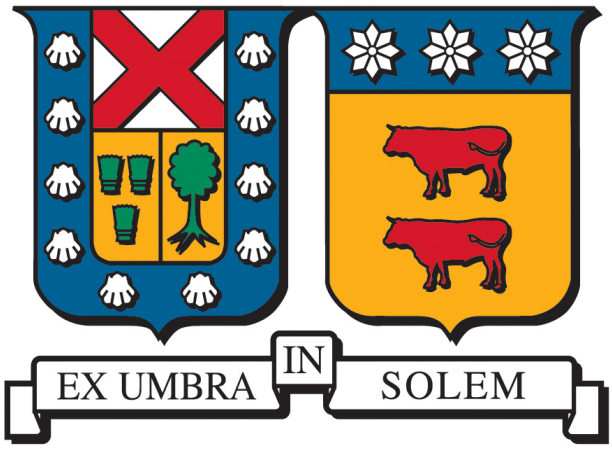
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE



"CONSTRUCCIÓN DE MODELO DE *FORECAST* PARA ESTIMACIÓN DE DEMANDA EN UNA EMPRESA MULTINACIONAL DE RETAIL"

ALFREDO IGNACIO CÉSPEDES URRUTIA

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: **JOSÉ LUIS MARTÍ LARA**

PROFESOR CORREFERENTE: **CECILIA REYES COVARRUBIAS**

NOVIEMBRE - 2017

Agradecimientos

Quiero agradecer a mi familia que fue un gran apoyo en todo sentido durante toda esta etapa, por el esfuerzo y darme la oportunidad de estudiar, ayudándome siempre con lo que pudieron. A mi polola, que sin su apoyo nada hubiera sido posible, todos esos informes redactados por ella e incluso aprender a programar para ayudarme en las cosas imposibles que pedían en la universidad, ella tiene una mención en informática después de todo este tiempo. También a su familia por ayudarme en todo lo que necesité y su preocupación cuando me veían que pasaba por situaciones difíciles. A mi profesor guía por su paciencia y buena disposición y a todos quienes de alguna u otra forma estuvieron involucrados en este largo e importante proceso universitario.

A mi hermana para que también logre con éxito esta etapa de su vida.

Resumen

El trabajo desarrollado tiene como objetivo encontrar un modelo para mejorar el *forecast* de una multinacional, creando una herramienta para poder predecir la demanda de productos. Para lo anterior, se utilizó la herramienta Azure ML junto con el lenguaje de programación R, con los que se usaron las funciones de red neuronal artificial para la predicción. Los resultados obtenidos fueron satisfactorios, según las pruebas realizadas en base a métricas de evaluación tradicionales.
Palabras Clave: BI, ANN, Forecast, R.

Abstract

The work developed aims to find a model to improve the textit forecast of a multinational, creating a tool to predict the demand for products. For this, the Azure ML tool was used together with the programming language R, with which artificial neural network functions were used for prediction. The results obtained were satisfactory, according to the tests carried out based on traditional evaluation metrics.
Keywords: BI, ANN, Forecast, R.

Tabla de Contenidos

Índice de Figuras

Índice de Tablas

Introducción

Mars es una empresa multinacional de origen estadounidense creada en el año 1911 por Frank C. Mars en Tacoma, Washington, estableciéndose como empresa productora de dulces. Hoy, con base en McLean, Virginia, Mars tiene ventas netas de más de 33 mil millones de dólares y seis segmentos comerciales: Petcare, Chocolate, Wrigley, Alimentos, Bebidas y Symbioscience con más de 72.000 empleados a lo largo del mundo.

En Chile, Mars es una unidad de negocio que no cuenta con plantas de producción, por lo que funciona como unidad de ventas importando sus productos desde plantas de fabricación en Argentina, México y Estados Unidos. Basándose en este modelo, una de las áreas de trabajo en que más tiempo y recursos están destinados en la compañía es en la estimación de la demanda de productos y venta futura de sus clientes.

Sin embargo, al día de hoy Mars Chile tiene grandes oportunidades de mejorar el cómo se ha estimado la demanda de los últimos años, presenciando grandes desviaciones frente al plan comercial y teniendo grandes consecuencias tanto monetarias como no monetarias. Por ende, el objetivo de este trabajo es establecer un modelo que permita a Mars Chile, predecir la demanda futura de sus productos y por consiguiente, mejorar su nivel de *forecast accuracy*. Este indicador, es uno de los indicadores más importantes de la compañía, el cual permitirá destinar correctamente el tiempo y recursos, y facilitar el cumplimiento de los objetivos de la compañía.

El documento a continuación, presenta el contexto de la compañía, la definición del problema y el objetivo principal a desarrollar. Como soporte, se investigará sobre temas relevantes que ayudarán a entregar una visión global y los primeros lineamientos que ayuden a alcanzar el objetivo propuesto.

En el desarrollo, se trabajará con la situación actual de la compañía, información disponible y la preparación de los datos que darán paso posteriormente a la etapa de desarrollo y validación del modelo de *forecast*. Esta etapa permitirá seleccionar la téc-

nica de modelado para construir el modelo y finalmente evaluarlos con los datos entregados por la compañía, permitiendo determinar la viabilidad del modelo construido y de cuán útil será en el futuro el uso de este modelo de predicción.

1. Definición del problema

En este capítulo, se abordan la descripción de la compañía con el fin de entender el negocio desde una visión global, el problema que presenta hoy y los objetivos propuestos para dar solución al problema.

1.1. Descripción de la empresa

Mars es una empresa multinacional de origen estadounidense creada en el año 1911 por Frank C. Mars en Tacoma, Washington, estableciéndose como empresa productora de dulces. Hoy, con base en McLean, Virginia, Mars cuenta con más de 72.000 empleados a lo largo del mundo, más de 20 plantas productoras, presencia en más de 78 países y ventas netas de más de 33 mil millones de dólares en sus cinco segmentos comerciales:

- Mars Chocolate
- Mars Petcare
- Mars Wrigley
- Mars Foods
- Mars Drinks

En Chile, Mars funciona como unidad importadora desde sus plantas de Argentina, México y Estados Unidos, comercializando 2 de sus 5 unidades de negocio: Mars Chocolate y Mars Petcare. Ambos segmentos de negocio cuentan con importantes marcas dentro del mercado, las cuales son:

- Mars Chocolate: Snickers®, M&Ms®, Milky Way®y Twix®son marcas mundialmente conocidas y si bien son del gusto de todas las personas, el segmento específico es para aquellas personas entre 14-29 años de edad.

- Mars Petcare: Pedrigee® y Whiskas® como productos destinados a retail (Supermercados y Mayoristas) que atienden preferentemente familias con ingreso

medio y que cuentan con mascotas que no necesitan un cuidado especial. Por otra parte, están Eukanuba®, Royal Canin® y Iams® como productos destinados a veterinarias, clínicas veterinarias y tiendas exclusivas de mascotas, para familias con ingresos altos, que quieran comprar alimento de la mejor calidad para sus mascotas y para aquellas mascotas que necesitan de un cuidado especial por alguna necesidad o enfermedad en particular.

1.2. Problema Detectado

Dado que Mars importa sus productos desde las tres plantas nombradas anteriormente, el proceso de solicitud de producto a dichas plantas puede variar dependiendo del volumen, distancia, aduanas y costos involucrados. ¿Qué se quiere decir con esto? No es lo mismo traer 1.000 toneladas de un producto desde Estados Unidos que hacerlo desde Argentina.

En términos generales, el tiempo estimado para productos importados desde Estados Unidos y México es de 2 a 3 meses, mientras que desde Argentina, es de 1 a 1.5 meses como máximo.

En términos más específicos, se tiene lo siguiente:

- Mars Chocolate: los productos de este segmento de negocio tienen una vida útil de 12 meses, y su importación es desde Estados Unidos y México. Por esto, se puede decir que los productos llegan a Chile con una vida útil de 8-9 meses. En la figura 1.1 se muestra la cadena de suministro para Mars Chocolate desde que el producto es producido en la planta hasta que es entregado al cliente.
- Mars Petcare: los productos de este segmento de negocio tienen una vida útil de 18 meses, y su importación es desde Argentina. Así, se puede decir que los productos llegan a Chile con más de 12 meses de vida útil. En la figura 1.2 se presenta la cadena de suministro para Mars Petcare desde que el producto es producido en la planta hasta que es entregado al cliente.

Figura 1.1: Cadena de suministros de Mars Chocolate.

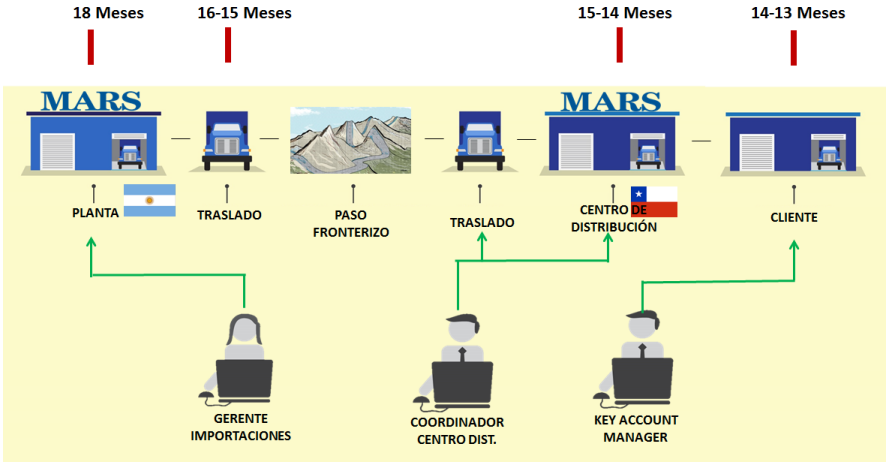


Figura 1.2: Cadena de suministros de mascotas.

Con lo anterior se podría pensar que hay tiempo suficiente para poder comercializar los productos sin problema por temas de vida útil. Sin embargo, los canales de comercialización (supermercados y mayoristas) aceptan los productos con un máximo de 6 meses de vida útil. Otro canal de comercialización son las tiendas tradicionales que aceptan los productos con un tiempo menor, pero que no tiene gran relevancia en el negocio de esta compañía.

Aquí nace el primer problema: los productos del segmento de Mars Chocolate sólo cuentan con un rango de 2 a 3 meses para ser comercializados, porque llegan con 8-9 meses de vida útil. Esto quiere decir que si se tiene un *forecast accuracy* fuera del objetivo que actualmente tiene la compañía y que es de un 70 %, todo el *stock* que se

importó en esa orden, pasará a un estado conocido como *prox vence*. A este problema, se pueden sumar algunas externalidades negativas que pueden influir en el *lead time* de los productos, como por ejemplo: problemas en paso fronterizo por mal clima, paro aduanero, inconvenientes en resolución sanitaria por cambios o nuevas leyes en curso (recientemente Ley 20.606 sobre etiquetado y publicidad de alimentos), entre otros.

Un segundo problema, es la gran variabilidad de la demanda que presenta hoy Mars Chile. Esta demanda se caracteriza por no ser regular, por tener periodos de mucho sobre *stock*, otros de muchos quiebres de *stock* e irregularidades por cambios en el mercado. Por lo anterior, se debe contar con un inventario de seguridad, el cual permita tener inventario disponible por cualquier eventualidad.

Hoy, el *forecast accuracy* de Mars Chile está en torno a un 55 % - 60 %, lo que significa 15 puntos porcentuales por debajo del objetivo mínimo. De lo anterior, se identifican importantes repercusiones como:

- Ineficiencias logísticas.
- Mayor gasto de bodegaje por sobre stock de producto.
- Costo adicional de liquidaciones.
- Devaluación del producto dado los grandes volúmenes de productos a liquidar.
- Pérdida de productos y de venta.
- Pérdida de participación de mercado al no tener productos disponibles para la venta.
- Poca credibilidad como compañía local hacia el resto de las unidades internacionales.
- Desgaste extremo por todas las áreas involucrados para armar un plan que permita vender estos productos, normalmente mediante un plan de liquidación.

En términos concretos, al cierre del año 2016, la compañía tiene almacenado más de un mes de inventario en productos próximos a vencimiento, sin contar todos los productos que ya se perdieron durante el año por vencimiento, todas las liquidaciones que llevaron a cabo para comercializarlos en tiendas tradicionales o "liquidadores", ni

tampoco todos los quiebres que existieron por pérdida de venta, en el minuto en que un producto específico pasa a estar próximo a vencimiento y no hay disponible uno "fresco"que pueda venderse.

Como se puede ver, el problema de *forecast* para Mars Chile es un tema relevante que involucra y afecta el trabajo de todas las áreas, no sólo impactando en los principales indicadores de venta (crecimiento total y por canal), sino también en los indicadores financieros (descuentos principalmente), indicadores logísticos (*fill rate,* sobre *stock*, distribución y entrega) y en el desgaste en el trabajo diario de todos aquellos involucrados en estas áreas.

1.3. Objetivos

Los objetivos van de la mano a solucionar el problema planteado anteriormente, incluyendo algunos objetivos específicos que ayudan a mejorar la forma en que trabaja la empresa.

1.3.1. Objetivo Principal

El objetivo de negocio para Mars Chile es mejorar la manera en que se predice la demanda, teniendo en cuenta la historia de venta de cada producto en cada cliente, donde se mejorará o mantendrá el indicador de *forecast accuracy* con el cual la compañía trabaja: un mínimo de 70 %.

1.3.2. Objetivos Específicos

- Conocer los procesos de la compañía y su funcionalidad, mediante la recolección de datos, comprensión y evaluación, con el fin de tener una base sólida para la construcción de un modelo de predicción y entendimiento de los resultados de dicho modelo.
- Investigar y revisar la teoría de redes neuronales para minería de datos, con el fin

de construir un modelo que ayude a predecir la demanda futura de la compañía, que tenga un mínimo de un 70 % de *forecast accuracy*.

- Utilizar el lenguaje de programación R como soporte al estudio estadístico de los datos, por ser una herramienta utilizada por la compañía y ser un *software* libre.

1.4. Alcance y Limitaciones

- **Alcance:** el alcance de este proyecto es la obtención de una herramienta de predicción de demanda que permita al área de ventas de la compañía cumplir con su estándar de exactitud de predicción para la comercialización de sus productos y la reducción de sus pérdidas por vencimiento y mal manejo de inventarios.

  Por consiguiente de lo anterior, esta exactitud dará como resultado optimizar el proceso de pedidos de venta, manejo de inventarios y gastos asociados a la distribución de los productos.

  En una primera instancia, el alcance se limita a esta área en específico pero que con el desarrollo de las pruebas y obtención de resultados, podría adaptarse a las necesidades de otras áreas de la compañía.
- **Limitaciones:** las limitaciones con las cuales se encuentra el proyecto son las siguientes:
  - El estudio se realizó con los datos disponibles desde el año 2014 a 2016, debido a que anteriormente a eso, la empresa contaba con otras fuentes de almacenamiento de los datos que no aseguran su confiabilidad.
  - Dada la complejidad de los datos y de la cantidad de ellos, se trabajó con los clientes y productos más importantes, y con el total de ellos con el fin de obtener un resultado total.
  - Dado el acuerdo de confidencialidad de la compañía los datos no representan la realidad de la compañía, sino más bien están mostrados a escala, lo que no influye en los resultados del modelo, pero sí en los valores reales de

la compañía.

2. Marco Teórico

El marco teórico es un trabajo necesario para poder entender en conjunto el desarrollo de este trabajo, haciendo una descripción detallada sobre los temas necesarios que conllevarán un mejor entendimiento y lectura de lo que se desarrollará.

Para contextualizar, los temas a tratar en el desarrollo de este trabajo son:

- Estadística
- Inteligencia de negocios
- Redes neuronales artificiales
- Series de tiempo
- Metodologías para proyectos de minería de datos
- *Machine learning* o aprendizaje automático

2.1. Estadística

La estadística es el estudio sistemático de los datos, transformándolos en información relevante que genera conocimiento. Dentro de sus usos sirve para poder organizar, clasificar, resumir y obtener información de los datos [1]. A continuación se describen los principales y más importantes conceptos:

- **Población:** es un conjunto de elementos que poseen alguna regla de pertenencia definida por el observador, de los cuales se quiere conocer sus características o comportamientos.

