumnos de cada curso perteneciente a los planes entregados. En combinación con el equipo de Data Science se generó un algoritmo que supliera el déficit que deja el algoritmo actual, brindando una nueva lógica para la estimación.

Para realizar la predicción el algoritmo construye modelos que crean perfiles de alumnos para cada plan de estudio. Los modelos consideran la historia de inscripciones y el estado en que los alumnos terminaron las asignaturas (cursando, no cursada, aprobada o reprobada) en los semestres anteriores. Posteriormente se aplican los modelos entrenados a los alumnos

candidatos con lo que se obtiene una probabilidad de tomar cada asignatura. Finalmente se calcula una esperanza de alumnos que tomarán las asignaturas el próximo semestre.

6.4.3 Pruebas Realizadas

Una de las pruebas que se realizó para la construcción del prototipo fue utilizando datos de una universidad mexicana. Estos datos contienen la información de estudiantes de una sola carrera y las inscripciones que realizaron. Se debía realizar una estimación para 64 asignaturas y se tenía información de cuatro periodos consecutivos.

Se muestra en la Tabla 11 los resultados obtenidos de la simulación, en donde se consiguió un 75% de asignaturas correctamente estimadas lo que representa 48 asignaturas. Las asignaturas que obtuvieron error en la predicción fueron 16 lo que equivale a un 25 % de las asignaturas. De éstas asignaturas con errores tan solo 5 tuvieron una sección de diferencia con lo real, equivalente a un 8% del total de asignaturas y 11 asignaturas con más de dos secciones de diferencia equivalente a un 17 %.

Caso	Indicador	Porcentaje	Cantidad de Cursos
Correcto	SC	75 %	48
Error	SE_1	8 %	5
Error	SE_2	17 %	11

Tabla 11: Resultados Universidad Mexicana

La diferencia entre secciones reales y predichas generadas es de 27 secciones lo que representa un SuE de 24 y un SoE de 3 secciones.

Los resultados obtenidos de la clasificación de dictar o no un curso comparando lo pronosticado versus lo real se encuentra a continuación:

		Real		
		Dictan	No Dictan	
Pronóstico	Dictan	54	0	100%
	No dictan	0	10	100%
		100%	100%	100%

Tabla 12: Tabla de Clasificación

De la Tabla 12 se puede ver que para esta simulación se obtuvo una correcta clasificación de todas las asignaturas. Por lo que se consiguió un *Recall*, *Specificity*, *Precision*, NPV y un *Accuracy* de 100%.

Para este periodo se pronosticó 12.530 inscripciones en comparación con las 11.885 inscripciones reales, lo que nos da un RI de 1,05.

A continuación en la Figura 30 se agrega el histograma de error en las secciones para cada asignatura.

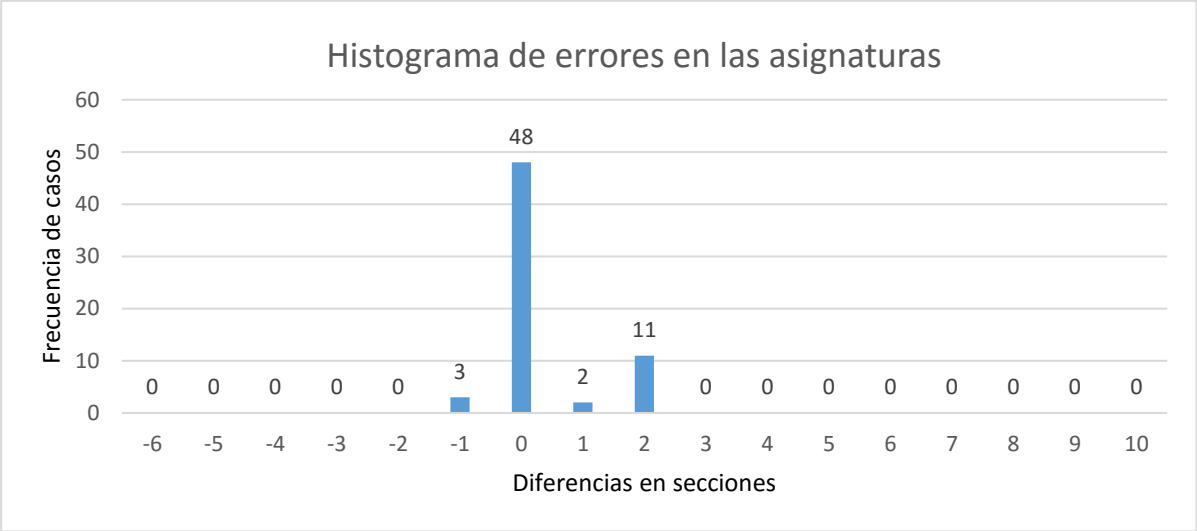

Figura 30: Histograma de errores en las asignaturas

En la Figura 30 se puede ver como se distribuyen los errores anteriormente mencionados en donde 48 asignaturas no tienen una diferencia en la generación de secciones con lo realmente ocurrido, mientras que 13 asignaturas tienen una sobreestimación y solo tres están sub estimadas.

Por último en la Figura 31 se muestra cómo se comportó el algoritmo en esta predicción para distintos tamaños de asignaturas. El algoritmo para la mayoría de las asignaturas no entrega un error mayor a 30 alumnos por lo que no se alcanza a ver reflejado en el error de secciones, pero se observa que existe un grupo de cursos para los cuales la diferencia es de 60 alumnos lo que provoca que se generen dos secciones de diferencia.

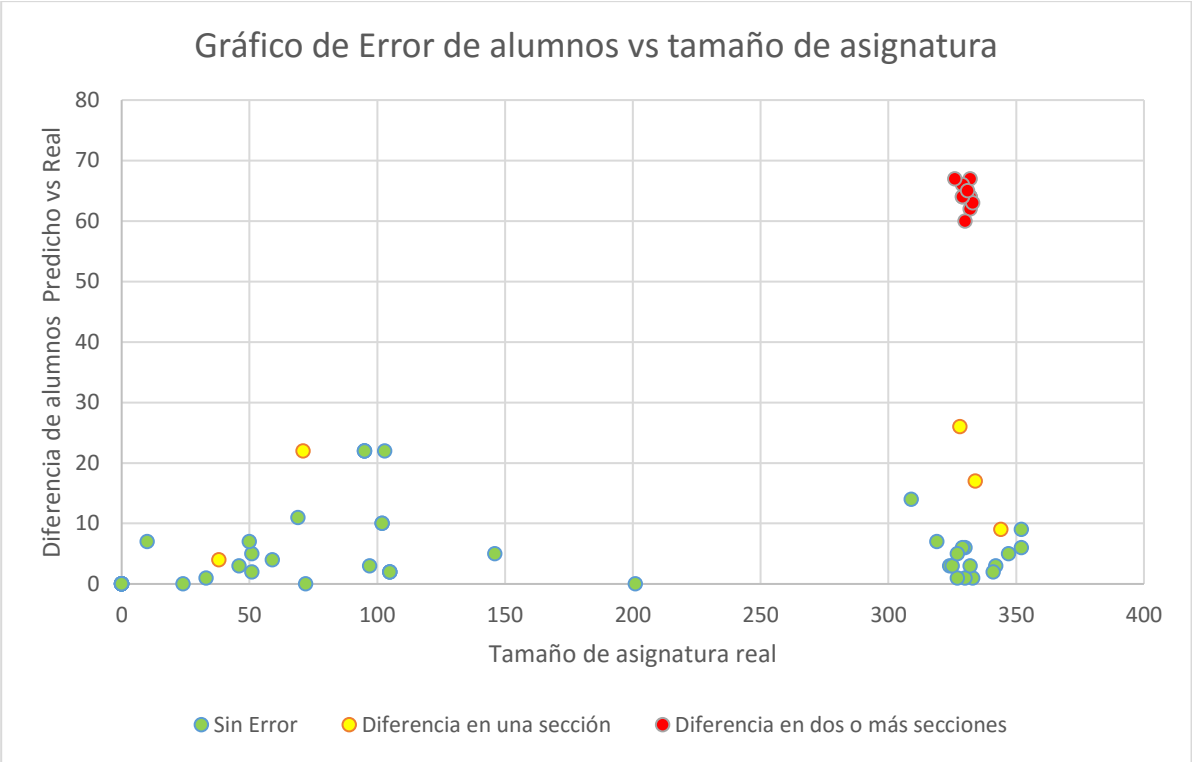

Figura 31: Gráfico de diferencia de alumnos vs tamaño de la asignatura

CAPÍTULO 7: EVALUACIÓN DEL PROYECTO

En el presente capítulo se muestra la evaluación completa del proyecto. Se comienza con una evaluación de los resultados entregados por los modelos en la prueba piloto y posteriormente se realiza un análisis de la factibilidad económica y rentabilidad del proyecto.

7.1 DEFINICIÓN DEL PLAN PILOTO

Se realizó un piloto para probar la herramienta tecnológica el cual consistió en una consultoría a una universidad de Perú. Esta consultoría tenía como objetivo estimar la demanda para el segundo semestre del periodo 2016. Para esta consultoría se debía estimar 1493 asignaturas tanto obligatorias como electivas. Esta universidad se caracteriza por tener una estructura de mallas curriculares compleja, lo que se demuestra en materias obligatorias con muchas equivalencias de ramos y una gran oferta de materias electivas, con la opción de generar *minors* o concentraciones.

Para la realización de la prueba se utilizó los datos de inscripciones históricas, estado histórico de los estudiantes, malla de los planes activos, equivalencias de ramos y convalidaciones de los alumnos.

El tiempo dispuesto para esta prueba fue de tres semanas, en donde ésta se dividió en tres etapas: una primera en donde se debió revisar los datos entregados, realizar una limpieza, encontrar inconsistencias en la información y devolver un reporte con todos los errores encontrados. En la segunda etapa se utilizó los algoritmos para hacer entrega de una estimación preliminar para todos los cursos. Y finalmente una última etapa en donde se entregó una estimación final luego de haber recibido una retroalimentación por parte del cliente acerca de los datos de la primera entrega. Para esta última etapa se tomaron en cuenta algunas reglas no consideradas en la primera estimación con lo que posteriormente se ejecutó el algoritmo nuevamente.

Se le entregó al cliente un reporte con los detalles de las estimaciones, el número de secciones mínimas que se requiere para albergar esa demanda y la variación de ambas con respecto a otros años.

Se intentó en una primera instancia utilizar ambos algoritmos de manera complementaria. El algoritmo actual para estimar la demanda de los cursos obligatorios y utilizar el algoritmo presentado en esta tesis para estimar demanda de cursos electivos (como se mencionó

anteriormente, el algoritmo actual solo funciona para asignaturas obligatorias de la malla). Dado el poco tiempo para el que fue planificada la consultoría y la alta complejidad que requiere el algoritmo actual para su ejecución se optó por solamente utilizar el algoritmo propuesto en esta tesis.

Tras la finalización de la consultoría se pudo concluir que la herramienta utilizada permite realizar una entrega de resultados más rápida en un ambiente de consultoría que el algoritmo actual, además también se logró realizar las predicciones utilizando una menor cantidad de datos de entrada.

Por último, para medir el rendimiento del algoritmo se realizó una ejecución para un semestre anterior y se comparó con lo realmente ocurrido. Para realizar esta ejecución se calibraron los modelos utilizando información hasta el segundo semestre del 2015 y luego se utilizaron estos modelos para predecir el primer semestre del 2016. A continuación se muestran los resultados e indicadores que se obtuvo de la prueba piloto.

7.1.1 Resultados Obtenidos

En esta sección se muestra en detalle los indicadores y gráficos del rendimiento del algoritmo en la consultoría, utilizando los indicadores propuestos.

El algoritmo para esta prueba obtuvo un 73,4% de asignaturas correctamente estimadas lo que significa que se pronosticó el mismo número de secciones para 1095 asignaturas. Esto implica un 26,6% de asignaturas que tuvieron algún grado de error, lo que representa a 397 asignaturas. Dentro de las asignaturas que tuvieron error 337 solo tuvieron una sección de diferencia lo que equivale al 22,6% de las asignaturas totales y 60 asignaturas con dos o más secciones de diferencia con lo real equivalente a un 4%.

Caso	Indicador	Porcentaje	Cantidad de Cursos
Correcto	SC	73,4 %	1095
Error	SE_1	22,6 %	337
Error	SE_2	4%	60

Tabla 13: Resultados de prueba piloto

Dentro del error anterior se generan secciones subestimadas o sobreestimadas, por lo que se produce una diferencia de 509 secciones, lo que representa un SuE de 155 y un SoE de 354.

A continuación se presenta la Tabla 14 con la matriz de cursos dictados y no dictados para esta predicción:

		Real		
		Se dictan	No se dicta	
Pronostico	Se dicta	245	279	47%
	No se dictan	21	948	98%
		92%	77%	80%

Tabla 14: Tabla de Clasificación prueba piloto

De la Tabla 14 se puede ver que el modelo tiene una *Precision* de 47%, es decir, que de las asignaturas que se pronosticó que se dictarán un 47% finalmente sí se dictaron. Por el otro lado tiene un NPV de 98% lo que significa que un 98% de las asignaturas que se pronosticó que no se dictarían finalmente no se dictaron. El total de casos bien clasificados brinda un *Accuracy* de 80%.

Existe un 92 % de asignaturas que realmente se dicta y se pronosticó que se dictarán (*Recall*) y un 77% de las asignaturas que realmente no se dictaron fueron pronosticadas que no se dictarían (*Specificity*).

Se puede ver que el error de tipo I es el que más afecta a la predicción. Esto se puede explicar debido a que el modelo al trabajar con probabilidades puede asignar una baja probabilidad a muchos alumnos lo que al calcular la esperanza podría llegar a obtenerse pocos alumnos que podrían inscribir esos ramos. Este es un error relativamente fácil de corregir debido a que se puede incorporar un mínimo de estudiantes para poder dictar cada asignatura, pero esto ya sería una responsabilidad que recae en el cliente. Otra de las razones que explican este error es que la universidad cambia códigos de los cursos o no reporta las equivalencias, lo que se comprobó luego de conversaciones con la universidad para mostrar el rendimiento de la predicción.

Para el periodo se obtuvo un total de 16.759 inscripciones pronosticadas en comparación con 16.826 inscripciones reales que ocurrieron en aquel periodo lo que brinda un RI de 0.99, indicador que refleja que las inscripciones obtuvieron un mismo orden de magnitud.

Se presenta en la Figura 32 el histograma de errores en secciones en donde refleja la distribución de frecuencia de asignaturas.

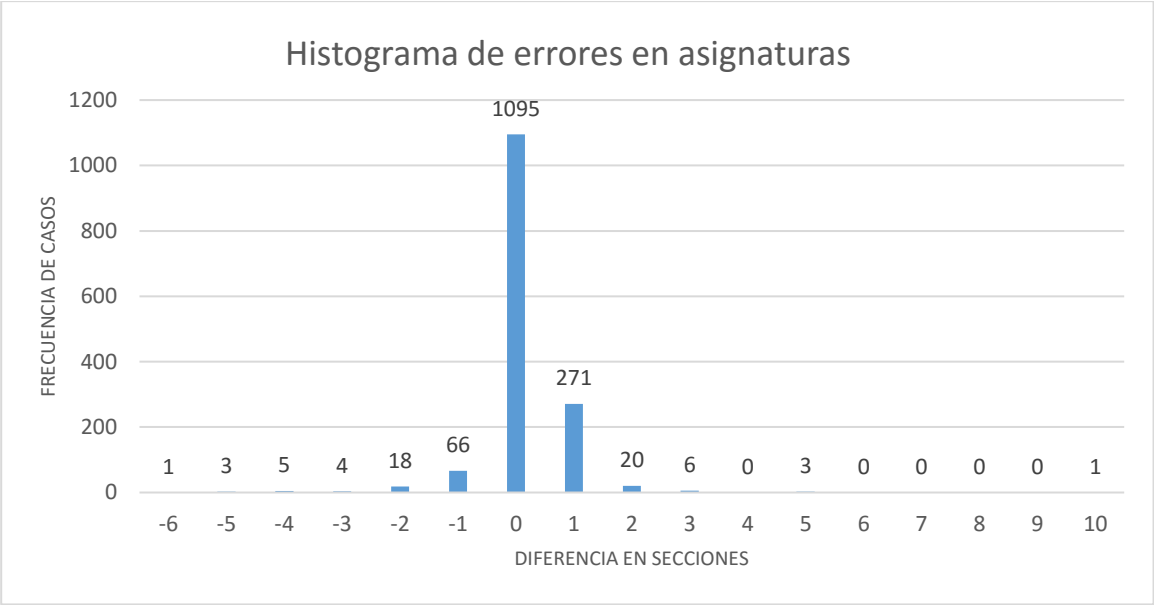

Figura 32: Histograma de errores en secciones

En la Figura 32 se puede ver que 1095 de las asignaturas no tiene una diferencia al generar secciones con lo realmente ocurrido. En donde se puede notar que se sobrestiman más los cursos de lo que se subestiman. Además, se concluye que el error en secciones tiene un comportamiento normal centrado en el cero, lo que es algo esperable.

Finalmente se muestra en la Figura 33 que la mayoría de los cursos correctamente estimados tienen una cantidad de alumnos menor a 100 y a medida que el tamaño de la asignatura aumenta el error comienza a variar más.

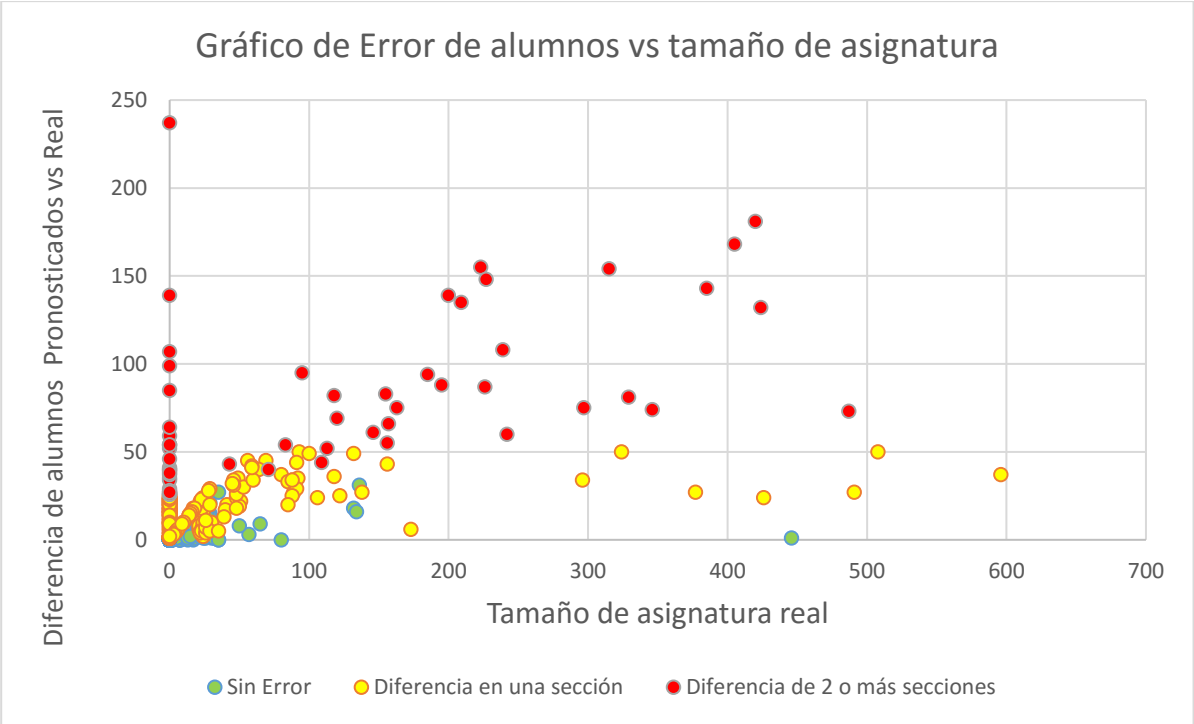

Figura 33: Gráfico de Error de Alumnos vs Tamaño de Asignaturas Prueba Piloto

Para continuar el análisis de los resultados se separaron entre ramos obligatorios y no obligatorios.

7.1.1.1 Ramos obligatorios

El algoritmo para este grupo de cursos obtuvo un 49,3% de asignaturas correctamente estimadas lo que representa un total de 100 asignaturas obligatorias. Las asignaturas que obtuvieron algún error en su predicción fueron el 51,6 % equivalente a 103 asignaturas. De estas asignaturas un 34% obtuvo solo una sección de diferencia que representa 69 asignaturas y 34 asignaturas con dos secciones o más de diferencia equivalente a un 16,8%.

Caso	Indicador	Porcentaje	Cantidad de Cursos
Correcto	SC	49,3 %	100
Error	SE_1	34 %	69
Error	SE_2	16,8 %	34

Tabla 15: Resultados de prueba piloto: Asignaturas Obligatorias

La diferencia entre secciones reales y predichas generadas es de 168 secciones lo que representa un SuE de 150 y un SoE de 18 secciones, por lo que se puede notar que existe una gran cantidad de secciones sub estimadas.

Los resultados obtenidos para la matriz de cursos dictados y no dictados, se encuentra a continuación:

		Real		
		Dictan	No Dictan	
Pronóstico	Dictan	118	5	96%
	No dictan	21	59	74%
		85%	92%	87%

Tabla 16: Tabla de Clasificación de Obligatorios

De la Tabla 16 se observa que el modelo tiene un poder de *Precision* de 96% y un NPV de 74 % lo que significa que los modelos pronostican mejor para los cursos que se dictan que para los que no se dictan. Por otro lado obtuvo un *Recall* de 85% y un *Specificity* del 92% lo que representa buenas estimaciones para la realidad de los cursos. Finalmente se obtiene un *Accuracy* del 87% lo que entrega un porcentaje global de aciertos con respecto al total.

Se puede notar que en estos cursos el error de tipo II es el que predomina, esto se debe a que algunas de estas asignaturas no tienen demanda en algunos periodos o demanda muy variable por lo que el modelo no pudo encontrar un patrón entre los cursos de las inscripciones de los alumnos.