  Esta población puede ser:

  - **Finita:** si los elementos pertenecientes a la población son finitos.
  - **Infinita:** si los elementos son infinitos.

  Lo anterior es relevante ya que muchas veces se trabaja con una muestra de la población, pues a veces es imposible obtener la información de toda una población, a lo que se le denomina viabilidad. Otro motivo por el cual se utiliza una muestra es la reducción de costos monetarios y no monetarios.

- **Variable:** es una característica que puede cambiar de una población a otra. En general, cuando se tiene una sola característica de un objeto, los datos se denominan como dato univariado, y al tener dos o más datos sobre el mismo, se denominan datos multivariados.
- **Estadística descriptiva:** la estadística descriptiva tiene métodos para resumir y describir características, dentro de los cuales se encuentran los cálculos numéricos como media, moda, mediana y desviación estándar. Esta es la primera etapa a desarrollar en cualquier trabajo de investigación y de análisis de información. Los métodos usados en esta investigación son:
  - **La media muestral:** es simplemente un promedio numérico. Dados los n números $x_1, x_2, ..., x_n$, la media se define como:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + ... + x_n}{n} \tag{1}$$

  - **La mediana muestral:** se obtiene ordenando todas las observaciones de menor a mayor. Su propósito es reflejar la tendencia central sin que influyan los valores extremos. Su fórmula es:

$$\widetilde{x} = \begin{cases} x_{(n+1/2)} & \text{si } n \text{ es impar} \\ x_{n/2} + x_{(n/2)+1} & \text{si } n \text{ es par} \end{cases} \tag{2}$$

  - **Cuartil:** son los valores que dividen en cuatro conjuntos porcentualmente iguales los datos. Esto sirve para encerrar la mediana dentro, con los rangos extremos percentil 25 (cuartil inferior) y percentil 75 (cuartil superior).

$$Q_1 = \frac{n+1}{4} \tag{3}$$

$$Q_3 = \frac{3(n+1)}{4} \tag{4}$$

  - **Rango intercuartil:** es la diferencia entre el tercer y el primer cuartil (antes

mencionados). Este valor se usa para construir un *Boxplot* que se define más adelante.

$$R_Q = Q_3 - Q_1 \tag{5}$$

- **ME**: llamado así por su sigla en inglés (Error Medio). Indica el promedio de los errores de predicción (se considera como error la diferencia entre el valor real y el valor predicho):

$$ME = \frac{\sum_{t=1}^{n} R_t - A_t}{n} \tag{6}$$

donde n es el número de observaciones, R_t es el valor real y A_t el valor ajustado del pronóstico.

- **RMSE**: llamado así por su sigla en inglés (Raíz Cuadrada de Error Medio). Es la raíz cuadrada del promedio de los errores al cuadrado:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (R_t - A_t)^2}{n}} \tag{7}$$

donde n es el número de observaciones, R_t es el valor real y A_t el valor ajustado del pronóstico.

- **MAE**: llamado así por su sigla en inglés (Error Absoluto Medio). Es el promedio del valor absoluto de los errores:

$$MAE = \frac{\sum_{t=1}^{n} |R_t - A_t|}{n} \tag{8}$$

donde n es el número de observaciones, R_t es el valor real y A_t el valor ajustado del pronóstico.

- **MPE**: llamado así por su sigla en inglés (Error Porcentual Medio). Es el

promedio porcentual de los errores.

$$MPE = \frac{100}{n} \sum_{t=1}^{n} \frac{R_t - A_t}{R_t} \tag{9}$$

donde n es el número de observaciones, R_t es el valor real y A_t el valor ajustado del pronoóstico.

- **MAPE**: llamado así por su sigla en inglés (Error Porcentual Absoluto Medio). Es el promedio porcentual de los valores absolutos de los errores:

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{R_t - A_t}{R_t} \right| \tag{10}$$

donde n es el número de observaciones, R_t es el valor real y A_t el valor ajustado del pronóstico.

- ***Forecast Accuracy***: es la precisión de la predicción en porcentaje, expresada por la fórmula:

$$FA = 1 - MAPE \tag{11}$$

- ***Boxplot***: esta herramienta gráfica sirve para visualizar la variabilidad de una variable y comparar la distribución de la misma; además sirve para ubicar los extremos de los datos o *outliers*. La figura 2.1 muestra ejemplos de gráficos.

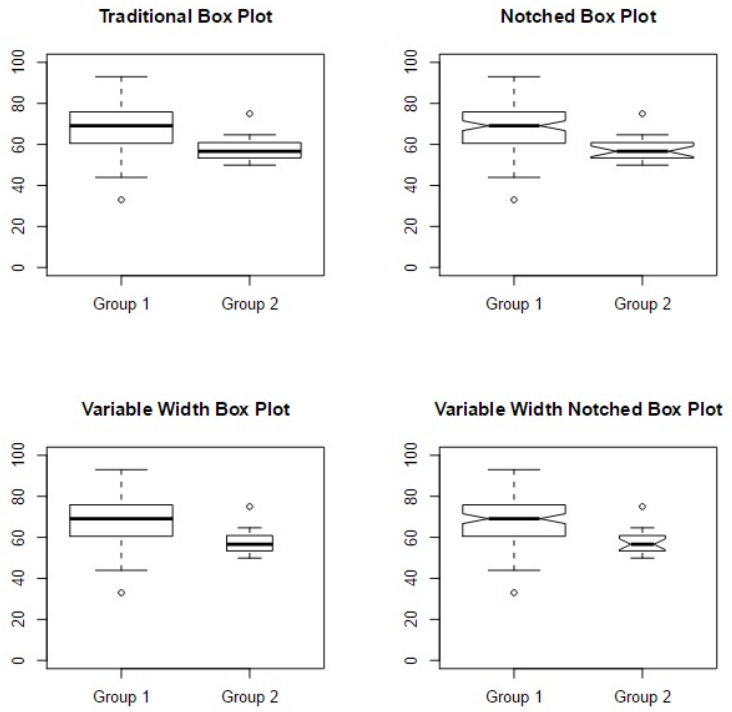


Figura 2.1: Ejemplo de gráfico de Boxplot [1]

Para construir un gráfico de este tipo es necesario realizar algunos cálculos que ayudan a obtener los distintos puntos necesarios en su construcción, pero que para este informe sólo basta con calcular los límites superiores e inferiores que se usarán posteriormente. Para esto, se utilizan las siguientes expresiones:

$$Lim_{inf} = Q_1 - 1,5R_Q \tag{12}$$

$$Lim_{sup} = Q_3 + 1,5R_Q \tag{13}$$

- ***Outlier***: denominado también valor atípico, es un valor numéricamente distinto al resto de los datos. Hay que mencionar que lleva trabajo entender cómo se comportan estos valores, dado que en ocasiones pueden entregar información valiosa del comportamiento de los datos y en otras, simplemente es mejor desecharlos. Para saber si un dato es un *outlier*, estos se

obtienen gracias a la siguientes expresiones:

$$Outlier < Q_1 - 1,5R_Q \tag{14}$$

$$Outlier > Q_3 + 1,5R_Q \tag{15}$$

A continuación, en la figura 2.2, se puede ver un ejemplo de *outlier* en un gráfico *boxplot*:

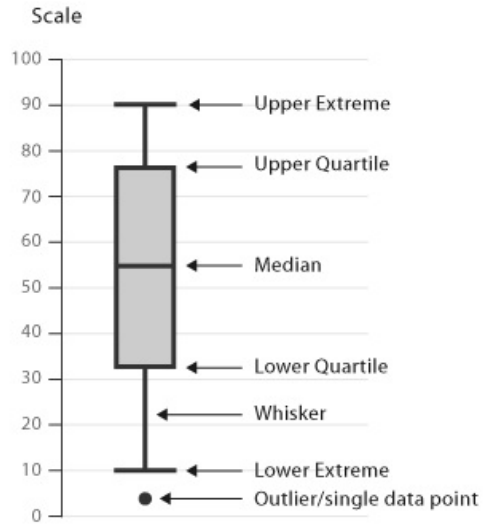


Figura 2.2: Ejemplo de Outlier en un gráfico Boxplot [1]

2.2. Inteligencia de negocios

Es un área que en diferentes tipos de organizaciones a la hora de tomar decisiones, ayuda de forma precisa y oportuna. Esta área garantiza que exista el conocimiento necesario que les permita escoger la alternativa que sea más conveniente para el éxito de la empresa. En los puntos siguientes se verán los temas relacionados a su aplicación y trabajos relevantes, con algunas herramientas que permiten hacer *business intelligence* (BI).

Hoy en día, la mayoría de las organizaciones poseen un sistema de información

que soporta gran parte de las actividades diarias del área en donde se desempeña. Este sistema puede ser sencillo o robusto, y con el paso del tiempo estas aplicaciones llegan a tener la historia de la organización almacenadas en una base de datos, y que pueden ser utilizadas como base para la toma de decisiones. El poder competitivo que puede tener una empresa se basa en la calidad y cantidad de la información que sea capaz de usar para la toma de decisiones. Mediante la implementación de BI se proporcionan las herramientas necesarias para aprovechar los datos almacenados en las bases de datos de los sistemas y generar conocimiento como respaldo a las decisiones, reduciendo el efecto negativo que puede traer consigo una mala decisión.

BI se define como la habilidad corporativa para tomar decisiones [3]. Esta habilidad se logra mediante el uso de metodologías, aplicaciones y tecnologías que permiten reunir, depurar, transformar datos, y aplicar en ellos técnicas analíticas de extracción de conocimiento. A continuación una breve descripción de las herramientas más utilizadas:

- ***Data Warehousing***: es el proceso de extraer datos de distintas aplicaciones, para que una vez depurados y estructurados, sean almacenados en una base de datos consolidada para el análisis del negocio.
- **OLAP**: es el procesamiento analítico en línea que permite obtener acceso a datos organizados y agregados de orígenes de datos empresariales. Además, organiza subconjuntos de datos con una estructura multidimensional de manera que represente un significado especial o responda a una pregunta en particular.
- ***Balanced Scorecard***: es una herramienta que permite alinear los objetivos de las diferentes áreas o unidades con la estrategia de la empresa y seguir su evolución. La ejecución de esta herramienta es tan amplia que puede llegar a cambiar la forma en que se presta un servicio o se comercializan productos.
- **Minería de datos**: es el proceso de seleccionar, explorar, modificar, modelar y valorar grandes cantidades de datos con el objetivo de descubrir conocimiento. El proceso debe ser semiautomático y los modelos hallados deben ser significativos

demostrando cierto patrón o regla de comportamiento.

La Inteligencia de Negocios da una visión actual e histórica del negocio, como también permite poder predecir el futuro. Para estas formas de visualizar la información existen diversos modelos que contribuyen en la búsqueda de lograr estos resultados [3]:

- **Predictivo:** analiza el histórico de la información existente para poder ver su comportamiento en el futuro.
- **Descriptivo:** ve la relación del conjunto de la información existente y clasifica en distintos grupos. Se enfoca en las relaciones entre clientes.
- **Decisión:** describe la relación entre todos los elementos de una decisión para poder predecir los resultados de las decisiones que involucran varias variables.

2.3. Redes neuronales artificiales

Las redes neuronales artificiales (RNA) nacen de la idea de que estos sistemas permiten imitar el funcionamiento del cerebro, el conocimiento y la percepción de las personas, logrando crear sistemas que tuvieran mejores resultados en varias tareas, como clasificación, pronóstico, entre otros [4].

Varios estudios demuestran que existen RNA en el cerebro humano y que aplicadas a diversas tareas tienen resultados exitosos. Las RNA poseen un tiempo de proceso de $O(10^{-13})$ mientras que los computadores tienen un tiempo de $O(10^{-9})$; haciendo que el procesador de un computador sea mucho más rápido en procesar que una neurona, con un factor de $O(10^{6})$ [4].

Las RNA tienen como objetivo poder ser usadas en aplicaciones reales, siendo las principales a emular [4]:

- Cálculo en paralelo.
- Memoria distribuida.

2.3.1. Componentes de una RNA

Los componentes generales de una RNA se ven reflejados en la figura 2.3:

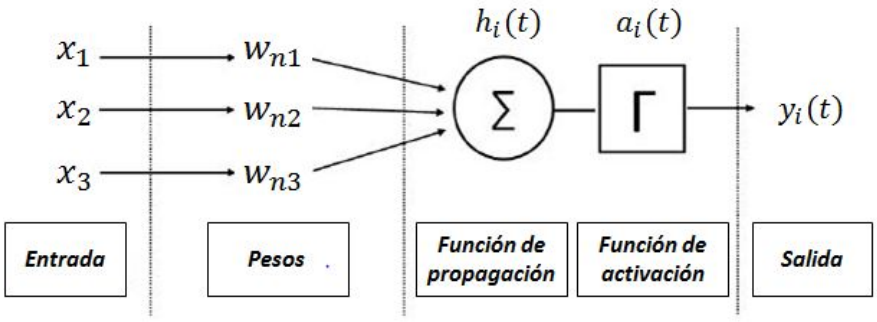


Figura 2.3: Ejemplo de una Red Neuronal Artificial (RNA)

A continuación se detallan cada uno de estos componentes[4][5]:

- **Entrada**: también llamado input x_j de la neurona, la que corresponde a los datos de entrada.
- **Pesos**: al recibir múltiples entradas en la neurona, éstas filtran gracias al peso de las entradas w_{nj}, ayudando a conocer la importancia y el efecto sobre el procesamiento de la neurona. Estos pesos pueden ser adaptados en la red para determinar distintas intensidades en la entrada a las neuronas artificiales. La forma en que normalmente se adaptan es el proceso de entrenamiento de la neurona, el que se verá mas adelante.
- **Función de propagación**: esta función tiene como objetivo obtener a partir de una entrada y un peso, el valor del potencial postsináptico h_i de la neurona. Esta función puede ser combinada de muchas formas, como por ejemplo usar el mínimo, máximo o diversos algoritmos de normalización. El algoritmo a usar va de la mano con la arquitectura elegida; dentro de las funciones más comunes se encuentra la suma ponderada de todas las entradas con sus respectivos pesos:

$$h_i(t) = \sum_j w_{ij} * x_j. \tag{16}$$

- **Función de activación**: es la salida real de la neurona desde la función de propagación, siendo de la forma:

$$a_i(t) = f_i(a_i(t-1), h_i(t)) \tag{17}$$

en donde $h_i(t)$ es el potencial postsináptico y es uno de los valores de los que depende esta función junto con los estados de activación anterior. Funciona con un valor umbral, en donde se determina la salida de la neurona. Si la suma es mayor al valor umbral, ésta genera una señal de salida y en el caso que sea menor, no hay reacción. La función de activación puede ser de tipo lineal y no lineal, donde el uso de ellas se limita al criterio de investigación y propósito de las RNA [6]. Ejemplos de algunas de estas funciones se pueden ver a continuación [7][4]:

  - **Función lineal**: en esta función la entrada es igual a la salida. Se utiliza mayormente para problemas de clasificación y es de la forma:

$$f(h_i(t)) = h_i \tag{18}$$

  - **Función logística o sigmoide**: la salida varía entre 0 y 1. Es la función de activación más usada, comúnmente en problemas de regresión:

$$f(h_i(t)) = \frac{1}{1 + e^{-h_i}} \tag{19}$$

  - **Función tangente hiperbólica**: se asemeja a la función logística pero su salida varía entre -1 y 1. Ésta es usada con frecuencia en redes multicapas.

$$f(h_i(t)) = \tanh(h_i) \tag{20}$$

- **Escalamiento**: el valor de salida de la función de activación puede ser procesada adicionalmente por un escalamiento o limitación, multiplicando el valor de la función de activación y luego sumando un desplazamiento.

- **Función de salida**: cada neurona tiene permitido sólo una salida $y_i(t)$, la que normalmente tiene como resultado el mismo valor de la función de activación.

$$y_i(t) = F_i(a_i(t)) = a_i(t) \tag{21}$$

- **Función de error**: en el entrenamiento la mayoría de los algoritmos necesitan calcular el error entre el obtenido y el esperado. Este error es calculado por la función de error.
- **Tasa de aprendizaje**: esta tasa depende de muchos factores controlables. Una tasa baja equivale a un largo entrenamiento para producir RNA bien entrenadas, y por el contrario, una tasa alta da como resultado que la red no sea capaz de discriminar un sistema adecuado. La mayoría de los algoritmos requieren que se le entregue un término, el cual normalmente es un valor positivo entre 0 y 1. El equilibro de la tasa de aprendizaje logra un buen tratamiento de las RNA.