A continuación se agrega el histograma de errores en las secciones para los ramos obligatorios.

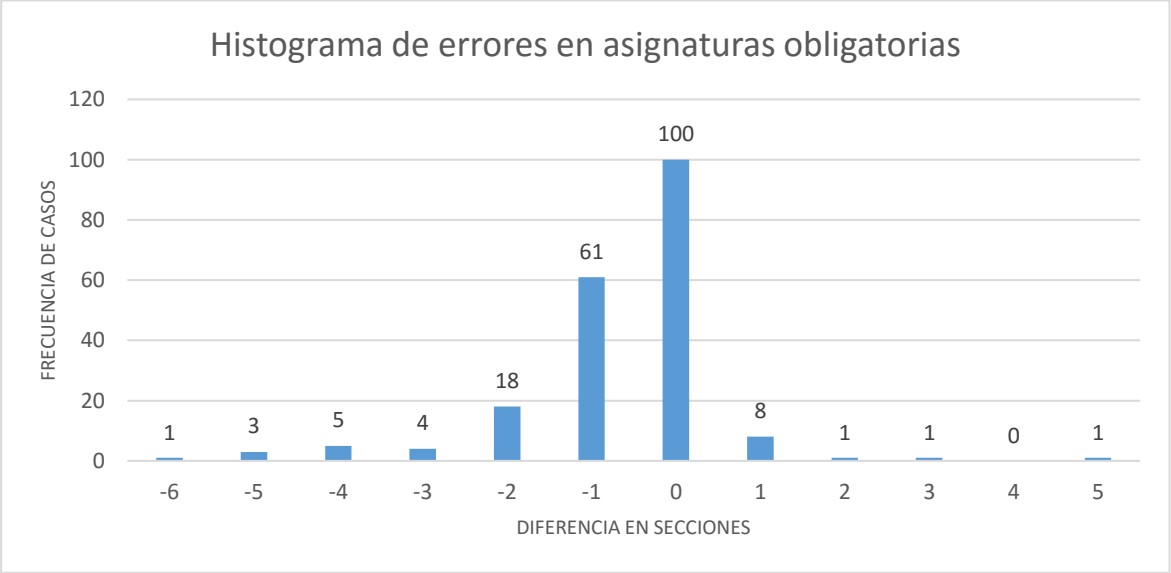

Figura 34: Histograma de asignaturas obligatorias

En la Figura 34 se puede ver que 100 de las asignaturas no tiene una diferencia en la estimación de secciones con lo realmente ocurrido. Se puede ver que la estimación de secciones continúa teniendo un comportamiento centrado en el cero pero esta vez inclinado hacia la sub estimación de cursos.

Luego en la Figura 35 se muestra el gráfico de error de alumnos vs tamaño de asignatura en la que la mayoría de los cursos correctamente estimados y con bajo error tienen una cantidad menor a 100 alumnos, pero a medida que el tamaño de la asignatura aumenta el error comienza a aumentar su varianza.

Existen también cursos que de un semestre a otro cambian sus códigos y que no fueron notificados por la universidad por lo que su demanda fue cero o cercana a cero. Lo anterior explica porque existen cursos con una alta diferencia de alumnos y un tamaño real cero.

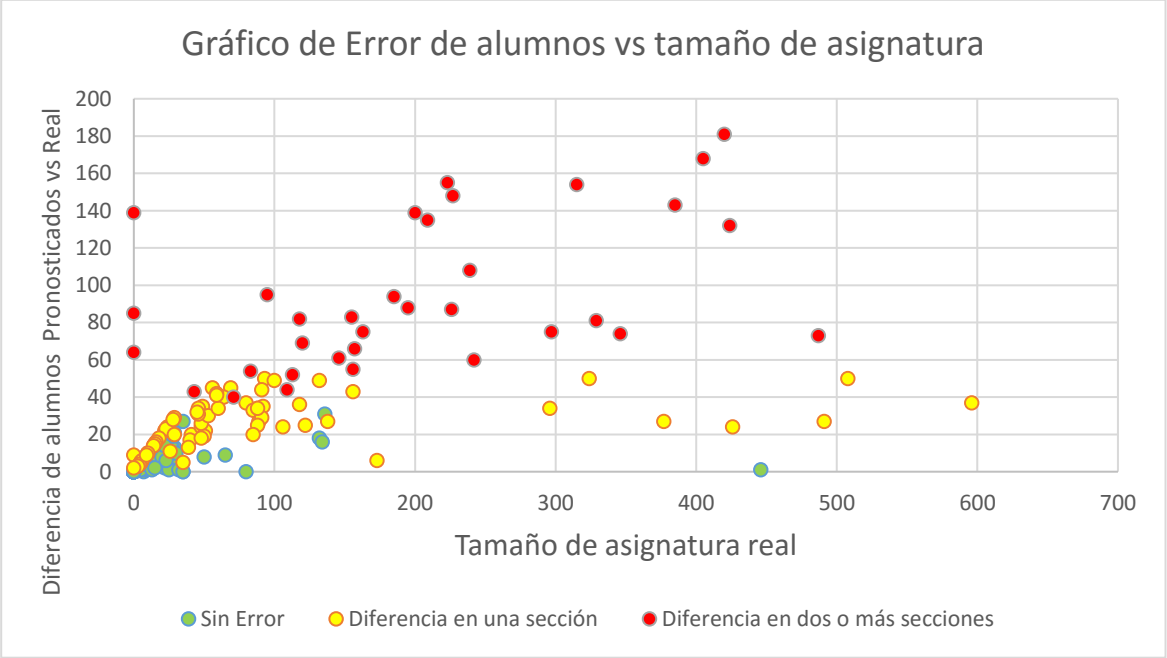

Figura 35: Gráfico de Error de Alumnos vs Tamaño de Asignaturas Obligatorios

7.2.1.2 Ramos no obligatorios

El algoritmo para este otro grupo de cursos obtuvo un 77 % de asignaturas correctamente estimadas lo que representa un total de 995 asignaturas no obligatorias. Las asignaturas que obtuvieron algún error en su predicción fueron un 23,1 % equivalente a 295 asignaturas de las cuales 268 solo tuvieron una sección de diferencia con lo real equivalente a un 21% del total de electivas y 27 asignaturas con diferencia igual o mayor a dos secciones equivalente a un 2,1%.

Caso	Indicador	Porcentaje	Cantidad de Cursos
Correcto	SC	77 %	995
Error	SE_1	21 %	268
Error	SE_2	2,1 %	27

Tabla 17: Resultados de prueba piloto: Asignaturas no obligatorias

La diferencia entre secciones reales y predichas generadas es de 341 secciones lo que representa un SuE de 5 y un SoE de 336 secciones.

Los resultados obtenidos para la matriz de cursos dictados y no dictados de cursos electivos se encuentran en la Tabla 18:

		Real		
		Dictan	No Dictan	
Pronóstico	Dictan	127	274	32%
	No dictan	0	889	100%
		100%	76%	79%

Tabla 18: Tabla de Clasificación de Electivos

De la Tabla 18 se puede ver que los modelos brindan una *Precision* del 32%, lo que quiere decir que los modelos no son tan confiables cuando dicen que un electivo se va a dictar. Por el contrario obtuvo un NPV de 100% por lo que los modelos son muy confiables al predecir que un electivo no se dictará.

Se obtuvo un *Recall* de 100% y un 76% de *Specificity* lo que significa que los modelos logran clasificar relativamente bien los casos reales. Para el total de los casos se obtuvo un *Accuracy* de un 79%.

Se puede notar que en estos cursos el error de tipo I es el que predomina, esto se debe a lo mencionado anteriormente para el caso general.

A continuación se agrega el histograma de errores en las secciones para los ramos no obligatorios.

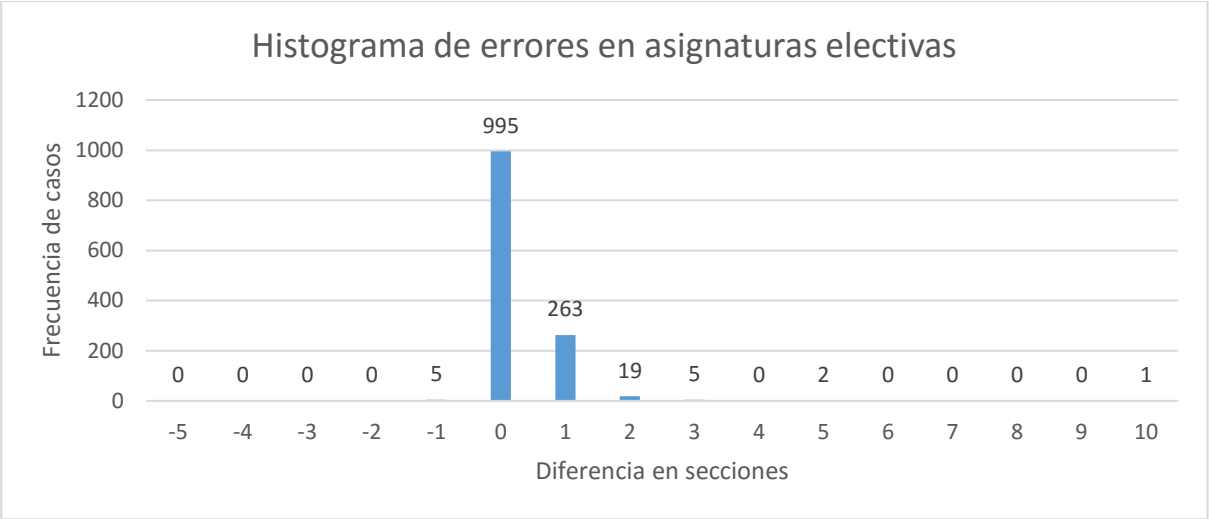

Figura 36: Histograma de asignaturas electivas

En la Figura 36 se puede ver que 995 de las asignaturas no tiene una diferencia en la predicción de secciones con lo realmente ocurrido. También se puede notar que la mayoría de los errores son sobre estimados solamente en una sección. Muchos de los casos que generan el error de tipo I en la Tabla 18 contienen un error bajo en secciones. Para este grupo de cursos se observa que la estimación de secciones continua teniendo un comportamiento centrado en el cero pero esta vez inclinado hacia la sobre estimación, lo que es esperable.

En el gráfico de error de alumnos vs tamaño de asignatura de la Figura 37 se muestra que la mayoría de los cursos al tener un tamaño pequeño de alumnos, menor a 60, contienen un error bajo.

Se puede notar que los cursos que no tuvieron demanda en la realidad para dicho periodo pero el algoritmo predijo que sí tendrían algunos alumnos, fueron los que causaron el mayor impacto en el error promedio. Esto se puede ver en la mayoría de cursos que están en rojo o amarillo.