2.3.2. Arquitectura

Se llama arquitectura a la estructura de una RNA. En general, las neuronas se agrupan en unidades que se llaman capas, las que en conjunto conforman una red neuronal. Hay diferentes tipos de capas, las cuales se describen a continuación:

- **Capas de entrada**: neuronas que reciben datos y tienen una conexión directa con el entorno.
- **Capas oculta**: neuronas que no tienen conexión directa con el entorno.
- **Capas de salida**: neuronas que proporcionan la respuesta de la RNA.

En la figura 2.4, se puede ver un ejemplo de arquitectura de una RNA:

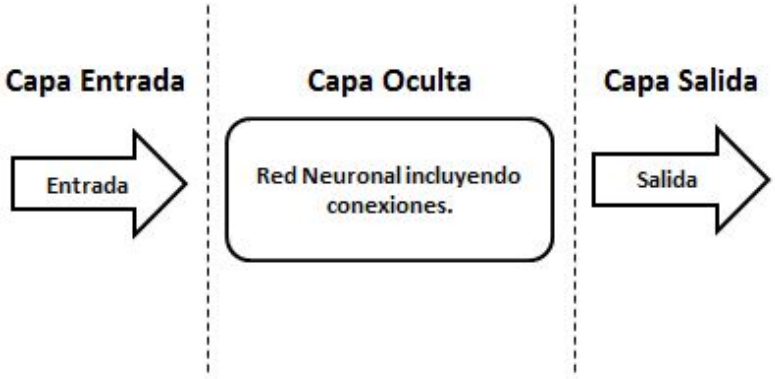


Figura 2.4: Arquitectura de una Red Neuronal Artificial (RNA)

Cada unidad neuronal está conectada con muchas otras y las conexiones entre ellas poseen varias características. una de ellas es que pueden incrementar o inhibir el entorno al cual se conecta. Este proceso puede aplicarse a las neuronas conectadas entre sí a través de las diferentes capas e incluso también en la misma neurona y conexiones entre dos capas diferentes. La capacidad de conocer el flujo de los datos entre conexiones, puede ser unidireccional (*feedforward*) o retroalimentadas (*feedback*). En esta última, la información circula en cualquier sentido.

2.3.3. Modo en que opera una RNA

Las RNA pueden trabajar de dos modos, los cuales son [4]:

- **Modo de entrenamiento:** en el entrenamiento supervisado se estudia la salida, cambiando los pesos de entrada, haciendo que en la salida se esté cerca del resultado deseado. Existe también el entrenamiento no supervisado que aún está en proceso de estudio.
- **Modo de recuerdo:** en general, el entrenamiento se apaga dejando los pesos de entrada y la estructura formada de manera fija, dejando la RNA lista para procesar datos.

2.3.4. Aplicaciones

Las RNA pueden ser utilizadas en un gran número de aplicaciones de diversas índoles. Se pueden desarrollar en un periodo de tiempo razonable con la capacidad de la tecnología de hoy en día, por lo que es una buena herramienta para llevar a cabo tareas en áreas tales como [5]:

- **Biología**: obtención de modelos de la retina.
- **Empresas**: explotación de base de datos; optimización de asientos y horario de líneas de vuelo.
- **Medio ambiente:** analizar tendencias y patrones; previsión del tiempo.
- **Finanzas:** previsión de la evolución de los precios; valoración de los riesgos de créditos.
- **Manufactura:** control de producción en líneas de procesos; inspección de calidad.
- **Medicina:** predicción de reacciones adversas en los medicamentos.
- **Militares:** clasificación de las señales de radar; optimización de recursos.

La mayoría de las aplicaciones consisten en reconocer patrones, por ejemplo patrones de serie de ejemplos, o completar señales a partir de valores parciales o reconstrucción de patrones.

2.4. Series de tiempo

Una serie de tiempo se define como una secuencia de datos estadísticos, medidos a lo largo del tiempo en un intervalo dado, llamados periodos, los cuales con frecuencia son de igual tamaño.

2.4.1. Componentes

En una serie de tiempo existen cuatro tipos básicos de variación de datos [5]:

- **Tendencia secular:** esta tendencia a largo plazo de una serie es el resultado de

un factor a largo plazo. En otros términos, ésta experimenta una tendencia con un patrón gradual propias de la serie. Los factores de esta tendencia pueden ser cambios en la población, demográficos, de ingresos, en la salud, tecnología, etc.

- **Variación estacional:** también llamado componente estacional, representa la variabilidad de los datos en donde las estaciones son las influyentes, lo que ocurre generalmente año tras año en los mismos meses o periodos, con casi la misma intensidad.
- **Variación cíclica:** esta variación se presenta generalmente en secuencias alternas de puntos bajos y altos de la línea de tendencia, que duran más de un año y que ocurre incluso después de eliminar las variaciones estacionales o irregulares. Un claro ejemplo son los ciclos comerciales cuyos periodos depende de la prosperidad o recesión del tiempo en el que se encuentre.
- **Variación irregular:** esta variación se debe a factores a corto plazo, imprevisibles y que no son recurrentes en una serie de tiempo. Dentro de éstas, existen dos tipos de variaciones irregulares: las que son provocadas por un acontecimiento puntual (protesta, elección, inundación) y las variaciones por casualidad, cuyas causas no se pueden señalar de una forma exacta, pero que finalmente no afectan la serie, ya que ésta tiende a equilibrarse.

2.4.2. Tipos de Tendencia

- **Tendencia lineal:** como antes se mencionaba, las tendencias vienen dadas por movimientos generales a largo plazo de una serie. Las tendencias a largo plazo, con frecuencia, se pueden aproximar a una recta, la cual muestra una serie que aumenta o disminuye a un ritmo constante. El método para obtener esta línea recta es conocida como método de mínimos cuadrados [3] .
- **Tendencias no lineales:** cuando la serie de tiempo tiene un comportamiento curvilíneo se dice que posee un comportamiento no lineal. Esta tendencia puede ser representada por distintas curvas, como por ejemplo polinomial, logarítmica o de

tipo exponencial.

2.4.3. Predicción

La predicción del comportamiento de diferentes variables siempre ha sido de interés científico en diversas áreas. Lo que se busca es encontrar un modelo que permita determinar con la mejor precisión posible valores futuros, para uno o más periodos hacia adelante.

En particular, las RNA al poseer características no lineales, son modelos en donde se encuentra una especial utilidad: el perceptor multicapa ha sido utilizado con mayor frecuencia en la predicción de series no lineales, debido a que aproxima cualquier función continua definida en un dominio acotado.

2.5. Metodologías para proyectos de minería de datos

2.5.1. CRISP-DM

Para el desarrollo de proyectos de minería de datos, se utilizan diferentes metodologías para llevarlos a cabo, donde la más utilizada es CRISP-DM.

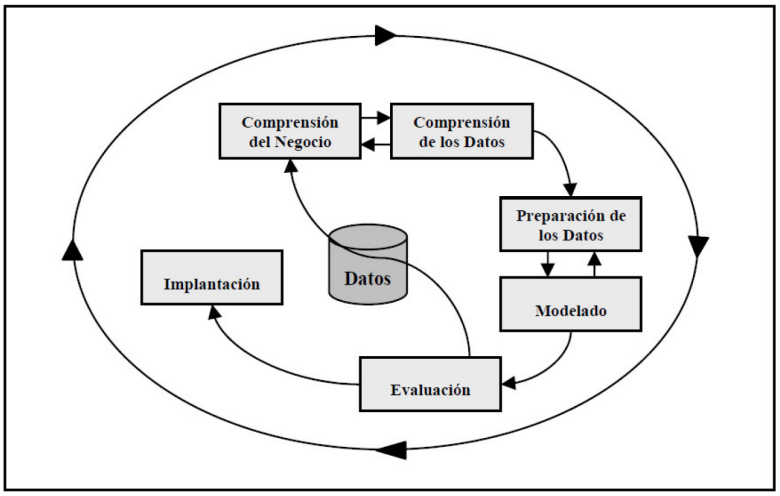


Figura 2.5: Modelo CRISP-DM [2]

CRISP-DM tiene 6 fases, como muestra la figura 2.5, que se detallan a continuación [2]:

- **Comprensión del negocio o problema:** esta etapa se conoce por ser la más importante, ya que comprende los objetivos y los requisitos del proyecto desde la perspectiva de la empresa, y los que luego se convierten en objetivos técnicos para armar un plan de proyecto. Sin esta primera fase, ningún algoritmo permitirá tener resultados confiables y no se obtendrán todos los beneficios que podría potencialmente entregar.
- **Comprensión de los datos:** es la recolección inicial de datos para establecer un primer acercamiento al problema, identificando la calidad de la información con el fin de establecer una hipótesis.
- **Preparación de los datos:** aquí se adaptan los datos para las técnicas, tales como visualización de datos, búsqueda de relaciones entre variables u otros.
- **Modelado:** en esta fase se seleccionan las técnicas de modelado más apropiadas para el proyecto, donde las técnicas se eligen en función de si: son apropiadas para el problema, se dispone de datos adecuados, se cumplen los requisitos del problema, el tiempo es adecuado para obtener el modelo y si hay conocimiento de la técnica empleada. Antes del modelado, se debe determinar un método de evaluación de los modelos para ver los beneficios de ellos. Una vez finalizado esto, se genera el modelo en base a parámetros que dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo.
- **Evaluación:** se evalúa el modelo teniendo en cuenta el cumplimiento de los criterios de éxito del problema y considerando que la fiabilidad para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para repetir algún paso anterior, donde se haya posiblemente cometido algún error.
- **Implementación:** se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea recomendando acciones o aplicando el modelo a

diferentes datos como parte del proceso. En esta etapa, además se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados.

2.5.2. SEMMA

Esta metodología corresponde al acrónimo de sus cinco fases de procesos, *Sample* (Muestreo), *Explore* (Exploración), *Modify* (Modificación), *Model* (Modelado) y *Assess* (Valoración) [8], las que se describen a continuación:

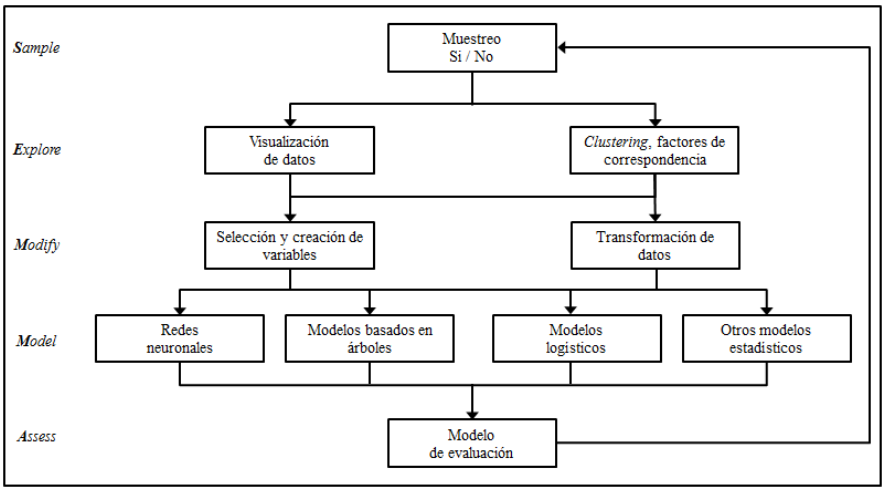


Figura 2.6: Modelo SEMMA

- **Muestreo:** se extrae una muestra representativa del conjunto de datos, donde se realiza un análisis respectivo. El método más común es el denominado "muestreo aleatorio simple", en donde cada elemento tiene la misma probabilidad de ser seleccionado.
- **Exploración:** en esta fase se trabaja con la muestra de los datos seleccionados, donde se realizan diversos análisis estadísticos y se usan herramientas de visualización que ayudan a ver las distintas relaciones entre las variables. Con esto se simplifica el modelo y se optimiza su eficiencia, ayudando a refinar la información en las siguientes fases.
- **Modificación:** esta fase modifica los datos que serán ingresados al modelo para uno que tenga un formato adecuado, mejorando la definición de estos.

- **Modelado:** se procede a modelar el conjunto de datos, haciendo que algún software realice una búsqueda completa de combinaciones de datos que ayuden a predecir los resultados esperados de manera confiable. Entre las técnicas usadas están lógica difusa, árboles de decisión, redes neuronales y computación evolutiva.
- **Valoración:** esta última fase valora los datos obtenidos para determinar el grado de confiabilidad de los mismos, y con esto poder evaluar el modelo, mediante comparaciones con otros métodos estadísticos o con una nueva muestra de datos.

2.5.3. Catalyst

Esta metodología, conocida también como P^3TQ, llamada así por su sigla *Product* (Producto), *Place* (Lugar), *Price* (Precio), *Time* (Tiempo) y *Quantity* (Cantidad) [3], propone dos modelos, uno de negocio y otro de explotación de información:

- **Modelo de negocio:** este brinda una guía de pasos para poder desarrollar y realizar un modelo que permita identificar un problema del negocio con sus requerimientos reales. Si no se encuentra una definición real al problema, se analizan las relaciones P^3TQ que existen en la cadena de valor de la organización. Este modelado presenta diferentes escenarios dependiendo en el cual se encuentre la organización con el fin de tomar las acciones pertinentes para implementar un proyecto. Los escenarios son: dato, oportunidad, prospectiva, definido y estratégico, los cuales se describen a continuación:
  - **Dato:** el proyecto inicia con un conjunto de datos y la premisa es explorar este conjunto para poder buscar relaciones interesantes.
  - **Oportunidad:** el proyecto inicia con un problema u oportunidad que debe ser explorado.
  - **Prospectiva:** el proyecto se diseña para poder descubrir dónde la explotación de la información puede agregar un valor en la organización.
  - **Definido:** el proyecto empieza con la premisa de crear la especificación del

modelo de explotación de datos con un fin específico.

- **Estratégico:** el proyecto comienza con una estrategia de análisis para dar soporte a un escenario planeado por la organización.

- **Modelo de explotación de información:** este modelo da una guía de pasos para poder ejecutar y realizar modelos de explotación de información a partir del modelo de negocio desarrollado anteriormente. Los pasos a seguir, a grandes rasgos, son los siguientes:
  - Preparación de los datos.
  - Selección de herramientas y modelado inicial.
  - Ejecución modelo.
  - Evaluación de los resultados.
  - Comunicación de resultados.

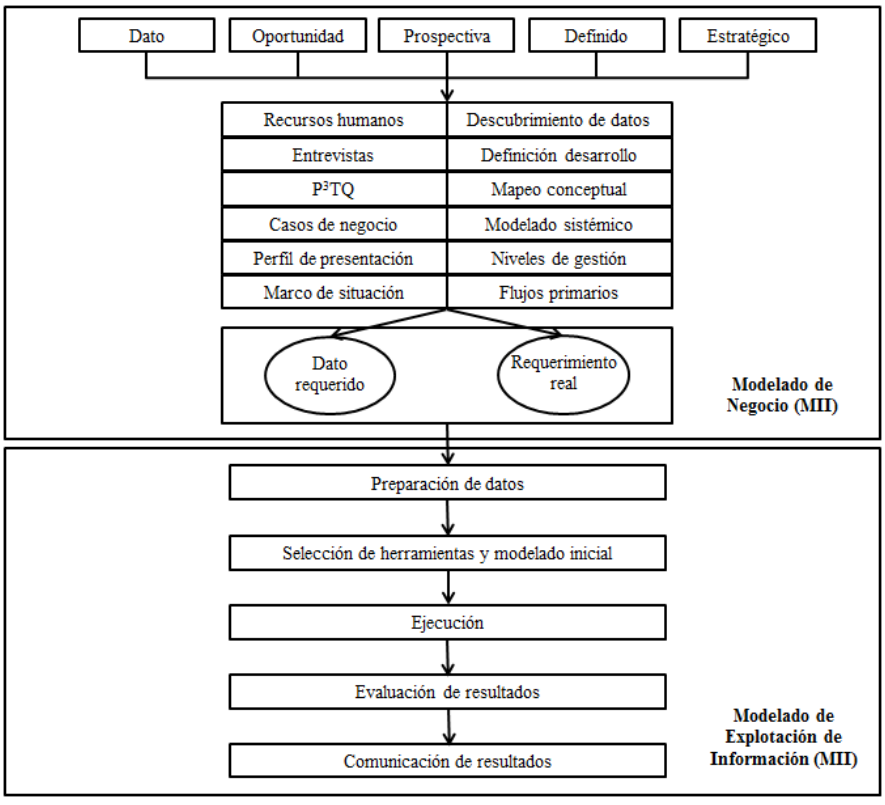


Figura 2.7: Modelo Catalyst

2.6. *Machine Learning* (aprendizaje automático)

Se define como una técnica de la ciencia de datos que ayuda a aprender sobre los datos existentes con el fin de analizar tendencias y comportamientos futuros; esta predicción del aprendizaje ayuda a que las aplicaciones o dispositivos sean más inteligentes. Como ejemplo, se tiene que al comprar *online*, el aprendizaje automático ayuda a reconocer los patrones del comprador y recomienda productos antes ya comprados por compradores de un perfil similar o de la misma índole.