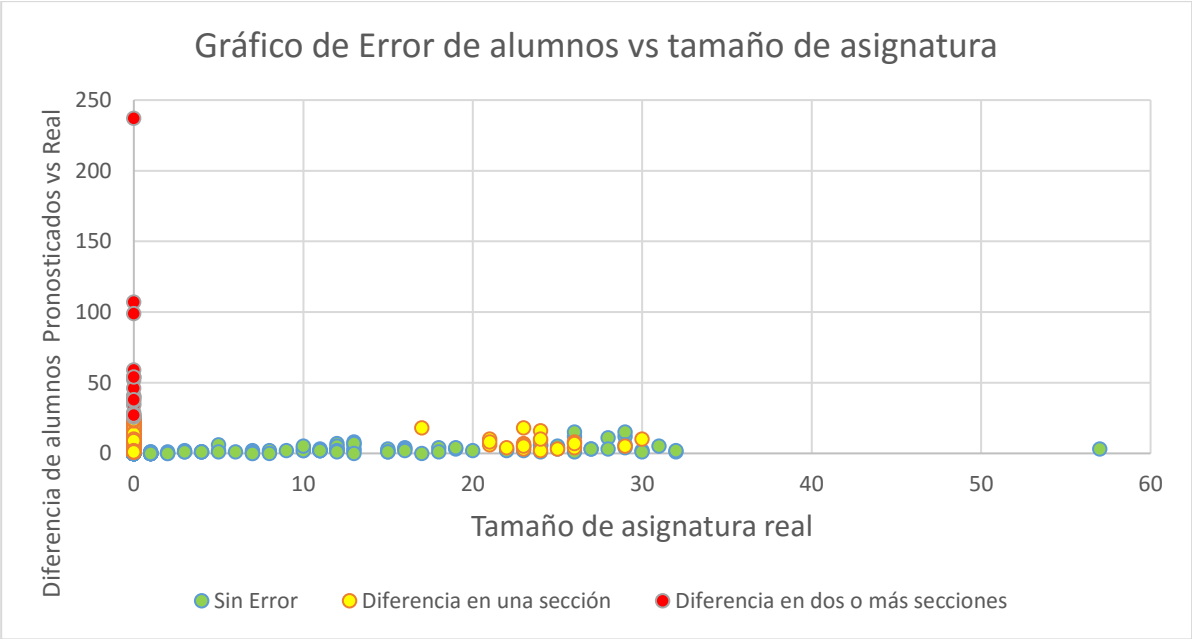

Figura 37: Gráfico de Error de Alumnos vs Tamaño de Asignaturas Electivas

Se puede notar finalmente que para esta prueba un bajo porcentaje de asignaturas (22,6%) puede generar un alto error en las secciones generadas (509). También se pude notar que el tamaño de las asignaturas tiene una relación directa con la precisión de la predicción.

7.2 PROPUESTA PARA LA UTILIZACIÓN DEL NUEVO ALGORITMO

Tras obtener los resultados de la predicción se propone una nueva metodología del uso del algoritmo con el fin de mejorar los resultados y el rendimiento en los índices. La metodología propuesta busca elegir el mejor resultado entre varios entregados por distintos modelos calibrados, para cada uno de los cursos. Se propone esta metodología para que en un trabajo futuro pueda programarse y automatizarse.

Se propone esta metodología debido a que varios de los cursos calculados para la prueba piloto poseen una clara periodicidad en su demanda. Y como se mencionó en la sección anterior los modelos fueron entrenados utilizando el segundo periodo del 2015 como periodo actual para la matriz de entrada de la Figura 20 (de ahora en adelante se llamará al periodo actual en la matriz de entrada como periodo objetivo para no incurrir en confusiones). Esta forma de calibrar entrega predicciones que a veces no son buenas, debido a que no se tiene suficientes casos de éxito (alumnos inscribieron el curso) en ese periodo.

A continuación se mencionan los pasos de la metodología propuesta para universidades con dos periodos por año:

1. Generar una calibración con el periodo pasado como periodo objetivo y generar una predicción.

2. Generar una calibración con el periodo antepasado como periodo objetivo y generar una predicción.

3. Calcular demanda histórica de los cursos.

4. Crear indicador de similitud, para esto se definen los siguientes conceptos:

$I_{ij} = Indicador\ de\ similitud\ para\ el\ curso\ j\ de\ la\ predicción\ i.$

$\mu_j = Promedio\ de\ los$ últimos $cuatro\ periodos\ distintos\ a\ cero\ del\ curso\ j.$

$\sigma_j = Desviación\ estandar\ de\ los$ últimos $cuatro\ periodos\ distintos\ a\ cero\ del\ curso\ j.$

Luego se calcula el indicador de similitud de la siguiente manera:

$$I_{ij} = \frac{|x_i - \mu_j|}{\sigma_j} \qquad (16)$$

5. Se elige la predicción que genere el menor indicador de similitud.

Para universidades que tengan más de dos periodos por año se deberán realizar una calibración y predicción por cada uno de los periodos.

Para finalizar se somete la predicción a una corrección con respecto a la demanda histórica en donde se toma los periodos equivalentes y se genera una relación con el periodo siguiente. Utilizando esta relación se calcula un valor esperado del siguiente periodo, un límite superior y un límite inferior para este valor.

Se define la relación como:

$$Y_{it} = g(t) + \varepsilon_{it} \qquad (17)$$

$Y_{it} = Demanda\ para\ la\ asignatura\ i\ en\ el\ periodo\ equivalente\ t.$

$\varepsilon_{it} = Error\ en\ la\ demanda\ para\ el\ periodo\ t\ de\ la\ asignatura\ i.$

En donde $g(t)$ es un polinomio. Para este caso se probó con $g(t) = \beta_i * t + \beta_{i0}$. Se deja como trabajo a futuro probar con diferentes funciones para $g(t)$.

Se define los parámetros

$\beta_i = Tendencia\ de\ la\ demanda\ de\ la\ asignatura\ i.$

$\beta_{i0} = Intercepto\ de\ la\ asignatura\ i.$

Con lo que se tiene la siguiente relación

$$Y_{it} = \beta_i * t + \beta_{i0} + \varepsilon_{it} \qquad (18)$$

Se trabaja bajo el supuesto que los errores siguen una distribución normal $\varepsilon_{it} \sim N(0, \sigma)$, utilizando las propiedades de la distribución normal se construye un intervalo de confianza de la siguiente manera [19]:

$$\beta_i * (t+1) + \beta_{i0} - z_{95\%} * \frac{\sigma}{\sqrt{t}} \leq Y_{i(t+1)} \leq \beta_i * (t+1) + \beta_{i0} + z_{95\%} * \frac{\sigma}{\sqrt{t}}$$

Este intervalo permite acotar la predicción para los casos en que se arrojen resultados muy fuera de lo corriente.

A continuación se mostrarán los resultados e indicadores de rendimiento utilizando la metodología propuesta.

7.2.1 Resultados utilizando la metodología propuesta

Para mostrar el resultado de la metodología se realizó una predicción para estimar la demanda del segundo semestre del 2016. Para esta predicción se calibraron los modelos con el primer semestre del 2016 y el segundo semestre del 2015 como periodos objetivos. Se mostrará una tabla comparativa de todos los indicadores para los cursos obligatorios.

Para enriquecer la comparación se generó una estimación utilizando modelos de series de tiempo ARIMA [20] para cada asignatura. Éstos fueron construidos solamente para realizar esta predicción y se usó el software R para su construcción. Se consideran un buen punto de comparación los modelos de series de tiempo, debido a que son modelos de fácil construcción que cualquier universidad podría realizar.

Indicador	Sin metodología	ARIMA	Con metodología
SC	49,3 %	54%	72%
SE_1	34 %	33,7%	23%
SE_{2+}	16,8 %	12,4%	5%
SuE	150	101	43
SoE	18	41	37
RI	0.73	0.88	0.99

Tabla 19: Comparación indicadores de rendimiento.

Aquí se muestra que los indicadores mejoran al utilizar la metodología, en donde las asignaturas estimadas correctamente suben de un 49.3% a un 72%, el error en una sección baja de un 34% a un 23 % y el error en dos secciones o más, baja de 16.8% a un 5%. Las secciones subestimadas (SuE) bajan de 150 a 43 secciones. El único indicador que empeora para estas asignaturas es el de secciones sobrestimadas que sube de 18 a 37 secciones. Finalmente la razón entre inscripciones (RI) sube a 0,99 lo que significa que con la metodología el número de inscripciones predichas se acerca al número de inscripciones reales.

Con respecto al modelo de series de tiempo, el algoritmo muestra una mejor predicción de secciones (SC) y menor porcentaje de error en secciones (SE_{2+} y SE_1) al utilizar la metodología. También se obtiene una menor cantidad de secciones subestimadas (SuE) y sobreestimadas (SoE). Se puede observar que al no utilizar la metodología el algoritmo pierde poder de predicción, en donde un modelo simple de series de tiempo logra generar mejores predicciones.

A continuación se muestra la matriz de cursos dictados y no dictados de cursos obligatorios.

		Real		
		Dictan	No Dictan	
Pronóstico	Dictan	142	3	98%
	No dictan	12	46	79%
		92%	94%	93%

Tabla 20: Tabla de Clasificación de Obligatorios con nueva metodología

Los resultados para la matriz de la Tabla 20 también mejoran con respecto a los observados en la Tabla 16 en donde el *Accuracy* sube de un 87% a un 93%, el poder de *Precision* sube de un 96% a un 98% y el NPV de un 74% a un 79%. El *Recall* sube de un 85% a un 92% y la *Specificity* sube de un 92% a un 94%. Como se puede ver con la ayuda de la metodología es posible clasificar de una mejor manera los cursos que se dictan y los que no se dictan.

A continuación se analiza el histograma de errores en secciones

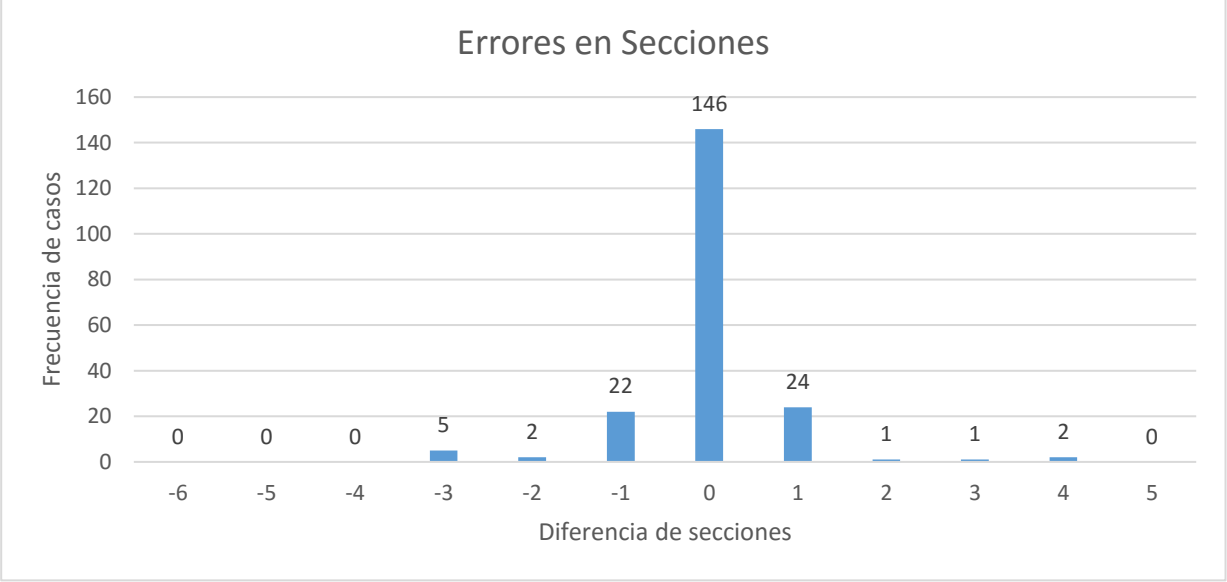

Figura 38: Histograma de errores en secciones con metodología propuesta

Al realizar la comparación con la Figura 34 se puede notar que las asignaturas correctamente estimadas suben de 100 a 146. Las asignaturas subestimadas en una sección bajan de 61 a 22 secciones y las asignaturas subestimadas en dos o más secciones bajan de 31 a 7 asignaturas. Solamente las asignaturas sobrestimadas en una sección suben de 8 a 24 asignaturas, mientras que las asignaturas sobrestimadas por dos o más secciones se mantienen.