En general, el concepto de aprendizaje automático puede ser un poco difícil de entender, por lo que a continuación se presentan diferentes conceptos que ayudan a un mejor entendimiento de esto:

- **Exploración de datos:** es el proceso de recopilar la información de un gran conjunto de datos, que a menudo no están bien estructurados, con el objetivo de encontrar características de análisis.
- **Aprendizaje supervisado.** se dividen los datos en un conjunto de entrenamiento y otro para su evaluación. Esta división es aleatoria y sirve para poder entrenar el modelo. se entrena el conjunto de entrenamiento y se supervisa la mayoría del aprendizaje automático. Un ejemplo claro es crear un modelo que pueda identificar transacciones fraudulentas desde un conjunto de datos; donde se etiquetan los cargos fraudulentos y cargos válidos conocidos.
- **Aprendizaje no supervisado:** se utiliza en los datos que no poseen etiquetas y el objetivo es buscar alguna relación entre los datos. Un ejemplo de esto es poder agrupar a clientes de un sector con hábitos de compras similares.
- **Datos de aprendizaje:** al entrenar un modelo, éste se hace a partir de datos conocidos y se realizan ajustes necesarios para que el modelo funcione con las características de los datos de una forma más precisa. En Azure, se crean modelos a partir de módulos de algoritmos que procesan los datos de entrenamiento.
- **Datos de evaluación:** con el entrenamiento del modelo, se evalúan los datos restantes de esta división en una prueba y se pueden utilizar datos en que se

conocen sus respuestas para saber la precisión del modelo.

A continuación, se detallan algunas de las herramientas en las cuales es posible implementar la técnica descrita anteriormente:

2.6.1. Microsoft Azure Machine Learning

Microsoft Azure Machine Learning es un servicio de análisis predictivo en la nube que permite crear e implementar modelos predictivos como solución de análisis [9]. Esta herramienta es capaz de trabajar con una biblioteca de algoritmos que están incluidos en ella o de implementar módulos externos de distintos lenguajes de programación como Python o R. La característica que más llama la atención es la capacidad de implementar estos modelos predictivos desde un *web service* listo para ser utilizado gracias a que todo está en la nube.

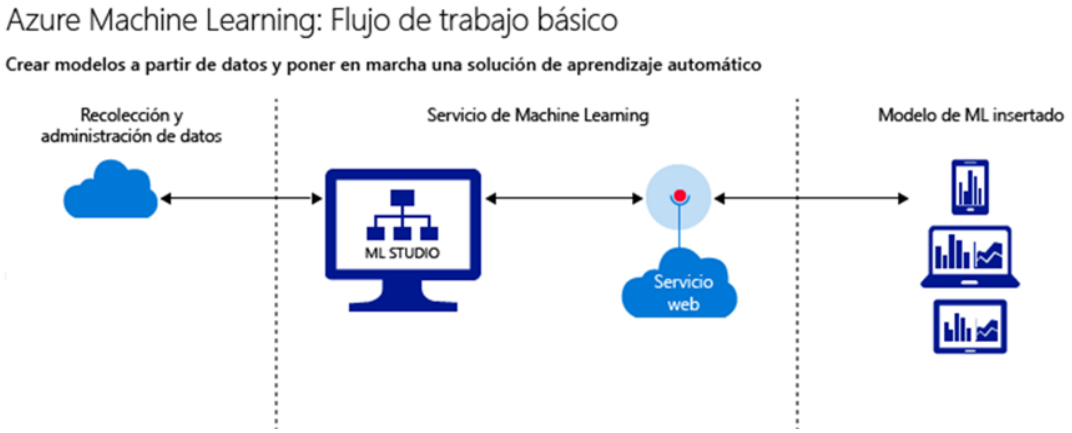


Figura 2.8: Ejemplo de flujo de trabajo de MS Azure ML

Los módulos son los responsables de trabajar los datos en esta herramienta y cada uno tiene una labor importante dentro del proyecto que se desee realizar:

- **Manejo de los datos:**
  - ***Format conversions:*** este módulo convierte los datos que se desean trabajar en datos capaz de ser reconocidos por esta herramienta.
  - ***Input and Output:*** lee los datos de una ubicación específica, como la nube,

Hadoop clusters, paginas web, etc.

- ***Tansformation:*** prepara los datos para su análisis, puede cambiar el tipo de dato, escalar o normalizarlos y mucho más.
- ***Learning With Counts:*** usa probabilidades para trabajar con gran cantidad de datos.
- ***Manipulation:*** provee varios métodos para los datos como remover, reemplazar, ubicar *missing values*, concatenar dos columnas, etc.
- ***Sample and Split:*** divide los datos por algún criterio, normalmente para crear *test* de entrenamiento.
- ***Scale and Reduce:*** transforma numéricamente los datos.
- ***Feature Selection:*** identifica las mejores características de los datos.

- **Herramientas:**
  - ***Machine Learning:*** contiene algoritmos de *machine learning* que son soportados por la herramienta.
  - ***OpenCV Library Modules:*** da un fácil acceso a librerías open surce para el procesamiento de imágenes y su clasificación.
  - ***R Language Modules:*** agrega código R al experimento.
  - R ***Python Language Modules:*** agrega código Python al experimento.
  - ***Statical Functions:*** calcula distribución probabilista, crea cálculos personalizados y cálculos relacionados a variables numéricas.
  - ***Text Analytics:*** permite trabajar con textos como reconocimiento, conversión de texto, etc.
  - ***Time Series:*** crea modelos predictivos usando algoritmos de series de tiempo.

2.6.2. R

El lenguaje R es uno de los lenguajes más utilizados para trabajar con datos, ya que cuenta con una amplia cantidad de operaciones con matrices y vectores, lo que facilita

su trabajo con bases de datos. También cuenta con una gran cantidad de bibliotecas para facilitar esta labor y junto con esto, al ser un lenguaje diseñado específicamente para la estadística, es muy preciso y exacto para estos tipos de análisis [10].

R tiene a su disposición distintas funciones, como por ejemplo, la creación de gráficos que aportan en la visualización de los datos y sus respectivos análisis. R ha implementando una gran cantidad de algoritmos referentes a *machine learning*, ya que gracias a sus herramientas estadísticas, muchas personas han optado por usar este lenguaje para sus investigaciones [10].

Usar R para *machine learning* tiene sus ventajas entre las cuales se destaca:

- **Es gratis:** se puede utilizar en cualquier proyecto o empresa, sin necesidad de permisos o políticas de esta, siendo una herramienta transportable que puede ser utilizada desde cualquier punto o estación de trabajo.
- **Actualizado:** la mayoría de las investigaciones estadísticas de *machine learning* utilizan el lenguaje R [11] [10], por lo que se van agregando constantemente algoritmos y bibliotecas para poder ser descargadas, junto con una amplia documentación.
- **Herramienta útil:** muchas empresas reconocen R como una herramienta importante dentro de las habilidades de un empleado, por lo que brinda una ventaja sobre el resto al tener los conocimientos suficientes sobre esta herramienta.

2.6.3. Python

Python es un lenguaje de *scripting* independiente de plataforma, que está preparado para realizar cualquier tipo de programa, y que en el último tiempo, el uso de este lenguaje se ha vuelto muy popular [12].

Una de las ventajas que ofrece Python sobre otros lenguajes de programación, es la gran comunidad de desarrolladores, la cual ha contribuido con una gran variedad de bibliotecas para extender la funcionalidad del lenguaje. Para el caso de *machine learning*, las principales bibliotecas que se utilizan son:

- ***Scikit-learn:*** es la principal biblioteca para trabajar con *machine learning*. Incluye la implementación de un gran número de algoritmos de aprendizaje.
- ***Statmodels:*** esta biblioteca tiene un enfoque en modelos estadísticos y es usada principalmente para análisis de modelos descriptivos y exploratorios.

2.6.4. Análisis crítico de herramientas para Machine Learning

Para un análisis crítico de las herramientas de *machine learning* se seleccionan algunos atributos relevantes que permiten entregar comparaciones de las tres herramientas descritas anteriormente y que proporcionan una guía sobre cuál de ellas puede ser utilizada dependiendo del proyecto a desarrollar. Estos atributos a comparar son: usabilidad, aprendizaje, costo, innovación y flexibilidad de la herramienta, los cuales pueden reflejarse en la tabla a continuación 2.1:

Atributo	**R**	**Python**	**Azure**
Usabilidad	Lenguaje orientado a la estadística.	Lenguaje multipropósito.	Herramienta, orientada a *Machine Learning*.
Aprendizaje	Curva de aprendizaje más lenta.	Curva de aprendizaje más rápida.	Curva de aprendizaje más rápida. No requiere conocimientos avanzados.
Costo	Gratuito.	Gratuito.	Versión gratuita y pagada.
Innovación	Es académico. Gracias a esto, es líder en esta área y su desarrollo es confiable.	Es el segundo lenguaje para desarrollar esta área. Dado que tiene contribuciones abiertas, hay probabilidades de encontrarse con errores en las últimas novedades.	Servicio en la nube, sin requerir instalación local, lo que ofrece información disponible en línea.
Flexibilidad	Gran flexibilidad gracias a su biblioteca.	Gran flexibilidad gracias a su biblioteca.	Módulos predeterminados, sin disponibilidad de adaptación.

Tabla 2.1: Comparativo de las diferentes herramientas usadas para machine learning

3. Desarrollo y resultados

En este capítulo se aplica lo expuesto en el marco teórico, profundizando en el conocimiento y funcionamiento de la compañía y en el diseño de un modelo que permita lograr los principales objetivos establecidos en las páginas anteriores.

Para comenzar a trabajar con los datos, se tuvo que agendar varias reuniones con el personal a cargo no sólo para entender el negocio, sino que también comprender los datos, estructurarlos y prepararlos para luego realizar el modelamiento, logrando identificar procesos y reglas importantes en el negocio que son necesarios para evaluar el producto final.

Todas las predicciones que se describen a lo largo de estas páginas, se generaron usando la herramienta R que luego fue incorporada en un módulo de Microsoft Azure Machine Learning con el fin de tener una herramienta fácil de usar y con potencial a futuro. Se trabajó con archivos .csv que son manejables por cualquier usuario en Excel, haciendo de la herramienta Microsoft Azure ML la escogida para no intervenir en el trabajo y formato en el cual se trabaja en la compañía.

Se encontrarán distintas maneras de mejorar las predicciones, cambiando la forma en que se trabaja en la compañía y contribuyendo a alcanzar el principal objetivo: mejorar el nivel de *forecast accuracy*.

3.1. Comprensión del negocio

3.1.1. Terminología de la compañía

Para una mejor comprensión del negocio, a continuación se detalla un glosario que permitirá conocer términos utilizados más adelante:

- ***Forecast***: pronóstico o estimación de demanda o venta para cada cliente / producto de la compañía, reflejado en una unidad de medida específica (toneladas, bultos, cajas, *display* y producto unitario).

- ***Forecast Accuracy***: es el porcentaje de asertividad de *forecast* por cliente / producto y que también se puede medir a total compañía.
- ***Prox Vence***: cuando al producto le quedan 6 meses de vida útil, el producto entra en estado *prox vence*, mejor conocido como próximo a vencimiento.
- ***SKU***: unidad de producto utilizado para la cadena de suministro. Cada producto está identificado por un código único que se conoce como SKU y que hace referencia a un producto específico almacenado o en tránsito en un determinado lugar.
- **Periodos:** cuando se haga referencia a esta denominación, se hablará de periodos de tiempo en que se desarrolla el negocio. Para Mars, el año cuenta de 13 periodos, donde cada uno se compone por 4 semanas de trabajo. Para el desarrollo de este trabajo, no se hablará de año calendario.
- **Categoría:** la compañía trabaja en Chile con dos unidades de negocio que se conocen como categoría de productos: Mars Petcare y Mars Chocolate.
- **Canales de comercialización:** los canales de comercialización que se trabajan en Mars son:
  - **Supermercados:** canal de comercialización moderno, al por menor y donde se venden alimentos y otros productos para el consumidor final.
  - **Mayoristas:** tiendas que venden con descuentos grandes unidades y que pueden ir dirigidos a almaceneros y consumidor final.
  - **Tiendas tradicionales:** se conocen como tiendas tradicionales a distribuidores regionales, botillerías, almacenes de barrios, quioscos, estaciones de servicio, entre otros.
- **Unidades de venta:** cada producto en cada cliente se comercializa de una forma diferente que se describen a continuación:
  - **Cajas:** unidad de comercialización más grande y que puede contener *display* o unidades de producto.
  - ***Display:*** unidad intermedia de comercialización que puede contener productos unitarios.

- **Unidad:** formato de comercialización más pequeño y se refiere al producto de manera individual.
- **Bultos:** unidad de medida similar a la caja, solo que su materialidad es distinta. En este caso, son bolsas que contienen varios productos unitarios internamente.

3.1.2. Objetivos del negocio

Mars Chile funciona sólo como unidad importadora desde sus plantas de Argentina, México y Estados Unidos, comercializando sólo dos de sus cinco unidades de negocio: Mars Petcare y Mars Chocolate. Dentro de lo que Mars quiere lograr como compañía, uno de sus objetivos más importantes es satisfacer a los clientes a través de la calidad de sus productos, realizando diversas operaciones para lograr este cometido, y dentro de estas, está entregar los productos en los tiempos de caducidad propuestos por la compañía. Se puede revisar en detalle el proceso logístico desde la producción hasta la llegada del producto al centro de distribución de la compañía en las figuras 1.1 y 1.2 ya presentadas, que permiten comprender el funcionamiento del negocio de manera global y lo que se requiere para tener disponible los productos para su venta. En resumen, los productos se encuentran disponibles para su venta con más de 12 meses de vida útil en el caso de Petcare y en el caso de Mars Chocolate, los productos se encuentran disponibles para su venta con sólo 8-9 meses de vida útil. Lo importante de esto es que los clientes sólo aceptan comercializar los productos con más de 6 meses.

Los criterios para el éxito de negocio radican fundamentalmente en el indicador usado en la empresa para predecir la demanda, ya que una mejor aproximación al valor real, indica una buena predicción de demanda. Junto con este indicador existen varios criterios que también son considerados al momento de evaluar la problemática que tiene hoy la organización en relación a los productos próximos a vencer, como: cantidad de productos vencidos en la bodega, cantidad de productos que son devueltos por los clientes por motivos de vencimiento y cantidad de productos que son utilizados en las

ofertas en las cadenas más grandes de clientes con motivos de vencimiento.

Junto con lo anterior, una de las formas en que la compañía logra el éxito en la calidad de sus productos, es el contrato que tiene con sus clientes, ya que al momento de no cumplir los productos con los estándares propuestos, estos son retirados del mercado, asumiendo el costo del producto y entregando a sus clientes nuevos productos con la calidad deseada.

3.1.3. Situación actual

La compañía gasta mucho tiempo y recursos para obtener la demanda futura, actuando a través de reuniones con los encargados de ventas y sus clientes para poder tener un plan de demanda de sus productos. Junto con este esquema de trabajo, el área de logística usa herramientas estadísticas como regresión lineal, logrando un aproximado que entrega finalmente, la cantidad de productos a importar de acuerdo a la demanda.