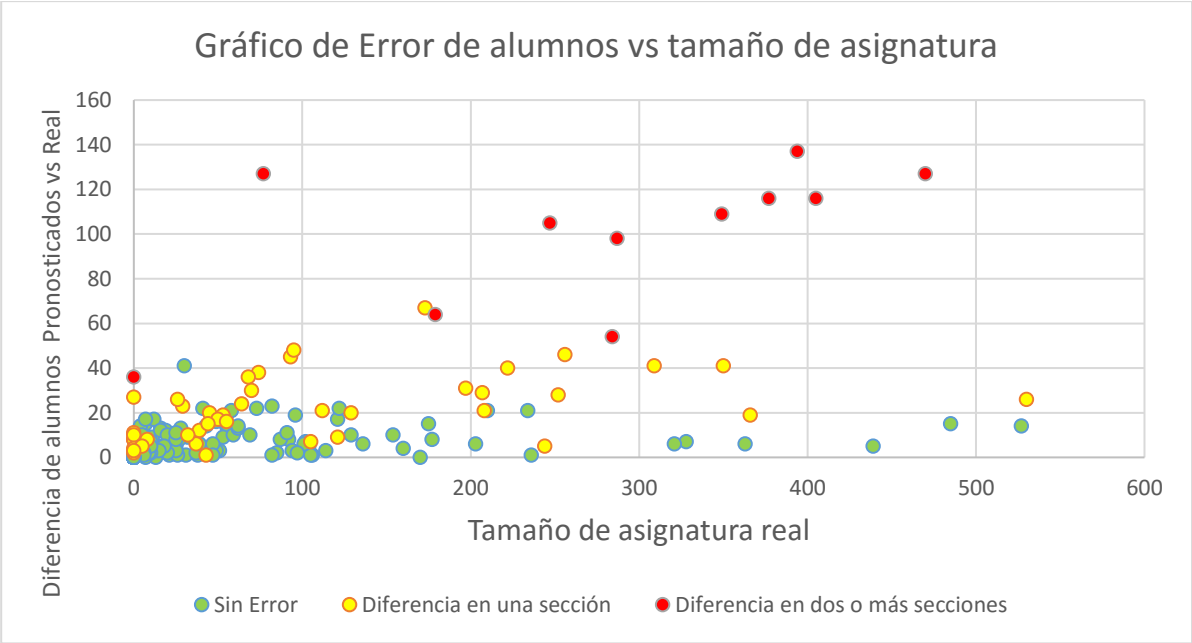

Figura 39: Gráfico de Error de Alumnos vs Tamaño de Asignaturas utilizando metodología propuesta.

Es posible observar que con la metodología los resultados de la Figura 39 mejoran en relación a la Figura 35. Una muestra de ello es que se pueden ver más asignaturas sin errores y tan solo once asignaturas con diferencia de dos o más en secciones (puntos rojos). Mientras que en la Figura 35 las asignaturas sin error son solamente asignaturas de pequeño tamaño y existían 34 asignaturas con dos o más secciones de error.

Se concluye finalmente que con la utilización de la metodología propuesta aumenta el potencial del algoritmo presentado en esta tesis, en donde se sube en un 22.7% la cantidad de asignaturas con secciones correctamente estimadas. Además, se consiguen las siguientes ventajas a partir del uso del algoritmo:

- **Modelos de fácil comprensión**: se generan modelos de árboles de decisiones, en donde una de las características principales de éstos es el ser sencillos de entender.

- **Utilización de equivalencias y convalidaciones de asignaturas**: el algoritmo propuesto soporta la utilización de equivalencias y convalidaciones, no así modelos simples como las series de tiempo.

- **No hay necesidad de prerrequisitos:** esto permite poder predecir cursos a universidades que poseen mallas más flexibles.

7.3 EVALUACIÓN ECONÓMICA

En la presente sección se presentan las principales variables que afectan la evaluación económica del proyecto. Al definir estas variables, se analizarán los beneficios del proyecto realizando distintos análisis de sensibilidad sobre éstas.

7.3.1 Análisis de Variables Relevantes

A continuación se analizan las variables más importantes para la realización de la evaluación económica del proyecto. Dentro de estas es posible distinguir las siguientes: plazo del proyecto, tasa de descuento, estructura de ingresos y estructura de costos.

- Plazo del proyecto: se toma como horizonte de tiempo para la evaluación económica un periodo de 5 años. Esto debido a que los contratos con los clientes, en general, duran esta cantidad de tiempo, por lo que los ingresos se van generando a lo largo de esos años.

- Tasa de descuento: se utilizan las tasas utilizadas por la empresa, la cual realiza evaluaciones del tipo privada. Cabe mencionar que la empresa no tiene una tasa de descuento definida para sus proyectos, es por esto que elige entre un rango posible de tasas que definen distintos escenarios para los proyectos.

- Estructura de ingresos: El proyecto busca aumentar la capacidad de la empresa para poder satisfacer a clientes cuyas mallas curriculares son muy flexibles y/o en donde los alumnos no siguen un avance de malla regular. Es por esto que los ingresos que se consideran son los que vendrán de las ventas de U-Forecast a clientes que antes no se podía satisfacer.

- Estructura de costos: La estructura de costos está conformada principalmente por los costos de inversión y el uso de los recursos humanos asociados a las integraciones de los sistemas que se deben realizar para el flujo automático de información. Las integraciones se realizan solo una vez al momento de adquirir uno de los productos. También existe en un costo menor asociado al arriendo de servidores. Por último cabe mencionar que todos los costos tanto de inversión como operacionales son financiados internamente por la empresa.

7.3.2 Inversión

La inversión del proyecto está compuesta principalmente por los recursos humanos utilizados para el desarrollo de la herramienta tecnológica, estos se detallan a continuación:

Para la construcción del algoritmo se necesitan distintos perfiles de profesionales, se debe contar con un jefe de proyecto que esté encargado del avance del proyecto, un analista de procesos TI encargado de definir y guiar el rediseño, un analista TI especializado en ETL el que ayudará en la obtención y limpieza de los datos, analista matemático encargado de la creación de los modelos y un jefe de desarrollo que ayudará a la integración de todas las partes.

A continuación se presenta la Tabla 21 en donde se muestra el sueldo promedio para cada uno de los perfiles, las horas hombre aproximadas utilizadas y el monto que significa la utilización de estos recursos

Inversión			
Desarrollo	**Sueldo Promedio**	**HH**	**Monto total**
Jefe de Proyecto	2,070,000	144	1.656.000
Analista de Procesos y TI	1,035,000	748	4.301.000
Analista TI ETL	1,035,000	72	414.000
Analista Matemático	1,242,000	432	2.980.800
Jefe de Desarrollo	1,380,000	72	552.000
		Total	9.903.800

Tabla 21: Inversión del desarrollo

7.3.3 Costos Operacionales

Existen costos asociados a la integración que se debe realizar para lograr una comunicación entre los sistemas de las universidades y la empresa para poder utilizar los productos. Es por esto que también se necesita utilizar recursos humanos para lograr la implementación en el cliente.

Para esto se hace necesario la disposición de un jefe de proyectos, un analista de TI, un jefe de desarrollo, un analista matemático y un analista TI experto en ETL.

Costos Operacionales			
Implementación	**Sueldo Promedio**	**HH**	**Monto Total**
Jefe de Proyectos	2.070.000	275	3.154.795
Analista TI	1.035000	212	1.219.000
Jefe de Desarrollo	1.380.000	148	1.134.667
Analista Mat	1.242.000	34	231.840
Analista TI ETL	1.035.000	40	230.000
		Total	5.970.302

Tabla 22: Costos Operacionales

Para algunas universidades es necesario arrendar servidores ajenos debido a que éstas no tienen la infraestructura para integrarse con U-planner. Es por esto que se deben hacer las instalaciones de forma remota en donde se alojarán tanto el sistema como los algoritmos. Para el estudio se hizo la suposición de que todas las universidades podrían tener este inconveniente. Este costo se estimó en 1.379.104 pesos anuales, el que se basó a partir de arriendos anteriores realizado por la empresa.

7.3.4 Ingresos

Como anteriormente se ha mencionado el precio de venta de los productos es proporcional a la cantidad de alumnos que tiene la universidad, lo que implica que universidades con mayor número de alumnos son más complejas por lo tanto tienen un precio mayor para un mismo producto.

Para realizar la estimación de los ingresos se aplicaron ciertos supuestos que ayudan al cálculo de los beneficios. El primer supuesto es que se le vendería el producto a universidades que tengan la característica de tener mallas complejas y difíciles de predecir. Estas universidades dentro de la evaluación tendrán el promedio de alumnos que tienen los actuales clientes de U-planner para el producto U-Forecast. Por lo tanto los ingresos por venta de este producto para cada universidad serían de 5.589.982 pesos chilenos.

7.4 FLUJO DE CAJA

En la Tabla 23 se puede apreciar el flujo de caja del proyecto evaluado en un plazo de 5 años. Se consideró una tasa de descuento del 15% obtenida a partir de proyectos similares realizados en la empresa. Todos los valores están expresados en pesos chilenos.

Para el proyecto se obtiene un **VAN de 9.800.000** aproximadamente, y una **TIR de 19,7%** para un escenario en donde se le vende a tres clientes el cual tiene unas altas posibilidades de ocurrir, considerando que este último año se desaprovecharon ventas con 10 clientes debido al problema abordado en esta tesis. Al observar los valores notamos que el proyecto es rentable.

Ítem	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
(+) Ingresos		16,769,947	16,769,947	16,769,947	16,769,947	16,769,947
(-)Costos Fijos		-22,307,795	-4,137,312	-4,137,312	-4,137,312	-4,137,312
(-)Pérdidas del Ejercicio Anterior			-5,537,848			
= Utilidad Antes de Impuestos		-5,537,848	7,094,787	12,632,635	12,632,635	12,632,635
(-) Impuesto de Primera Categoría (25%)		0	-1,773,697	-3,158,159	-3,158,159	-3,158,159
= Utilidad Después de Impuestos		-5,537,848	5,321,090	9,474,476	9,474,476	9,474,476
(+)Pérdidas del Ejercicio Anterior			5,537,848			
= Flujo de Caja Operacional		-5,537,848	10,858,938	9,474,476	9,474,476	9,474,476
(-)Capital de Trabajo	-9,903,800					
= Flujo de Caja	-9,903,800	-5,537,848	10,858,938	9,474,476	9,474,476	9,474,476

Tabla 23: Flujo de Caja

7.5 ANÁLISIS DE SENSIBILIDAD

7.5.1 Sensibilidad sobre la cantidad de clientes

Se realizó un análisis para identificar la sensibilidad de la variable de cantidad de clientes. Para esto se analizaron posibles escenarios para distintas cantidades de clientes. Se definieron los escenarios posibles con ventas desde un cliente hasta diez clientes durante el primer año.

Se observa que el escenario en el que se realiza una sola venta se obtiene un **VAN negativo de 3.319.000 aproximadamente y un TIR negativo de 10,1%**, por lo que es importante tener en mente esta posibilidad para tomar acciones pertinentes y superar este escenario. En la Figura 40 se muestra la variación del VAN y TIR para distintas cantidad de clientes y se puede observar que para escenarios con ventas mayores a dos clientes el proyecto comienza a ser rentable.

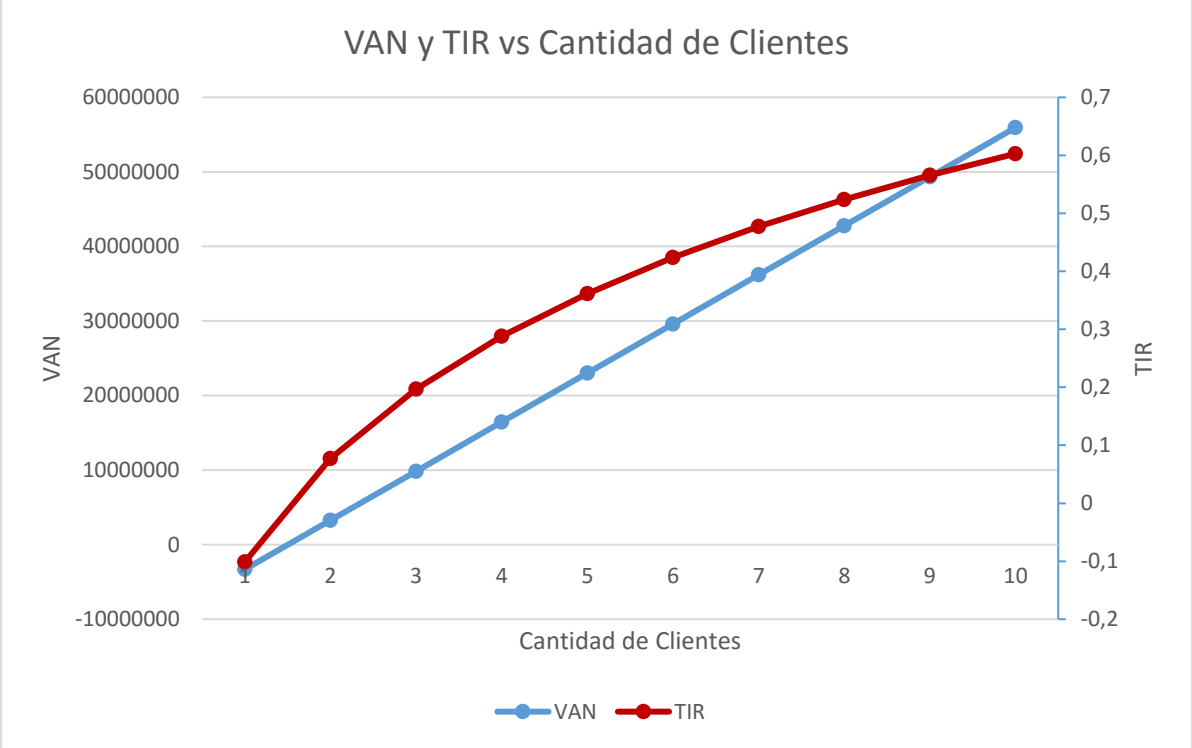

Figura 40: VAN y TIR vs Cantidad de Clientes

Se puede observar de la Figura 40 que el VAN tiene un comportamiento lineal proporcional respecto a esta variable. Mientras que la TIR es creciente, pero va disminuyendo su pendiente a medida que aumentan los clientes. Analizado lo anterior se concluye que el factor crítico de éxito económico del proyecto es que este logre atraer a más clientes.

7.5.2 Sensibilidad de las tasa de descuento.

Se realiza un segundo análisis sobre la sensibilidad en la rentabilidad respecto a la tasa de descuento que se aplicará en el proyecto.

Se hace necesario analizar esta variable debido a que la empresa no cuenta con una tasa de descuento predeterminada para la evaluación de sus proyectos. A continuación se muestra en la Figura 41 la variación del VAN y la TIR con respecto a la tasa de descuento para un escenario moderado en donde se le venderá a tres clientes.

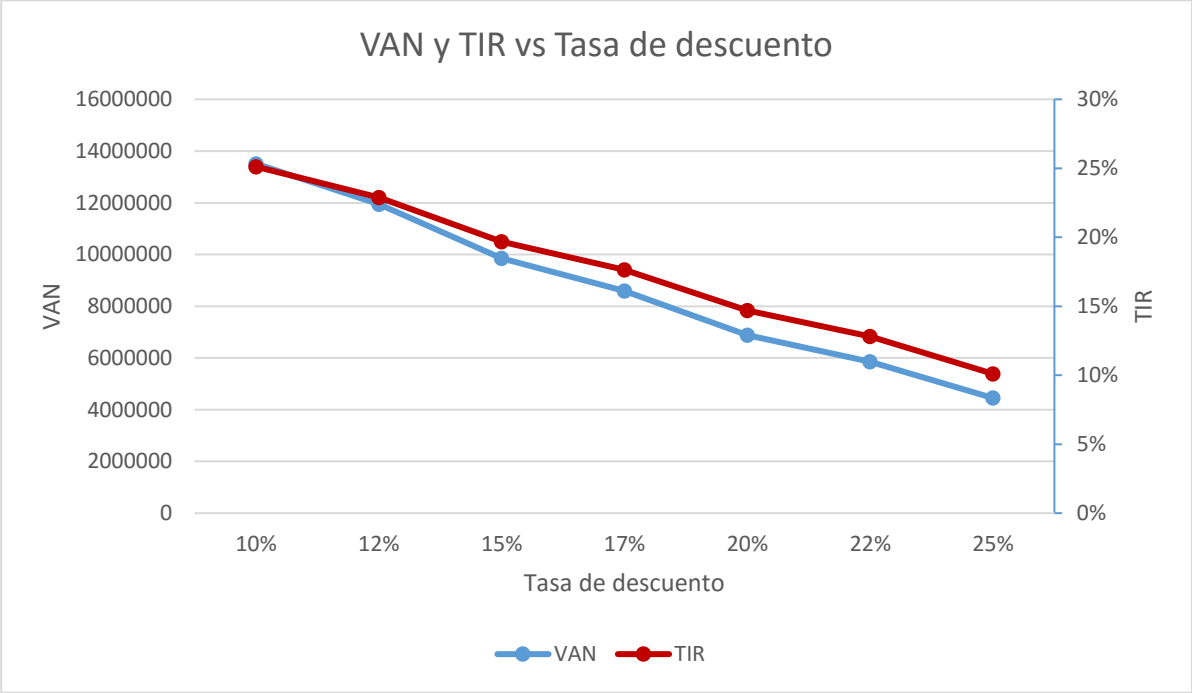

Figura 41: VAN y TIR vs Tasa de descuento

El VAN y la TIR tienen un comportamiento lineal y decreciente con respecto a la tasa de descuento, lo que es esperable. Se puede concluir a partir de la Figura 41 que para las tasas de descuento probables, el proyecto siempre es rentable.

CAPÍTULO 8: CONCLUSIONES

8.1 Estimación de Demanda

El problema de saber cuántos alumnos tendrán las universidades en las salas de clase es un tema que ha sido muy poco estudiado en la literatura y en general las universidades deben enfrentar este problema utilizando su propia experiencia, basado sólo en información de semestres anteriores y utilizando soluciones específicas para su estructura de cursos.

La empresa U-planner mediante su producto U-Forecast ha tomado un primer paso en la búsqueda de un producto que pueda enfrentar esta problemática de forma genérica para distintos clientes. Sin embargo se torna complicado estimar la demanda de universidades utilizando una sola lógica. Esto debido principalmente a la amplia estructuras de mallas curriculares de las universidades en Latinoamérica. Por lo que se hace necesaria una actualización del producto U-Forecast.

Este proyecto de tesis busca ser una mejora para el producto U-Forecast, permitiendo a la empresa mejorar y medir el rendimiento del actual algoritmo y poder vender este producto a universidades que antes no se les podía vender, debido a la complejidad de sus mallas curriculares.

También el proyecto es un aporte en la materia dado que introduce técnicas de machine learning en una problemática para la cual no se había realizado antes.

Para este trabajo se logra la realización de una prueba piloto que consiste en una consultoría, un evento cercano a la realidad, en donde se puso a prueba el algoritmo propuesto. Esto dio como resultado un 73.4% de asignaturas correctamente estimadas y 22.6% de asignaturas que tienen una sección de error. Para las asignaturas obligatorias se obtuvo un 49% de asignaturas correctamente estimadas, 34% de asignaturas con una sección de error y un 16% de asignaturas con dos o más secciones de error. Para mejorar estos resultados se propuso una metodología que elige el mejor modelo de predicción para cada curso y corrige de acuerdo a la historia. Utilizando esta metodología se subió el porcentaje de asignaturas correctamente estimadas a un 72%, se bajó el porcentaje de asignaturas con una sección de error a 23% y se bajó el porcentaje de asignaturas con dos o más secciones de error a un 5%.

Con la utilización del algoritmo y la metodología se logra las ventajas de modelos de fácil comprensión, poder utilizar un modelo que tome un conjunto completo de información sin tener que usar datos no estrictamente necesarios como los prerrequisitos.

Se presentaron además métricas para medir el rendimiento de las predicciones, lo que facilita posteriormente la comparación entre estimaciones y algoritmos. Además estas métricas permiten determinar cómo se comportan los algoritmos y qué tipo de errores se está obteniendo.

8.2 Trabajo Futuro

Queda como trabajo a realizar la integración de la lógica propuesta con la existente, en donde se definan reglas que escojan los casos en los que se usará cada algoritmo y se pueda ejecutar de manera automática. Además en un futuro se deberá automatizar el proceso de mantención y creación de los modelos, que deberá incluir reglas de la metodología propuesta en la sección 7.2. Por último se deberá realizar la integración del algoritmo con la aplicación de U-planner para que ésta pueda ser utilizada por los clientes directamente.

Este proyecto abre la alternativa de poder seguir generando lógicas nuevas que ayuden en la disminución del error al estimar la demanda y que puedan fácilmente integrarse en el producto U-Forecast.

Esto abre las puertas para un siguiente paso en el cual se genere más investigación de las aplicaciones de la ingeniería de negocios y la minería de datos para los procesos y productos que actualmente tiene U-planner y los que se construirán en un futuro.