Dentro de la situación actual de la empresa también existen diversos agentes externos que influyen de cierta medida en la demanda, siendo el más importante el impacto de venta de productos en periodos cruciales de la compañía, por ejemplo, imponiendo metas anuales, que se deben alcanzar haciendo que la demanda varíe en un parámetro arbitrario dictado por los jefes de cada área. Este problema está interiorizado dentro de los encargados de la empresa, pero al ser una empresa multinacional, generalmente las metas y el crecimiento, vienen exigidos por un modelo regional que maneja América Latina, haciendo de esto una obligación para obtener los resultados esperados a fin de año. En general, este problema de demanda se ve amortiguado, ya que al ser esta empresa fabricante e importadora a la vez, al aumentar los números de demanda, también lo hacen en el área de fabricación y disminuyen en la importación al perder productos por vencimiento. Por este motivo, hasta ahora no ha sido un problema tan grave globalmente pero un equilibro en estos puntos sin duda que aumentaría la ganancia de la empresa, impactando directamente en el área local.

Con esto, el costo beneficio que se produciría al mejorar la forma en que se predice

la demanda sería sacrificar la venta que pudiera estar proyectada en meses posteriores y que pudiera ser superior al que se podría entregar con el nuevo modelo. Esta mejora, evitaría generar sobre *stock* de productos en los clientes, problemas futuros por vencimientos y costos de descuentos por liquidaciones. Implementar una solución a la manera de predecir la demanda, ayudará a la compañía a entender realmente las oportunidades que tiene en el mercado, potenciar sus productos ganadores y planificar de una mejor manera la importación, evitando quiebres por mala organización o por sobre *stock* a un cliente en específico, dejando sin *stock* al resto de ellos.

3.1.4. Determinación de los objetivos de minería de datos

Para el proyecto las metas son claras, mejorar la manera en que se realiza la predicción de la demanda utilizando inteligencia de negocios con una herramienta fácil de usar por los usuarios de la empresa. En el caso de minería de datos el objetivo es normalizar los datos presentes en la empresa junto con su trabajo con el objetivo de encontrar características de éstos y presentarlos en una forma comprensible para el uso de los usuarios.

3.1.5. Plan de proyecto

Para alcanzar con éxito los objetivos propuestos, se llevó a cabo la planificación de un plan de proyecto por etapas que permita construir y lograr paso a paso los objetivos principales y específicos planteados en páginas anteriores.

Cada una de estas etapas se describen a continuación:

- **Definición de métricas para alcanzar los objetivos:** para alcanzar el objetivo principal del proyecto se definieron métricas que permitirán darle seguimiento a los avances que se van dando:
  - **Volumen de venta:** se tomará como métrica inicial el análisis de la venta de la compañía para determinar en qué clientes y en qué productos se llevará a cabo el proyecto.

- **Asertividad en la estimación:** tal y como se muestra en páginas anteriores esta es la principal métrica que contribuirá a determinar el éxito del proyecto. El *forecast accuracy* o asertividad en la estimación debe estar por sobre el 70 %.

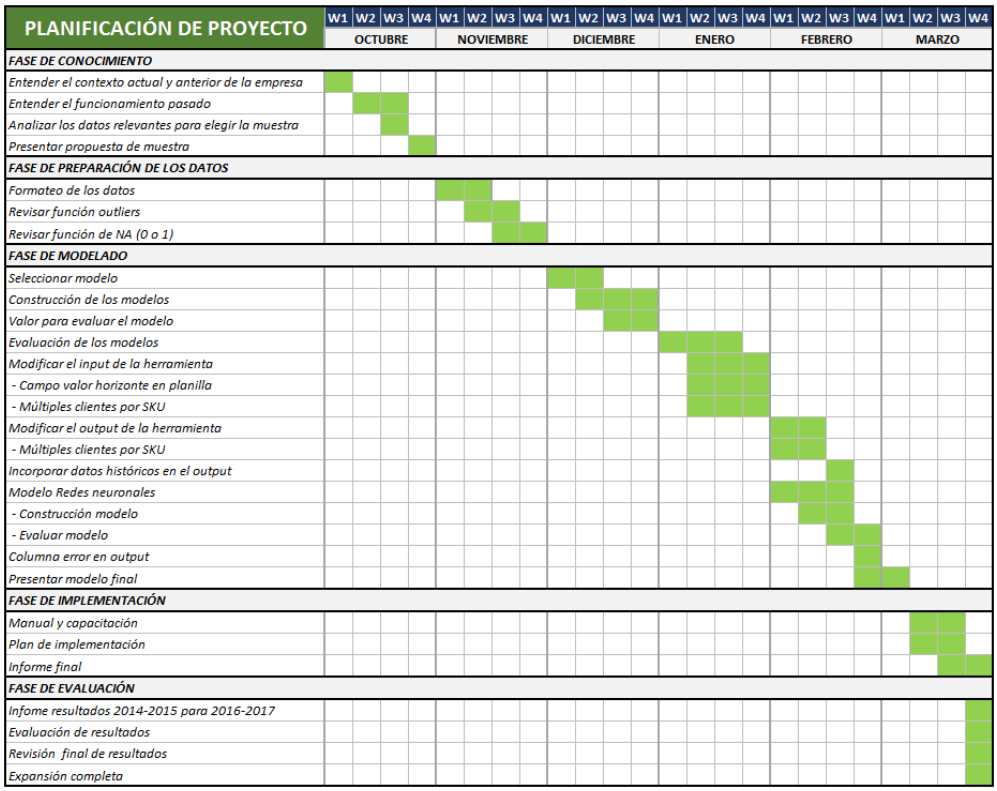


Figura 3.1: Carta Gantt del proyecto.

- **Entregables del proyecto:** para determinar los entregables del proyecto, se define una lista de actividades que se deben llevar a cabo para cumplir cada uno de los objetivos junto con fechas estimadas de entrega:
  - **Fase de Conocimiento:** se detalla el contexto de la compañía, historia y datos relevantes.
  - **Fase de preparación de datos:** se presenta el formateo de datos y revisión de funciones.
  - **Fase de Modelado:** en esta etapa se define y desarrolla el modelo a trabajar.
  - **Fase de implementación:** se lleva a cabo la implementación del modelo en

la compañía, capacitación y desarrollo de informe final.

- **Fase de Evaluación:** en esta etapa final se evalúan los resultados del modelo y se define su expansión.

Área Funcional	***Cargo***	***Etapa de Participación***
Ventas y Operaciones	Manager	Proceso Completo.
Ventas y Operaciones	Analista	Proceso Completo.
Ventas y Operaciones	Memorista	Proceso Completo.
Operaciones	Gerente Importaciones	Fase de conocimiento. Fase de implementación.
Operaciones	Gerente Distribución	Fase de conocimiento. Fase de implementación.
Operaciones	Analista Importaciones	Fase de conocimiento. Fase de implementación.
Ventas	Gerente Ventas Chocolate	Fase de conocimiento.
Ventas	Gerente Ventas Mascotas	Fase de conocimiento.
Ventas	Key Account Manager	Fase de conocimiento. Fase de implementación.
Marketing	Brand Manager	Fase de Conocimiento.

Tabla 3.1: Participación usuarios en el proyecto

- **Planificación del proyecto:** esta etapa de planificación se separa en dos temas

relevantes a tratar: planificación de actividades y gestión de personas. La primera parte tiene relación con detallar las actividades numeradas en el punto anterior y que se desglosan en aquellas más específicas que contribuyen a llevar a cabo cada actividad, las cuales pueden verse reflejadas y resumidas en la carta gantt de la figura 3.1.

La segunda parte tiene relación con aquellos usuarios que deben participar en la construcción del modelo y de quienes simplemente deben ser usuarios activos una vez que se realicen las pruebas, junto como también en la etapa de despliegue, en la cual las áreas y su participación en este proyectos serán indispensables para el exito de este proyecto. La lista de los encargados pueden verse resumidas en la tabla 3.1.

3.2. Comprensión de los datos

Para poder comprender los datos, lo primero es revisar cómo éstos se presentan en la compañía y cómo se trabajan. Bajo este contexto, la empresa cuenta con varias maneras de presentar su información, tales como: base de datos, planillas de texto, planillas de Excel y gráficos. Por otra parte, los datos se trabajan bajo un área específica de demanda quien maneja toda la información relevante para llevar a cabo el proyecto. La persona encargada del área es indispensable no sólo en la etapa de comprensión, sino que también en las subetapas que se detallan a continuación:

3.2.1. Recolección de datos iniciales

Si bien la compañía cuenta con varias formas de presentación, todas ellas tienen una única fuente en común, la cual es la herramienta SAP. Desde ahí se puede obtener toda la información necesaria no sólo de demanda, sino que de todas las áreas de la compañía.

Los archivos de demanda generados se descargan en formato de texto plano, en

cambio, los archivos relacionados con los productos y clientes, en los cuales están los SKU e información de cada cliente, son obtenidos directamente del área de Marketing y Ventas. Estas dos áreas además, actualizan esta información en base a cambios en la compañía en casos como por ejemplo: un cliente cambia su razón social y su demanda histórica es homologada hacia la nueva razón social con el objetivo de no perder la historia que permita generar una estimación de demanda correcta. Lo mismo, puede ocurrir con algún producto de la compañía por cambios de gramaje, empaque o descripción.

Hay que destacar que en general, todos los datos obtenidos por agentes externos a SAP son revisados meticulosamente junto con los encargados de cada una de esas áreas. El objetivo de esto es buscar cualquier error que pudiera existir, ya que cada área de la compañía posee sus datos en archivos maestros diferentes.

Sin embargo, para obtener esta información se presentan problemas inherentes al proceso que se detallan a continuación:

- La poca capacidad de SAP para descargar información histórica por problemas de velocidad de respuesta y cantidad de datos disponibles. Para el desarrollo del proyecto es indispensable contar con este histórico de ventas de al menos 2 años, por lo que que se optó por obtener históricos por rangos de meses y luego juntar los datos con un *script*.
- No todas las ventas clientes/producto estaban en la misma unidad de medida. Esto quiere decir, que algunos productos se vendieron como cajas, otros como unidades, y otros como *display* o toneladas.
- Todos los datos venían en meses calendario y la predicción debía hacerse en base a periodos Mars.
- Hay productos discontinuados o con SKU diferentes por motivos de promociones o cambios de componentes. Esta información se recolectó directamente del área de Marketing el cual contaba con un histórico con la equivalencia de los SKU de los productos discontinuados con los SKU actuales que se usan en la empresa.

- Gran cantidad de tiempo destinado a recolectar del área de clientes, los códigos de los clientes históricos y actuales, y se homologaron según correspondían. Esta homologación se lleva a cabo dado que, por ejemplo, un cliente en específico cambia su razón social y SAP arroja la venta como si fueran dos clientes diferentes cuando en realidad es sólo uno.

Se debe mencionar que un problema habitual al momento de trabajar con datos sensibles para alguna compañía es el acuerdo de confidencialidad que se firma al momento de recibir toda la información de ésta, lo que a veces complica el poder tener libre elección con los datos. Sin embargo, en este caso el acuerdo sólo establece en distorsionar los valores reales de la demanda por un valor creado a partir de una constante aleatoria establecida anteriormente en las reuniones de equipo. Lo anterior, no afecta al propósito del trabajo, ya que las curvas estarán trasladadas en cierto valor, pero su tendencia y comportamiento se apreciarán sin distorsión alguna.

Los archivos de datos obtenidos son, en primera instancia archivos de texto plano desde SAP y archivos maestros de área de ventas y marketing en planilla Excel, entre los cuales están los siguientes archivos:

- Histórico de demanda: considera la demanda desde 2014 a 2016 en archivo texto plano, descargado desde SAP.
- Histórico de producto: incluye los productos actuales y discontinuados a la fecha en archivo planilla Excel desde el área de Marketing.
- Histórico de clientes: Considera a quienes se les ha realizado por lo menos una venta entre los años 2014 a 2016, en archivo planilla Excel desde el área de Ventas.

La descripción de los datos se mencionarán en la sección siguiente.

3.2.2. Descripción de los datos

En esta etapa se presentan la descripción de los datos en donde se obtuvo la información sobre este proyecto, esto está separado en tres archivos principales descritos a

continuación:

El archivo histórico de demanda, obtenido de SAP, posee 301.660 filas, cuyos campos se encuentran descritos en la tabla 3.2. Como resumen sobre la demanda, se tiene que el promedio en el archivo es de 35.485, su valor mínimo es 0 y el valor máximo que alcanza es 9.360.

Nombre columna	**Descripción**	**Importancia**	**Tipo de valor**	**Esquema de codificación**
Fecha	Contiene la fecha de demanda de un cliente "x".	Alta	Date	DD/MM/AAAA
Canal	Tipo de cliente: Supermercado, Mayorista y Tradicional.	Baja	Entero	Numérico 2 dígitos
Número de Pagador	Código referencial del cliente.	Alta	Entero	Numérico 8 dígitos
Nombre de Pagador	Nombre de la razón social del cliente.	Media	Cadena Caract.	Alfanumérico
RUT del pagador	RUT asociado al cliente.	Media	Cadena Caract.	Formato RUT punto y guión
Categoría de Producto	Tipo de producto que se vende, en este caso, mascota y chocolate.	Media	Cadena Caract.	CONFECT o PETCARE
Número de Producto	Equivalente al SKU del mismo, dígitos numéricos.	Alta	Entero	Numérico 8 dígitos
Nombre de Producto	Descripción del producto.	Media	Cadena Caract.	Alfanumérico
Demanda en Bultos	Demanda del producto, en unidades	Alta	Decimal	Numérico

Tabla 3.2: Campos incluidos en el archivo histórico de demanda.

El archivo histórico de clientes, obtenido desde el área de ventas posee 221 filas, en las cuales están presentes los siguientes campos reflejados en la tabla 3.3.

Nombre de Columna	**Descripción**	**Importancia**	**Tipo de valor**	**Esquema de Codificación**
RUT Cliente	Identificador universal del cliente	Baja	Cadena Caract.	Formato RUT punto y guión
Número de pagador	Código referencial del cliente	Alta	Entero	Numérico 8 dígitos
Nombre del pagador	Nombre de la razón social del cliente	Media	Cadena Caract.	Alfanumérico
Región	Región a la cual pertenece	Baja	Cadena Caract.	Numérico romano de región más ciudad XX-Ciudad
Área	Área a la cual pertenece: Región Metropolitana o Región	Media	Cadena Caract.	REG o RMS
Canal	Tipo de cliente: Supermercado, Mayorista y Tradicional	Baja	Entero	Numérico 2 dígitos
Número vendedor	Identificador del vendedor a cargo.	Baja	Entero	Numérico 8 dígitos
Nombre Vendedor	Nombre del vendedor a cargo.	Baja	Cadena Caract.	Alfanumérico

Tabla 3.3: Campos incluidos en el archivo histórico de cliente.

Por ultimo el archivo histórico producto, obtenido desde el área de Marketing, tiene 397 filas y contiene los campos de la tabla 3.4.

Nombre de columna	Descripción	Importancia	Tipo de Valor	Esquema de Codificación
SKU	Código único de identificación de producto	Alta	Entero	Numérico 8 dígitos
Nombre producto	Descripción del producto.	Alta	Cadena Caract.	Alfanumérico
Categoría	Tipo de producto que se vende, en este caso, mascotas y chocolates	Baja	Cadena Caract.	Alfanumérico
Segmento	Tipo de producto vendido para una categoría en específico.	Baja	Cadena Caract.	Alfanumérico
Unidad de Venta	Formato en que se vende el producto	Media	Cadena Caract.	Alfanumérica
EAN	Identificador único del producto dado por la Cámara de Comercio de Santiago	Baja	Entero	Numérico 13 dígitos
DUN	Identificador único del envase que lleva el producto dado por la Cámara de Comercio de Santiago	Baja	Entero	Numérico 14 dígitos
Estado	Estado en que se encuentra el producto.	Media	Cadena Caract.	Alfanumérico

Tabla 3.4: Campos incluidos en el archivo histórico de producto.