CAPÍTULO 9: BIBLIOGRAFÍA

1.- Barros, O. (2004). Ingeniería e-Business: Ingeniería de Negocios para la Economía Digital. JC Sáez Editor.

2.- Barros, O. (2009). Ingeniería de Negocios Diseño Integrado de Negocios, Procesos y Aplicaciones TI. Tercera Parte.

3.- Barros, O. (2009). Ingeniería de Negocios Diseño Integrado de Negocios, Procesos y Aplicaciones TI. Segunda Parte.

4.- Zambrano, A. C. (2005). Herramienta para el análisis de requerimientos dentro de la pequeña empresa desarrolladora de software en Bogotá. Proyecto de grado de ingeniería de sistemas, Pontificia Universidad Javeriana, Bogotá.

5.- Barros, Ó. (2013) Ingeniería de Negocios: Diseño Integrado de Servicios, sus Procesos y Apoyo TI.

6.- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.

7.- Betancur-Calderón, D., & Moreno-Cadavid, J. (2012). A Multi-Agent Approach for the Extract-Transform-Load Process Support in Data Warehouses. Tecno Lógicas, (28), 89-107.

8.- Ríos, S. A., Velásquez, J. D., Yasuda, H., & Aoki, T. (2006, September). Conceptual classification to improve a web site content. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 869-877). Springer Berlin Heidelberg.

9.- GestioPolis. (2014, Julio 11). *Gestión de la calidad total en organizaciones educativas*. [En línea] Recuperado de http://www.gestiopolis.com/gestion-de-la-calidad-total-en-organizaciones-educativas/. Fecha de última consulta 27 de enero de 2017.

10.- The Guardian (2013, Noviembre 12). *Does university management matter?* [En línea] Recuperado de https://www.theguardian.com/higher-education-network/blog/2013/nov/12/university-management-teaching-research-impact. Fecha de última consulta 27 de enero de 2017.

11.- Hax, A. C. (2009). The delta model: reinventing your business strategy. Springer Science & Business Media.

12.- Barros, O., & Julio, C. (2011). Enterprise and process architecture patterns. Business Process Management Journal, 17(4).

13.- «IDEF0, » [En línea]. Available: http://www.idef.com/.

14.- Osterwalder, A., & Pigneur, Y. (2010). Business model generation: a handbook for visionaries, game changers, and challengers. John Wiley & Sons.

15.- Porter, M. E. (2008). Competitive strategy: Techniques for analyzing industries and competitors. Simon and Schuster.

16.- Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

17.- Kaplan, R. S., & Norton, D. P. (1995). Putting the balanced scorecard to work. Performance measurement, management, and appraisal sourcebook, (vol. 66, pp. 17511).

18.- «Object Management Group Business Process Model and Notation, » [En línea]. Available: http://www.bpmn.org. Fecha de última consulta 27 de Enero de 2017.

19.- Velasco Sotomayor, G., & Wisniewski, P. M. (2001). Probabilidad y estadística para ingeniería y ciencias.

20.- Wei, W. W. S. (1994). Time series analysis. Reading: Addison-Wesley publ.

21.- Ranking Web of Universities (2016, Julio). *Number of Universities and HEIs*. [En línea]. http://www.webometrics.info/en/node/24. Fecha de última consulta 27 de Enero de 2017.

22.- BPMNPoster [En línea]. http://bpmb.de/index.php/BPMNPoster. Fecha de última consulta 27 de Enero de 2017.

23.- Ríos, S. A., & Muñoz, R. (2014). Content patterns in topic-based overlapping communities. The Scientific World Journal, 2014.

24.- Ríos, S. A., & Aguilera, F. (2010). Web intelligence on the social web. In Advanced Techniques in Web Intelligence-I (pp. 225-249). Springer Berlin Heidelberg.

25.- Rıos, S. A. (2007). A study on web mining techniques for off-line enhancements of web sites (Doctoral dissertation, Ph. D Thesis).

26.- Erazo, L., & Ríos, S. A. (2014). A benchmark on automatic obstructive sleep apnea screening algorithms in children. Procedia Computer Science, (vol. 35, pp. 739-746).

CAPÍTULO 10: ANEXOS

ANEXO A: Notación BPMN para Modelamiento de Procesos

En la Figura 42 se aprecia la notación para diagramar actividades en notación BPMN. Las principales actividades corresponden a tareas generadas por actores del proceso.

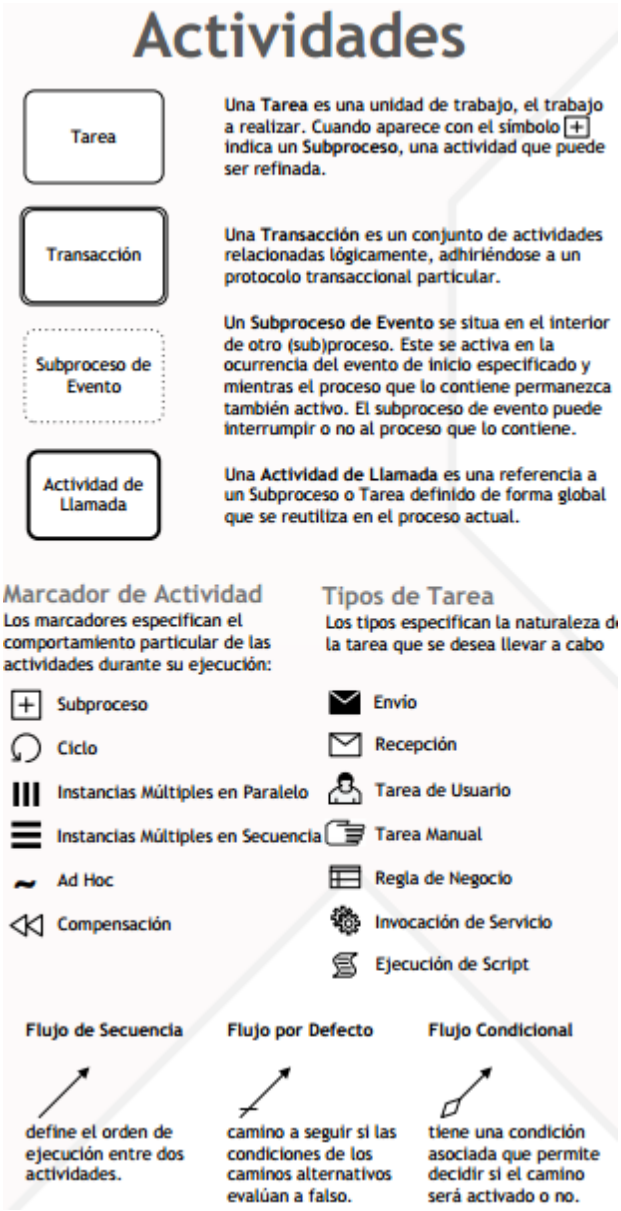

Figura 42: Actividades Notación BPMN
Fuente: BPMN Poster [22]

En la Figura 43 se aprecia la definición de las compuertas utilizadas en notación BPMN. Éstas funcionan como punto de bifurcación, o bien direccionan la ocurrencia de ciertos eventos.

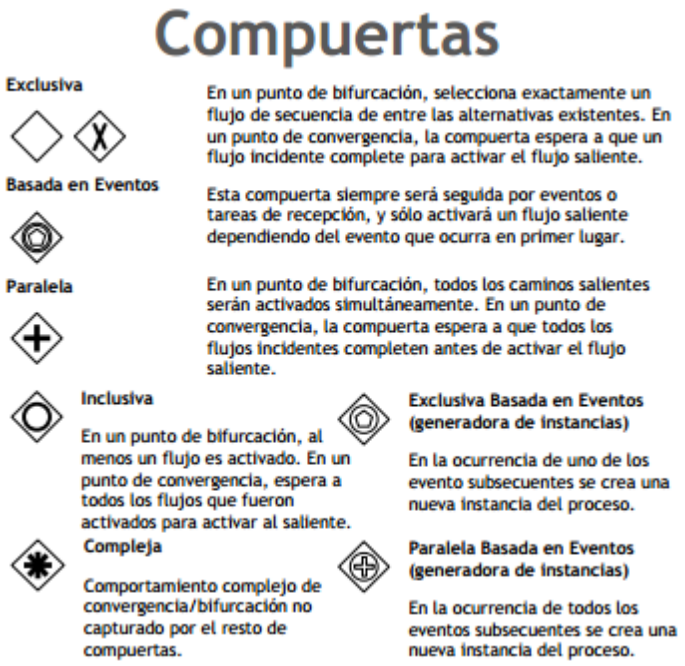

Figura 43: Compuertas en Notación BPMN.
Fuente: BPMN Poster [22]

Figura 44: Contenedores Notación BPMN
Fuente: BPMN Poster [22]

En la Figura 44 se aprecia la explicación de los contenedores en la notación BPMN. Éstos sirven para representar a las entidades responsables de las actividades de un proceso.

Los eventos son caracterizados en la Figura 45. Estos eventos pueden incluir la generación de mensajes, eventos temporales, eventos que capturan errores u otras condiciones pre-definidas, entre otros.

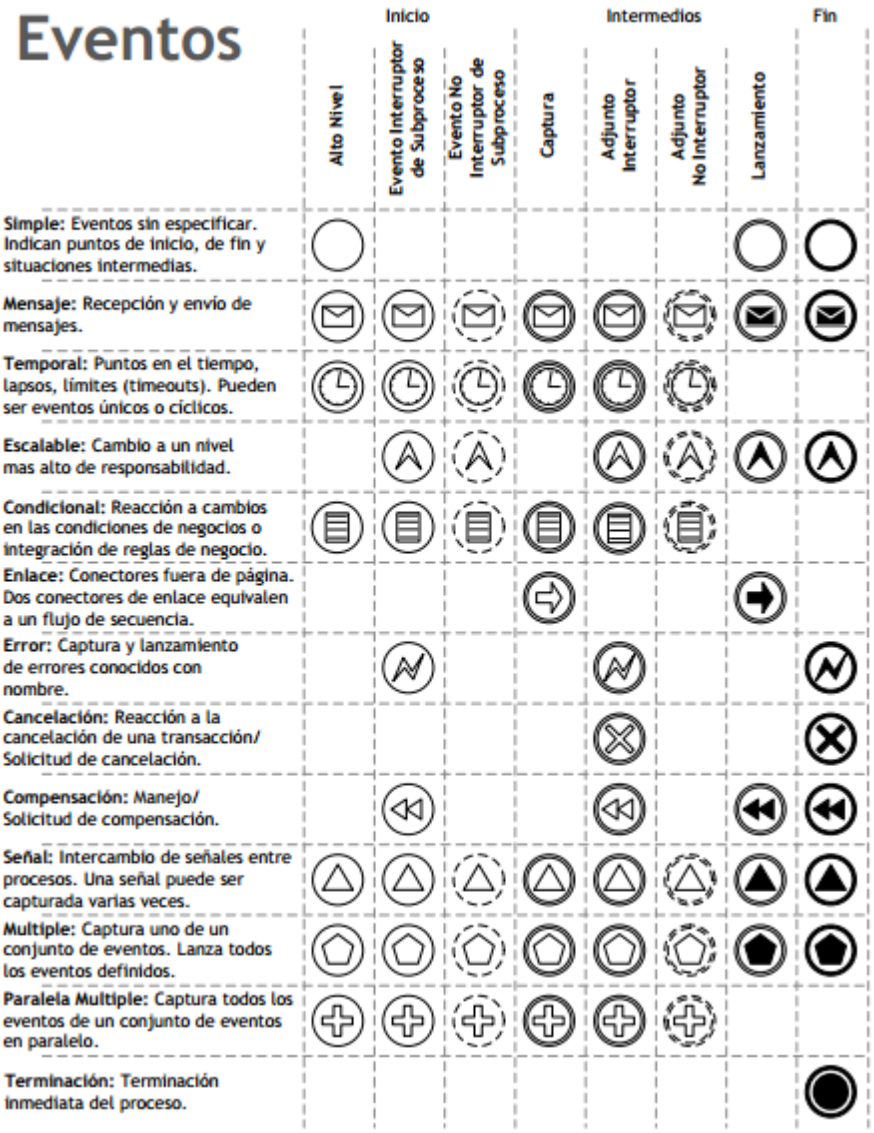

Figura 45: Eventos Notación BPMN
Fuente: BPMN Poster [22]

ANEXO B: PROCESOS ETL

En la Figura 46 se muestra un proceso ETL para la carga de periodos académicos generado en el software pentaho[5] para limpieza y carga de los datos que entran a la calibración de los modelos.