De las tablas 3.2 a 3.4, hay cuatro tipos de datos que son utilizados como datos de entrada de la herramienta:

- **Fecha:** es indispensable para realizar la predicción, ya que van de acuerdo a periodos del año.
- **Demanda:** es el número a predecir, esta demanda se usa como la variable en las

predicciones.

- **Cliente y Producto:** son los filtros en los cuales se trabaja durante todo el proyecto, haciendo cruces de información y analizando estos casos.

3.2.3. Verificación de la calidad de los datos

En el proceso de la calidad de los datos, hay que tener presente qué es la calidad de estos y cómo se comportan estos. La calidad de los datos tiene relación a la consistencia según el cual se forman los datos. Los valores individuales de los campos, la cantidad y distribución de los valores contribuyen a medir de qué tan buena calidad son, permitiendo además encontrar valores fuera de rango, los cuales pueden obstruir el proceso.

Para verificar la calidad de estos datos se presentan a continuación los factores para determinar la calidad, junto con una tabla con datos y sus respectivos factores. Hay que mencionar también que la calidad, en este caso, apunta solamente a los datos de alta importancia indicados anteriormente, es decir, fecha, demanda, clientes y producto, por lo que los datos con baja importancia no son verificados a continuación. Los factores a considerar son los siguientes:

- **Datos perdidos:** valores vacíos o sin respuesta.
- **Errores de datos:** errores tipográficos realizados generalmente al introducirlos en el sistema.
- **Errores de medición:** datos introducidos correctamente pero basados en mediciones erróneas.
- **Incoherencia de codificación:** valores no estándares o valores incoherentes.

Dato	Datos Perdidos	Errores de medida o tipo	Incoherencia de codificación
Fecha	Cantidad NA: 68	No aplica	Formato Fecha DD-MM-AAAA, en vez de DD/MM/AAAA. Cantidad: 118.519
Demanda	Cantidad NA: 1.146	Valor decimal en unidades valor entero Cantidad: 546	No aplica
Cliente	Cantidad NA: 8	No aplica	No aplica
Producto	Cantidad NA: 2	No aplica	No aplica

Tabla 3.5: Resumen de calidad de los datos.

De la tabla 3.5 se desprende que la calidad de los datos presentes para este estudio poseen ciertos problemas en relación a errores de medición e incoherencia de codificación; estos errores son ciertamente más simples de modificar y corregir. Otro de los problemas es que la demanda, en una considerable cantidad de datos, está representada en valores decimales, cuando debiera estar en unidades enteras, siendo el mínimo un valor igual a 0 y variando siempre en números enteros. Por otra parte, está la incoherencia de codificación en el dato fecha. La cantidad de estos y el error es muy frecuente, ya que hay varios estándares de escritura de una fecha y es aceptada en la mayoría de los sistemas en más de una forma.

Al ser una recopilación oficial de la demanda de los productos de la compañía, los datos obtenidos representan fielmente el comportamiento de la compañía por lo que al momento de ser obtenidos de diversas fuentes, éstos son consolidados y enviados a SAP.

Este paso anterior, que es manual, puede generar un error y repercutir en todas las áreas funcionales de la compañía, debido a que todos poseen acceso a la misma información. Dado su nivel de exposición, se asume que los datos rescatados anteriormente son de la mejor calidad posible.

Así mismo el rango de los valores de demanda varían desde 0 hasta infinito, siendo

verificado que no existan datos menores a 0 lo que produciría una demanda negativa que, por definición, es imposible.

3.3. Preparación de los datos

Esta etapa es sumamente importante, ya que en primera instancia se necesitan los datos de los productos más importantes de la empresa, pero hay que dejar cabida para obtener también del resto. Para preparar los datos, inicialmente se ordenaron según clientes y categorías de productos (mascotas o chocolates). Esto, debido a que los productos están relacionados directamente con los clientes y con la categoría, se crearon filtros para obtener la información sobre estos dos campos, por lo que los datos obtenidos son la demanda de un producto " x " de un cliente " y ".

A continuación, se detallan las subetapas de la fase de preparación de los datos.

3.3.1. Selección de datos

En esta etapa se describen los datos seleccionados para ser utilizados en el proyecto:

- **Selección de elementos:** este estudio se limita a la selección de tres clientes, entre los cuales destacan Walmart, Cencosud y Promerco; junto con esto también se trabaja con el total de los clientes de la compañía.
- **Selección de atributos:** los atributos en este caso, son el número de producto y la demanda de estos.

3.3.2. Limpieza de los datos

Esta actividad, es una de las más difíciles de realizar, ya que un problema que surge en el negocio es cuando la demanda aumenta o disminuye considerablemente, por motivos externos al normal funcionamiento del mercado. En muchas empresas las metas de venta son el objetivo número uno, por lo que quienes trabajan en esta área tienen un tiempo límite para poder cumplirlas, el que puede ser mensual, semestral o

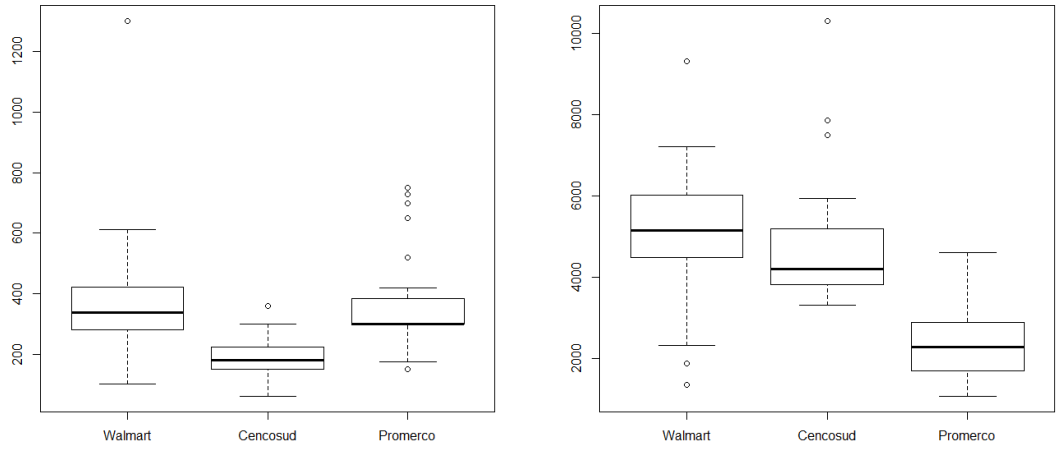
anual. Este cumplimiento se impone generalmente a principios de año con un análisis de crecimiento de mercado, de la empresa, del producto, y de muchos factores que no vienen al caso profundizar; pero hay que tener en cuenta que estas metas muchas veces entregan valores de ventas reales, pero de una forma difícil de estudiar, dada su posible irregularidad.

En este caso particular del negocio, la empresa tiene como tiempo final de cumplimiento de metas el término del año, por lo que las ventas en los últimos periodos del año se incrementan considerablemente por muchos factores como promociones, empuje de cantidad de productos en meses donde hay menor la demanda, ocasionando sobre *stock* a un cliente para llegar a la meta, sin pensar en las consecuencias futuras que traen estos actos.

Lo anterior es un problema para el proyecto, ya que dichos actos son por decisiones puntuales y no por un comportamiento regular del negocio, por lo que es difícil registrar cuál es la demanda normal en el histórico o cuál es una sobre demanda por estos actos.

Dichos datos de demanda pueden ser considerados como *outliers*, tal y como se aprecia en las figuras 3.2 y 3.3. Estos datos de demanda, se alejan por sobre el límite superior, ya que el límite inferior siempre será 0, porque por definición no existe una demanda negativa. Para trabajar con este problema y limpiar bien los datos que serán utilizados para el modelado, es necesario aplicar alguna técnica que encuentre estos valores y los elimine de alguna forma del *set* de datos.

Para este caso, se utiliza la herramienta *BoxPlot* que muestra, a grandes rasgos, la concentración de los datos y cuáles datos sobresalen de los límites del gráfico.



(a) Snickers. (b) Pedigree 15Kg.

Figura 3.2: Gráfico de Boxplot para Snickers y Pedigree con sus respectivos outliers de demanda.

(a) Snickers. (b) Pedigree 15Kg.

Figura 3.3: Gráfica de Boxplot para el total de productos con sus respectivos outliers de demanda.

A pesar de que estos valores considerados *outliers* sean esporádicos y no se aprecien

de manera habitual en las ventas de la compañía, no se puede tomar la determinación de eliminarlos y/o reemplazarlos por una demanda igual a 0. El motivo de esto, es porque este cambio distorsionaría el trabajo con los datos, perdiendo información relevante y provocando inconsistencia en los datos históricos y en las predicciones futuras.

Dado lo anterior, para estos valores *outliers*, se determina que se reemplazarán por un valor igual al límite superior de los datos de demanda posible en la herramienta *Boxplot*.

El uso de *Boxplot* ayuda a normalizar los datos para una mejor aproximación al utilizar algún modelo. Además, se tiene un parámetro que indica el porcentaje de *outliers* para cambiar. Este valor puede ser sensibilizado para mejorar los resultados, cambiando el número de *outliers* que pueden presentar los datos.

Terminada la limpieza del caso particular de los *outliers*, a continuación se describe la limpieza en relación a los problemas encontrados en la calidad de los datos:

- **Datos perdidos:** estos tienen que ser separados en dos partes, ya que el significado de los datos es diferente. Fecha, cliente y producto son datos sensibles en el sentido de que si son nulos (NA), éstos no pueden ser utilizados, ya que no tienen significado y la mejor forma de tratarlos es eliminar la fila directamente de la base de datos porque no es posible obtener información alguna de estos campos en blanco. Esta es la opción más rápida y fácil de realizar ya que existe la posibilidad de obtener alguna información cruzando algunos datos con las ventas realizadas por un vendedor, pero esta tarea es muy engorrosa y tomando en consideración la cantidad de datos que presentan este problema, se decide simplemente eliminarlos. Por otra parte, los datos perdidos en la demanda, pueden ser tratados como un valor igual a 0; esto se concluye debido a que al no existir un dato en esta columna, lo cual significa implícitamente que es igual a un valor 0 de demanda, o sea que no existe una demanda del producto.
- **Errores de medidas:** las mediciones obtenidas que son de carácter decimal, son el problema a tratar, ya que al ser este campo en unidades, deberían ser números

enteros y no decimales. Para este tratamiento de datos, se tomaron en cuenta los valores y el producto en el cual existe este problema. Por políticas de la empresa, los productos son vendidos en *display*, esto quiere decir que son productos empaquetados que contienen una cierta cantidad y no pueden ser vendidos en menor cantidad abriendo este empaque.

Sin embargo, al hablar de clientes, hay sólo uno en específico al que se le vende en unidades decimales. Luego de investigar, se llega a concluir que estos productos son muestras que son inventariadas e ingresadas como ventas al sistema pero con precio $0,001$.

Teniendo en cuenta esto y también que los decimales son entre el rango de $]0, 1[$, la solución parece más simple de lo que parece, cambiar todos los números decimales al valor entero más próximo haciendo función techo, esta función es la apropiada, ya que al abrir un *display* y entregar sólo 1 unidad como muestra, en el fondo la compañía pierde los demás productos debido a que no es posible vender un *display* abierto.

- **Incoherencia de codificación:** la incoherencia de los formatos de fecha presentes en la base de datos, es tratada simplemente eligiendo un solo formato y realizando un cambio de los demás. En este caso, el formato elegido para la fecha es DD/MM/AAAA, por lo que los demás formatos son cambiados a éste. Hay que mencionar que solo el problema se reduce a un cambio de signo (cambiar el signo $-$ por $/$), por lo que esto es simple de realizar. Si hubiera existido un problema de formato fecha de otro tipo, por ejemplo un formato de fecha en inglés (MM/DD/AAAA), el tratamiento hubiera sido más complejo de realizar y tendría un detalle distinto de cómo fue tratado.

3.3.3. Construcción de nuevos datos

La creación de un nuevo campo es indispensable para el proyecto, ya que en la base de datos el campo fecha no es representativo al negocio. Este campo, da como

información el día, mes y año de la demanda de un cliente y producto, pero para el proyecto, la forma en que se quiere obtener la demanda es en periodos Mars. En la empresa existen los llamados periodos Mars, con duración de 4 semanas partiendo de la primera semana del año. Este periodo puede ser obtenido haciendo una función que sea capaz de reconocer la fecha del año y asignarle el periodo y número de semana Mars en la cual se encuentra. Con esto se estructura una nueva columna llamada periodo, la que es de la siguiente forma AAAAPPSS, donde las primeras cuatro letras equivalen al año, las siguientes dos corresponden al periodo y las últimas dos a la semana. Como ejemplo, si se quiere referir a la semana 2 del periodo 5 del año 2017, el formato será 20170502.

3.3.4. Integración de los datos

Como se mencionó anteriormente los registros son rescatados en rangos de meses, ya que SAP no permite obtener el rango de años necesarios en una sola descarga. Por este motivo, lo primero fue recopilar todos los archivos de las fechas por meses y unirlos con un *script* para tener un archivo único con todos los datos necesarios desde el 2014 a la fecha. Estos archivos están en formato de texto, donde los campos están separados por un " ; " entre sí; al llevar el único archivo a Excel, para facilidad de tratar con los datos, se dejó en cada columna un único campo, eliminado el " ; " que los separaba.

Para poder integrar los datos desde los distintos archivos que existen es necesario entender el significado de los datos para poder integrarlos de la mejor forma posible. Primero que todo se tiene que uno de los datos a integrar es el *SKU* de los productos del histórico de demanda junto con el de productos.

- ***SKU* producto:** se actualizan los *SKU* de los productos a la fecha, ya que como se comenta, la base de datos es desde el 2014, por lo que dentro de todos estos años, han habido varios cambios de *SKU* y de descripción de productos, así como también cambios en el empaque o en la cantidad que posee cada producto. Todos estos problemas conllevan a describir un *SKU* de producto equivalente.

- ***SKU* producto equivalente:** la idea es normalizar los *SKU* de los productos, ya que éstos van cambiando dependiendo de las promociones o *packs*. Para poder registrar las promociones en SAP, la empresa crea uno nuevo que contenga el mismo producto con alguna característica especial, por ejemplo, un regalo por la compra del producto. Esto se hace para distinguir la venta del producto sin promoción o con promoción. Los *SKU* equivalentes también tienen efecto con productos discontinuados, por lo que se optó por utilizar el *SKU* más recientemente creado sin promoción e integrarlo a todos los productos de igual o parecidas características.
- **Código cliente equivalente:** al existir clientes que poseen diferentes códigos a lo largo del tiempo por diversos motivos como cambio de razón social, alianza con otros inversionistas, se utilizó la base de datos actual de clientes y se identificó si históricamente tuvo algún cambio para llevar a cabo la homologación.

  Con estas descripciones y el entendimiento del resto de los campos, estos se integran a la base principal la cual es la de demanda, en donde están presentes los *SKU* de producto y los códigos de los clientes. Estos dos datos son revisados meticulosamente en conjunto con los históricos de productos y clientes para lograr una fusión de la información, y poder actualizar a la fecha de hoy cualquier cambio pasado que haya sufrido cualquiera de estos campos.

  Con la integración de los datos realizada se da paso a la siguiente etapa correspondiente al formateo de los datos ya integrados.

3.3.5. Formateo de los datos

Dado que se trabajan con unidades de demanda, éstas pueden sobrepasar el valor de 1.000 por lo que se eliminó el punto en estos campos, es decir, se trabajó de la siguiente forma 1000. Para todos los datos decimales, se utilizó una función "techo"para que como resultado, se tengan números enteros (punto descrito y explicado en la calidad de los datos).

Dado que ya se le dio estructura al periodo de venta, y no trabajar con las fechas incluidas en la base de datos, no hubo necesidad de formatear este campo sacando los "/". Por otro lado, el mínimo valor de los datos para la demanda es 0, lo cual indica que no hubo ventas de ese producto, lo que en la base de datos se representa como un valor nulo, por lo que el formato del valor *null* se cambia a 0 para utilizar el mismo formato durante todos los valores de la demanda (valores numéricos).

También hay que mencionar que para la fase de modelado, el valor 0 presenta un problema en el *input* de la herramienta, por lo que se cambió por un valor cercano a 0, en este caso 1. Esto se elige ya que no influye en gran medida en el promedio de la demanda, y no distorsiona las predicciones ni las ventas. Cabe mencionar que estos cambios ocurren porque son modelos matemáticos y el valor 0 hace que la función tome este mismo valor en cada iteración.

4. Desarrollo y validación del modelo de *forecast*

En este capitulo se escoge el modelo a desarrollar, las pruebas necesarias para la construcción del modelo y el desarrollo de cada uno de ellos. En la última parte del capítulo se evaluarán los modelos desarrollados y cómo se comportan frente al objetivo de mejorar el *forecast* de la compañía.

4.1. Modelado

En esta actividad es importante seleccionar un método que sea capaz de cumplir y llevar a cabo el objetivo principal: predecir la demanda futura de la empresa. Para esto, se hizo un estudio de las RNA, encontrando en ella una herramienta para resolver el objetivo principal.

4.1.1. Selección de la técnica de modelado

Las RNA son el método seleccionado para poder predecir la demanda de los productos de la empresa. Los motivos de esto es porque existe una gran cantidad de trabajos de investigación, en los que cuentan con una importante habilidad de aprendizaje, puede entregar respuestas en tiempo real, manejar excepciones y entradas de datos anormales, así como también, es una técnica que puede ser utilizada para diferentes áreas de la minería de datos (clasificación, predicción, entre otros).

Este modelo cuenta con diversos requerimientos para no afectar sus resultados. A continuación se describen los puntos principales que requiere este modelo de RNA para su óptimo funcionamiento [7] :

- **Datos suficientes para resultados fiables:** para que el modelo tenga resultados fiables y se asemejen a los valores requeridos, tiene que haber un mínimo de datos, en este caso un mínimo de periodos de demanda, para ser utilizados en el modelo. La cantidad de datos disponibles, están descritos anteriormente, abarcan un periodo de datos mensuales históricos desde el 2014 a 2016, por lo que la

cantidad de datos es la suficiente para obtener resultados esperados. Hay que considerar también que los datos son escogidos desde el 2014 ya que es el inicio de la llegada de la compañía a Chile, por lo que es imposible obtener datos más antiguos que estos. Los resultados podrían mejorar considerablemente entre más periodos existan en los datos.

- **Calidad de datos:** afecta directamente los resultados del modelo, ya que con una mala calidad, los resultados son menos precisos. Teniendo en cuenta esto, la calidad trabajada en los datos en etapas anteriores, ayudan a obtener mejores resultados en el modelo.

Para trabajar con las RNA es necesario contar con alguna herramienta que sea capaz de realizar los estudios correspondientes. La escogida para usar RNA es R, la cual posee un enfoque de estudio de análisis estadístico y varias bibliotecas con numerosos modelos predictivos [11]. Además:

- es una de las herramientas más usadas para trabajar con datos, gracias a su alta capacidad.
- es gratuita.
- posee una amplia variedad de bibliotecas que ayudan a simplificar la implementación del modelo de RNA, ya que gran parte del trabajo duro, como es la programación del modelo, está resuelto.
- está orientada a la estadística, lo que da como resultado un mejor manejo de datos y su trabajo.
- al ser un lenguaje de programación ayuda a tener libertades, lo que contribuye a definir parámetros y jugar con variables libremente.

4.1.2. Plan de prueba

Para el plan de prueba existen diversas métricas que pueden ser útiles para problemas de esta índole (regresión), entre las cuales están: ME, RMSE, MAE, MPE y MAPE. Estas métricas serán presentadas durante la evaluación de los resultados, aun-

que se hará hincapié al error absoluto medio ($MAPE$), el cual como indicador, ayudará también a calcular el *forecast accuracy*. El $MAPE$ será utilizado en dos tipos de resultados; en el primer caso será en el *fitting* del modelo en comparación a la curva real de los datos, y en el segundo caso, será el error de la predicción generada por el modelo y los valores reales de los datos. Para identificar estos casos, el $MAPE$ se designará como $MAPE_f$ cuando se haga referencia al *fitting* y $MAPE_p$ a la predicción. Es importante mencionar que mientras más bajo sea el valor $MAPE$, el modelo debería entregar mejores resultados en la predicción.

Junto con este valor también está el valor de *forecast accuracy*, el cual indica qué tan precisa es la predicción. Para el desarrollo de este trabajo, el valor exigido es de un 70 % o más, lo que indica que cualquier modelo que pueda predecir arriba de un 70 % está en condiciones de ser un buen modelo a usar en la predicción. Este valor especifico se da, ya que en la empresa, las predicciones que se hacen están en el orden de un 55 % a 60 %, por lo que la meta de este trabajo es mejorar estos valores, pero siendo realistas con los números con los cuales se trabaja actualmente. Los resultados del modelado están basados en el cálculo del error. Dado esto, se escoge la regla de la pirámide geométrica [7]. Si el resultado no es satisfactorio, se debe adicionar una nueva neurona al modelo. Para trabajar en las pruebas del modelo, se decide hacer un **conjunto de entrenamiento**, donde van a estar los datos desde el inicio del 2014 hasta el fin del 2015 y donde se construye el modelo basado en estos datos y luego se realizan las **pruebas del modelo**, que toma el rango de datos desde el inicio del 2016 hasta el término del mismo año, y donde se mide la calidad de los resultados del modelo basado en el datos reales de demanda para ese año. Esta decisión va de la mano de la acotada cantidad de datos recopilados para realizar este proyecto, por lo que de esta forma con el conjunto de entrenamiento se logrará obtener modelos de buena calidad.

4.1.3. Construcción de modelo

La construcción del modelo está dado por la herramienta R, la cual con una biblioteca específica *(forecast)* genera un modelo automático y acorde a los datos entregados. Dentro de esta biblioteca, se encuentran diferentes métodos, en el que se encuentra nnetar que genera una RNA prealimentada con una capa oculta [13] y un *lagged input*, que significa utilizar el input anterior para una predicción de una serie de tiempo univariada.

Este modelo se construye de la siguiente forma:

$$NNAR(p, P, k)[m] \tag{22}$$

donde:

- p: *lags* de 1 a p, que significa comparar el periodo actual con el periodo anterior.
- P: número de *season lags*, es la cantidad de periodos a comparar.
- k: neuronas de la capa oculta.
- m: frecuencia.

Estas variables toman diferentes valores dependiendo de los datos introducidos en la función en R ya que la herramienta, procesa los datos, hace iteraciones y entrega la mejor alternativa para el conjunto de datos y al existir diversos productos a estudiar, el modelo cambia los valores dependiendo de los datos. El valor de p es obtenido automáticamente por el algoritmo en R y el valor de P por *default* es 1 en todo el trabajo.

Los resultados del modelado están basados en el cálculo del error. Dado esto, se escoge la regla de la pirámide geométrica [7] para modificar las neuronas. Si el resultado no es satisfactorio, se debe adicionar una nueva neurona al modelo. Para calcular la cantidad de neuronas de la capa oculta se utiliza esta regla, la cual consiste en definir el número máximo de neuronas mediante la ecuación:

$$k = \sqrt{m * n} \tag{23}$$

donde:

- k: número de neuronas de capa oculta.
- m: número de datos de salida.
- n: número de datos de entrada.

se tiene que:

$$k = \sqrt{1 * 27} = 5{,}19 \tag{24}$$

Dado lo anterior, se elige como máximo 6 neuronas en la capa oculta para hacer los experimentos, logrando ver que a medida que aumenta la cantidad de neuronas los modelos tienden a tener un grado de *overfitting*, haciendo que el modelo pase a ser específico y no general. Dentro de los resultados esperados, los modelos poseen una gran exactitud demostrando que un problema tiene que tener un equilibrio de neuronas, tal y como se aprecia en las tablas 4.1, 4.3 y 4.4. Esto, debido a que un número bajo de neuronas logra un *underfitting* y un número muy alto, *overfitting*.

Por último dentro de este modelo, existe un componente llamado función de activación. Esta función, es de diferentes tipos y la elección va a depender del proyecto que se quiera llevar a cabo. Para este caso, la función de activación utilizada se conoce como logística o sigmoide, que se caracteriza por ser una de las más utilizadas y que entrega mejores resultados al trabajar las RNA para predicciones [6]. Además, se realizará un pequeño experimento en donde se cambiará esta función a una lineal, para corroborar que la elección de la función sigmoide es la que mejor se adapta a este trabajo.

A continuación, se trabajan con distintos parámetros para la construcción del modelo final. Los experimentos realizados se dividen en cuatro partes:

- Variando el valor de k (neuronas en capa oculta).
- Variando el valor p (*lags*).
- Variando simultáneamente el valor p y k.
- Variando la función de activación.

Con estos cuatro tipos de experimentos, se podrá concluir, gracias a los errores que

se obtendrán, cuáles serán los mejores parámetros para construir los modelos. Cabe destacar, que la forma en elegir estos parámetros van de la mano con el $MAPE_f$ obtenido, debido a que entre más pequeño sea este valor, se da a entender que el modelo se ajusta de mejor forma a los datos.

4.1.4. Variando el valor *k*

En la primera etapa de construcción del modelo, se toman los valores predeterminados por el algoritmo en los parámetros *p* y *P*, en los cuales en todos los casos es 1 y se trabaja variando el valor de *k* de acuerdo a la regla de la pirámide geométrica, haciendo variar este parámetro de uno en uno hasta el valor máximo de 6. Con esto se puede ver en la tabla 4.1, cómo varía el error al momento de agregar neuronas de la capa oculta al modelo.

	Walmart		**Cencosud**		**Promerco**		**Total**	
k	**Snickers**	**Pedigree**	**Snickers**	**Pedigree**	**Snickers**	**Pedigree**	**Snickers**	**Pedigree**
1	30.639	21.852	16.384	15.201	16.288	18.127	25.437	9.993
2	18.512	14.505	9.391	*8.034*	9.387	10.475	17.744	5.687
3	10.016	9.948	2.120	4.132	5.852	6.751	8.524	2.978
4	6.277	4.924	**0.642**	2.788	3.172	**0.612**	5.333	1.728
5	4.147	1.454	0.779	2.576	**2.597**	1.888	3.105	**0.492**
6	**2.975**	**1.314**	0.763	**1.354**	2.689	0.867	**2.071**	0.528

Tabla 4.1: Rendimiento de la función $NNAR(1,1,k)_{[13]}$.

La tabla anterior muestra los distintos errores obtenidos con los parámetros del modelo, variando solamente *k*, en donde los valores en negrita son los valores de errores más bajos para cada caso de cliente/producto. En general, a medida que van agregando

neuronas a la capa oculta, los errores van mejorando y convergiendo rápidamente en las últimas filas de la tabla 4.1.

4.1.5. Variando el valor p

En la segunda parte de la construcción del modelo se toman diferentes valores de p y se dejan los parámetros P y k con valor igual a 1. Con esto se ve cómo la variación de este parámetro afecta en los errores del modelo. La tabla 4.2 muestra esta experiencia.

	Walmart		**Cencosud**		**Promerco**		**Total**	
p	**Snickers**	**Pedigree**	**Snickers**	**Pedigree**	**Snickers**	**Pedigree**	**Snickers**	**Pedigree**
1	30.634	20.580	23.535	15.196	17.042	25.752	25.018	11.005
2	26.988	16.689	16.691	10.650	9.401	22.926	26.040	10.032
3	26.784	18.742	17.232	9.799	8.460	17.402	20.362	**7.866**
4	26.020	**11.116**	14.691	**6.231**	**7.516**	15.945	15.179	8.134
5	**24.064**	11.716	**14.525**	9.786	7.593	**14.207**	**14.936**	9.086

Tabla 4.2: Rendimiento de la función $NNAR(p, 1, 1)_{[13]}$.

Con la tabla 4.2 se presentan los errores en negrita que indican el error más bajo para cada cliente/producto en relación al valor de p.

4.1.6. Variando simultáneamente los valores de k y p

En el tercer experimento se construye el modelo variando ambos valores k y p y dejando P como valor 1. Los rangos donde se mueven estas variables, son determinadas por los resultados obtenidos junto con la regla sobre el valor de k, ya que como se aprecia en las tablas 4.3 y 4.4, el valor de p a medida que aumenta, hay un punto en el cual sigue aumentando este valor, pero los errores obtenidos no disminuyen. En ese

instante es donde se detuvo la experiencia. Esto sirve para determinar si los modelos, variando ambos parámetros, obtienen mejores resultados. De esta manera, se logra tener una estimación sobre que parámetros influyen más en los distintos modelos generados.

Se muestran las tablas 4.3 y 4.4 de total clientes sólo como ejemplo de la forma en que se realiza el experimento y los demás resultados se pueden observar en las tablas del anexo.

Total Snickers	**k**						
p		1	2	3	4	5	6
	1	25.018	14.748	7.086	4.260	2.904	1.732
	2	26.040	9.195	7.169	3.363	2.567	0.601
	3	20.362	9.984	5.685	3.049	0.338	0.389
	4	15.179	5.286	1.547	0.060	0.025	0.021
	5	14.936	5.169	1.207	0.644	0.020	**0.019**

Tabla 4.3: Rendimiento de la función $NNAR(p, 1, k)_{[13]}$ *total Snickers.*

Total Pedigree	**k**						
p		1	2	3	4	5	6
	1	11.005	7.434	4.780	3.286	2.080	1.004
	2	10.032	5.693	2.593	0.960	0.777	0.270
	3	7.866	2.810	1.041	0.331	**0.015**	0.027
	4	8.134	3.688	0.879	0.212	0.686	0.029
	5	9.086	3.759	1.034	0.512	0.157	0.151

Tabla 4.4: Rendimiento de la función $NNAR(p,1,k)_{[13]}$ total Pedigree.

Las tablas 4.3 y 4.4 muestran el error obtenido variando los valores de p y k simultáneamente para el total de clientes para los dos productos. Al igual que en las tablas anteriores, los valores en negrita, son los valores en donde el error es el menor obtenido en el modelo.

4.1.7. Variando la función de activación

Como última parte de la construcción del modelo, se realiza un experimento sobre los valores que entrega el modelo cambiando la función de activación. Hay que tener en consideración que la función elegida en la etapa de modelado es la **sigmoide**, pero se realiza un estudio con una función de activación **lineal** para corroborar que la función escogida es la adecuada para este trabajo. Se construye el modelo variando ambos valores k, p y dejando P como valor 1. Se hace este experimento solamente con el total de clientes, ya que como se comenta anteriormente, esto es para corroborar la elección de la función sigmoide y no de realizar un trabajo paralelo con una función de activación diferente.

Total Snickers	**k**						
		1	2	3	4	5	6
p	1	76.185	56.820	28.445	25.558	23.480	15.223
	2	78.265	28.759	25.755	15.557	13.888	10.430
	3	54.586	31.542	27.563	10.049	9.555	**7.886**
	4	68.756	24.899	15.445	12.025	11.102	10.257
	5	75.256	58.563	15.236	9.605	10.778	10.483

Tabla 4.5: Rendimiento de la función NNAR$(p, 1, k)_{[13]}$ total Snickers para función de activación lineal.

Total Pedigree	**k**						
		1	2	3	4	5	6
p	1	45.409	42.558	23.969	16.501	12.114	9.854
	2	53.751	20.562	21.593	18.332	9.445	10.452
	3	45.250	35.687	12.041	15.536	9.105	**9.022**
	4	56.448	47.852	14.775	18.445	11.572	12.029
	5	51.468	20.735	22.556	13.025	11.554	11.265

Tabla 4.6: Rendimiento de la función NNAR$(p, 1, k)_{[13]}$ total Pedigree para función de activación lineal.

Las tablas 4.5 y 4.6 muestran el error obtenido variando la función de activación a una función lineal, junto con los distintos valores de k y p. El mejor resultado del error

está mostrado en negrita en las tablas. Con estos resultados a simple vista se infiere que la función de activación lineal no posee las mejores características para este modelo, ya que aunque siendo un error razonable, estos no se acercan a los errores obtenidos en las tablas 4.3 y 4.4. Otro punto a destacar es que el modelo disminuye el error al agregar neuronas, pero no hay una real conclusión variando p.

Como último punto hay que recalcar que los valores de errores presentes en todos estos experimentos, son el $MAPE_f$, que equivale al error en el *fitting* de las curvas, comparando las curvas generadas por los modelos con las curvas de los valores reales.