[5] http://www.pentaho.com/

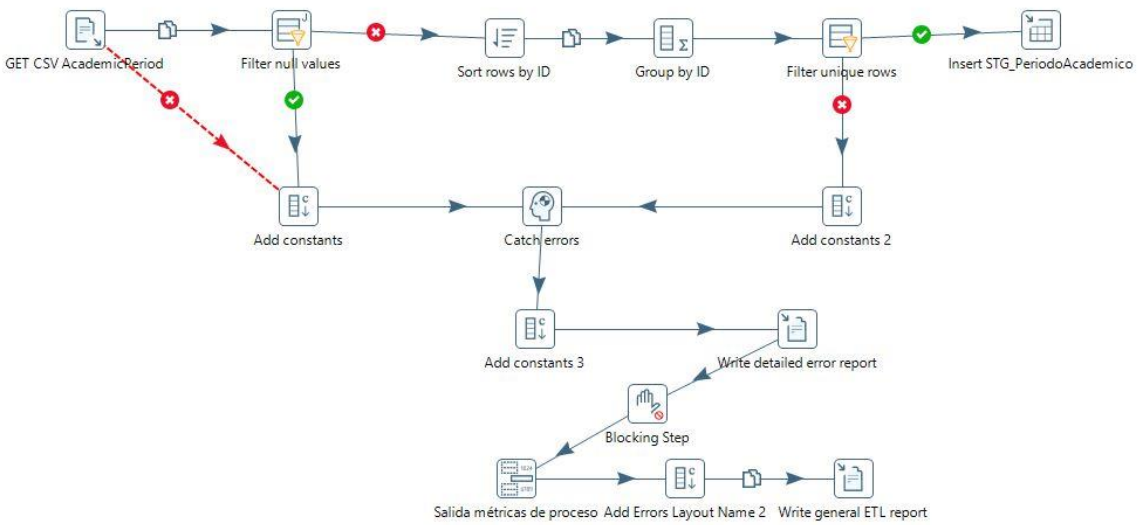

Figura 46: Ejemplo de Proceso ETL para carga de Periodos Académicos

En la Figura 47 se muestra otro ejemplo de proceso ETL para la carga de alumnos e inscripciones de alumnos, en donde se valida consistencia, duplicidad de registros y columnas mínimas que debiese tener los datos y que genera un reporte de carga.

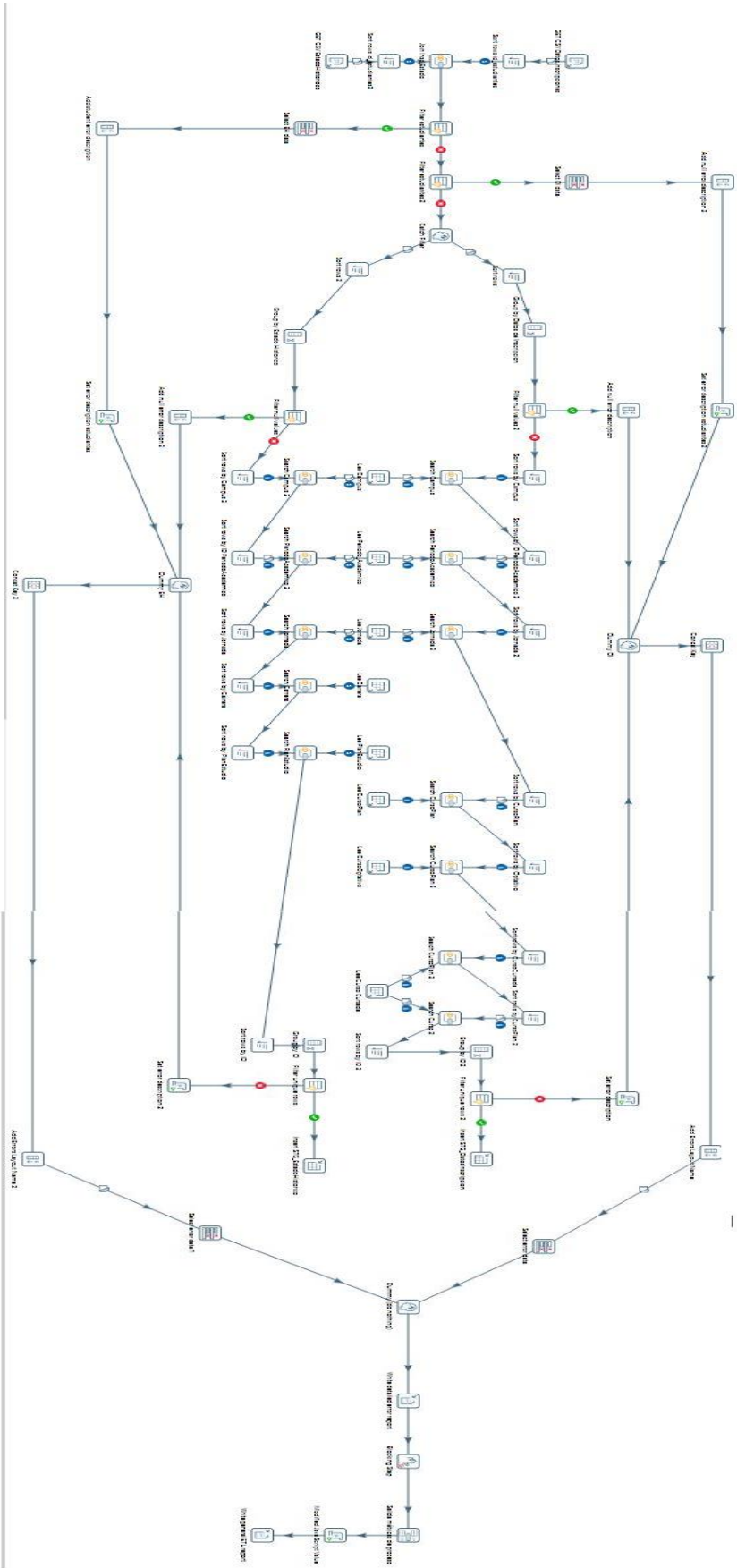

Figura 47: ETL de carga de alumnos e inscripciones

ANEXO C: PROCESO DE CALIBRACIÓN DE MODELOS

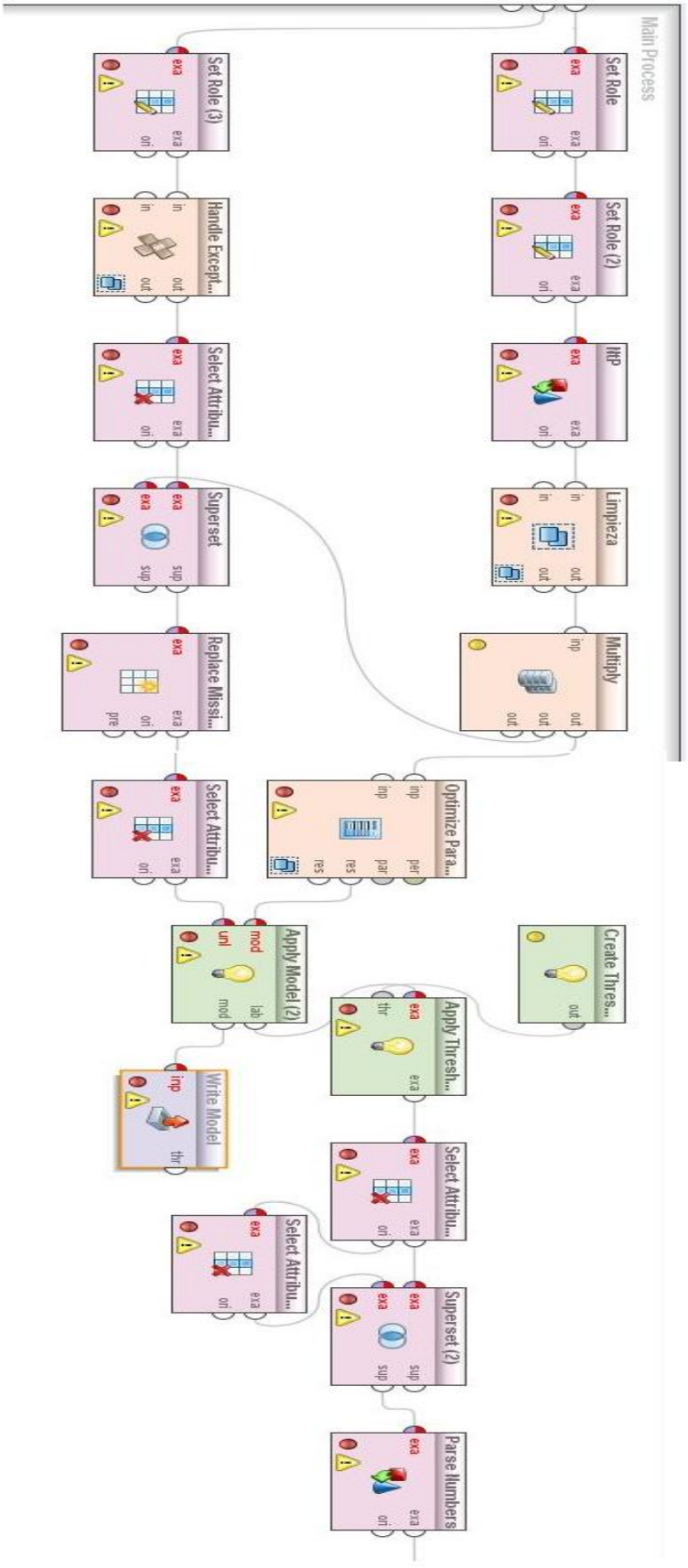

Figura 48: Proceso de calibración de modelos

En la Figura 48 se muestra el proceso creado en Rapidminer[6] para realizar la calibración de los modelos en donde se entrenan los datos y luego se aplica el modelo para generar una predicción en datos conocidos.

[6] https://rapidminer.com/