4.1.8. Modelos generados

Con todas las tablas generadas variando los parámetros k y p, en conjunto con las tablas 4.1 y 4.2, se construyen los modelos para cada cliente/producto, en los cuales se indican los parámetros elegidos en relación al menor error obtenido, y que se puede apreciar en la tabla 4.7:

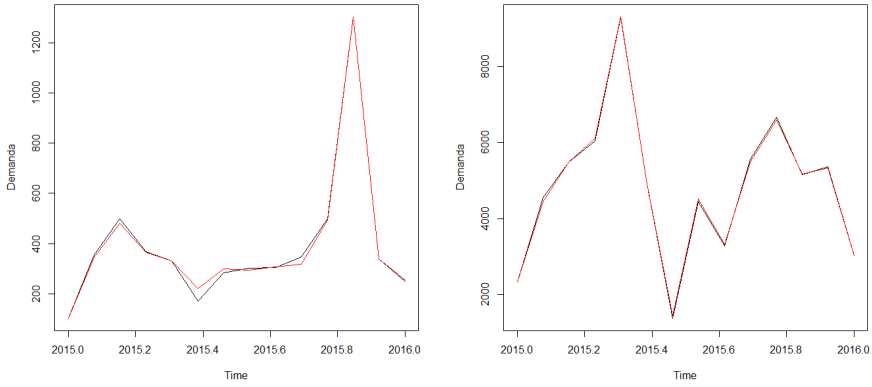
Cliente	Producto	p	P	k	m	$MAPE_f$
Walmart	**Snickers Single**	5	1	6	13	0,021
	Pedigree 15Kg	4	1	5	13	0,012
Cencosud	**Snickers Single**	4	1	5	13	0,013
	Pedigree 15Kg	4	1	6	13	0,010
Promerco	**Snickers Single**	4	1	5	13	0,017
	Pedigree 15Kg	5	1	5	13	0,009
Total	**Snickers Single**	5	1	6	13	0,019
	Pedigree 15Kg	3	1	5	13	0,015

Tabla 4.7: Resultado del modelo en relación al menor error obtenido para la función $NNAR(p, P, k)_{[m]}$.

A partir de esto, se escogen para cada cliente/producto, los parámetros descritos en la tabla 4.7 en la cual se crea el modelo a desarrollar. Para probar cada modelo, se generan los gráficos correspondientes a las figuras 4.1 a 4.4 en relación al *fitting* de cada curva. Hay que tener presente que al ser modelos generados gracias a las RNA, estos pueden variar cada vez que se entrenan, por lo que los errores obtenidos son un reflejo de un entrenamiento en particular y que puede variar en algunas décimas, si se quisiera entrenar los modelo de nuevo.

Los gráficos 4.1 a 4.4 muestran dos curvas en donde se aprecia que la curva en color rojo es la curva creada por el modelo y la curva color negro son los datos reales. En general las figuras anteriores, están representadas muy parecidas a las curvas reales, ya que es difícil apreciar la diferencia entre los colores en los gráficos. Esto se debe al

poco error que se obtuvo de los modelos generados (tabla 4.7).

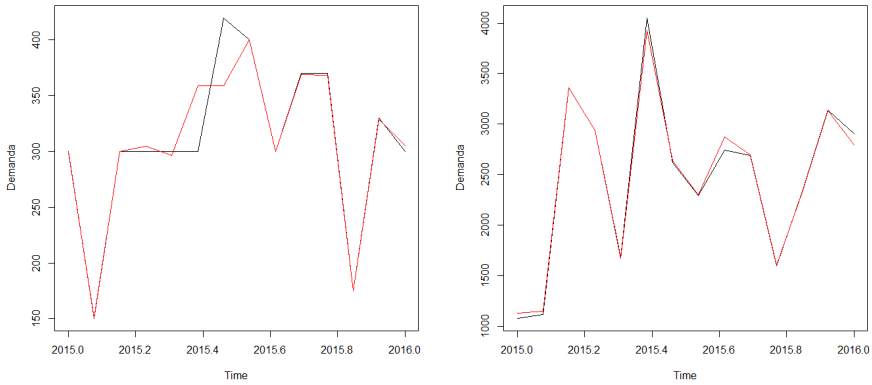


(a) Snickers (b) Pedigree

Figura 4.1: Gráfico del modelo RNA para Walmart.

(a) Snickers (b) Pedigree

Figura 4.2: Gráfico del modelo RNA para Cencosud.



(a) Snickers (b) Pedigree

Figura 4.3: Gráfico del modelo RNA para Promerco.

(a) Snickers (b) Pedigree

Figura 4.4: Gráfico del modelo RNA para total cliente.

4.2. Evaluación de los Modelos

Para poder evaluar los modelos presentados anteriormente, se utilizan las métricas descritas en el plan de prueba. Junto con esto se trabaja con el *forecast accuracy* para obtener el valor que se exige dentro de los objetivos del proyecto, el cual indica qué tan acertada es la demanda de las predicciones obtenidas.

La evaluación de los modelos, seguirá la siguiente forma:

- Generar la predicción con el modelo descrito.
- Obtener métricas de las predicciones generadas.
- Calcular el *forecast accuracy* de las predicciones generadas.

Utilizando los modelos generados anteriormente, se tienen las curvas de las figuras 4.5 a 4.12 en las cuales: la curva de color rojo, representa la demanda real, y la curva de color azul la predicción hecha por el modelo, junto con las métricas y *forecast accuracy*. Dentro de las métricas el valor en negrita es el $MAPE_p$ el cual será la métrica elegida y se utilizará en toda la etapa de evaluación para calcular el *forecast accuracy*.

Para una mejor lectura, se separa esta parte de la evaluación, en todos los modelos generados a partir de la etapa anterior en conjunto de los errores obtenidos y sus respectivos gráficos.

4.2.1. Evaluación del modelo Walmart Snickers $NNAR(5,1,6)_{[13]}$.

Modelo	ME	RMSE	MAE	MPE	MAPE
Walmart Snickers	94.866	228.539	267.419	2.014 %	**81.663 %**

Tabla 4.8: Resultados de las métricas para evaluar el modelo Walmart Snickers

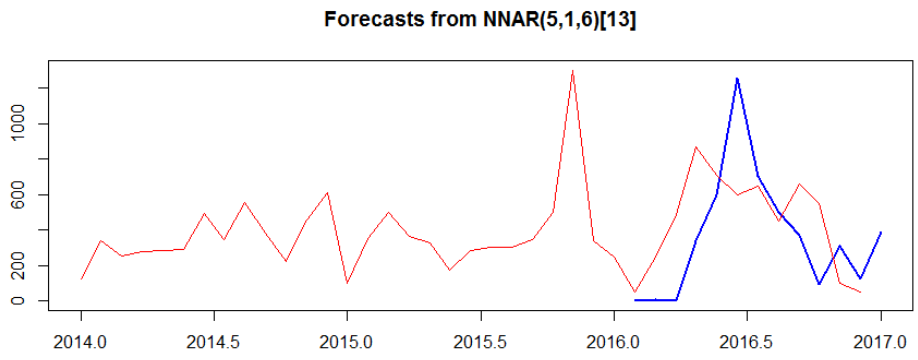


Figura 4.5: Predicción del modelo Walmart Snickers.

De las figura 4.5 se desprende que en el modelo generado para Walmart, el resultado está bajo los objetivos del proyecto, muy lejano al 70 %. El resultado de este cliente es un caso interesante de analizar, ya que aunque no cumple con el objetivo, este valor es el peor valor obtenido en los modelos de predicción, por lo que es motivo de estudio más allá del modelo, si no del trabajo realizado específicamente con los datos.

Dado lo anterior, y entendiendo el contexto en que se desenvuelve la organización, el comportamiento del cliente y del producto, este resultado tan lejano tiene directa relación con los ciclos promociónales que tiene el cliente propiamente tal. Walmart es el cliente más importante para la compañía y para la mayoría de las multinacionales del país. Estos ciclos promociónales se llevan a cabo en todas las categorías de productos y en las que sólo participa un proveedor por cada categoría, lo que significa que en muchas oportunidades se debe tomar la decisión de participar o no, a pesar de que muchas veces los descuentos que se exigen están muy por sobre lo que puede soportar el estado de resultados para obtener ganancias. No obstante, se consigue un bloqueo de la competencia y se mantiene la participación de mercado y el peso de la marca en el cliente. Estos ciclos promociónales generan valores atípicos muy fuertes en los periodos de ventas, impactando negativamente al modelo de predicción. En estos periodos, la demanda está muy por sobre lo normal y los volúmenes de venta se elevan a tal punto en que se triplica e incluso se cuadriplica el volumen, impactando directamente en el mes en curso y en los meses posteriores.

4.2.2. Evaluación del modelo Cencosud Snickers $NNAR(4,1,5)_{[13]}$.

Modelo	ME	RMSE	MAE	MPE	MAPE
Cencosud Snickers	28.144	49.354	40.301	12.571 %	**15.667 %**

Tabla 4.9: Resultados de las métricas para evaluar el modelo Cencosud Snickers

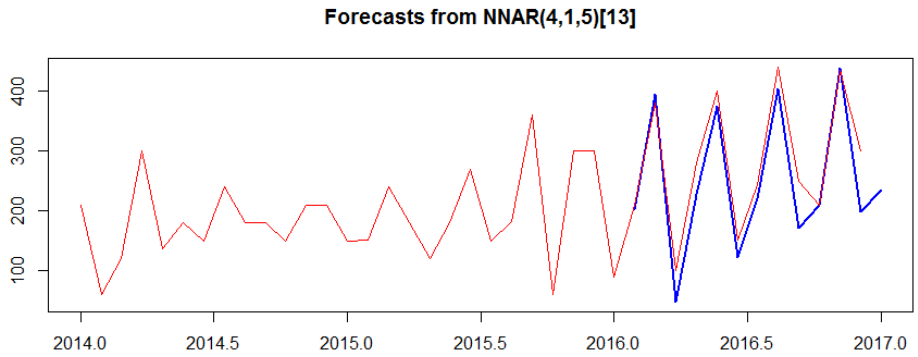


Figura 4.6: Predicción del modelo Cencosud Snickers.

De la figura 4.6 se ve como el modelo en Cencosud alcanza un valor por sobre el 80 %. Dado lo anterior, el resultado entregado para Cencosud es un buen indicador porque actualmente la compañía para este producto/cliente en específico se encuentra en torno al $60 - 65$ %. En la realidad, la prueba del modelo con este resultado puede ser desechado o aceptado, dependiendo de la decisión que tome el área responsable del proceso de predicción, comparando este modelo con su forma de predecir la demanda. La tendencia debería inclinarse a aceptar el modelo dado el resultado obtenido en Cencosud y el hecho de que este cliente represente de mejor forma al resto de los clientes.

4.2.3. Evaluación del modelo Walmart Pedigree $NNAR(4,1,5)_{[13]}$.

Modelo	ME	RMSE	MAE	MPE	MAPE
Walmart Pedigree	-341.836	1354.834	709.701	-14.940 %	**17.982 %**

Tabla 4.10: Resultado de las métricas para evaluar el modelo Walmart Pedigree

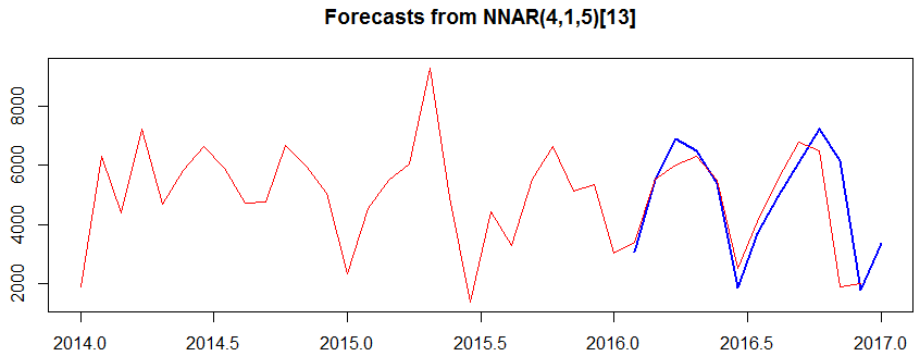


Figura 4.7: Predicción del modelo Walmart Pedigree.

4.2.4. Evaluación del modelo Cencosud Pedigree $NNAR(4,1,6)_{[13]}$.

Modelo	ME	RMSE	MAE	MPE	MAPE
Cencosud Pedigree	70.575	1636.370	1265.939	2.349 %	**17.622 %**

Tabla 4.11: Resultado de las métricas para evaluar el modelo Cencosud Pedigree

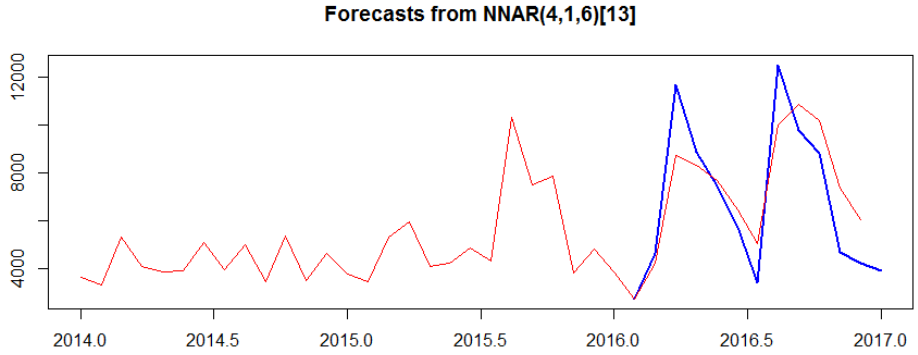


Figura 4.8: Predicción del modelo Cencosud Pedigree 15Kg.

Siguiendo los resultados para Pedigree y Snickers en el cliente Cencosud, el modelo cuenta con un $MAPE_p$ aceptable para ambos casos, cumpliendo satisfactoriamente

con el objetivo. Estos resultados a diferencia del anterior, muestran una realidad completamente distinta. El producto en cuestión, se desenvuelve en una categoría en que la marca tiene la mayor participación de mercado y por ende el mayor peso de venta. Por lo anterior, al ser un producto indispensable para el cliente, el comportamiento de este producto, es estable a lo largo de los periodos, y el impacto en venta por ciclos promocionales no afecta el desempeño del producto, lo cual se ve reflejado en los gráficos 4.7 y 4.8, apreciando una curva cíclica de cada medio año, lo cual se ve bien reflejado en el modelo creado.

4.2.5. Evaluación del modelo Promerco Snickers $NNAR(4,1,5)_{[13]}$.

Modelo	ME	RMSE	MAE	MPE	MAPE
Promerco Snickers	-6.234	46.211	36.212	-4.3879 %	**10.570 %**

Tabla 4.12: Resultados de las métricas para evaluar el modelo Promerco Snickers

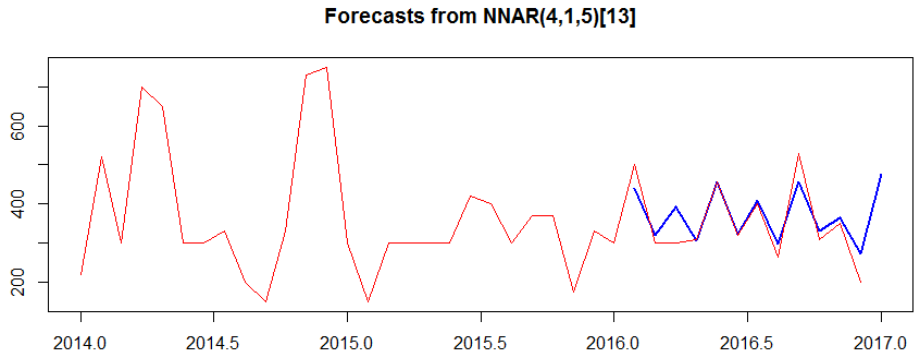


Figura 4.9: Predicción del modelo Promerco Snickers Single.

4.2.6. Evaluación del modelo Promerco Pedigree $NNAR(5,1,5)_{[13]}$.

Modelo	ME	RMSE	MAE	MPE	MAPE
Promerco Pedigree	72.554	549.177	402.220	2.377 %	**14.222 %**

Tabla 4.13: Resultados de las métricas para evaluar el modelo Promerco Pedigree

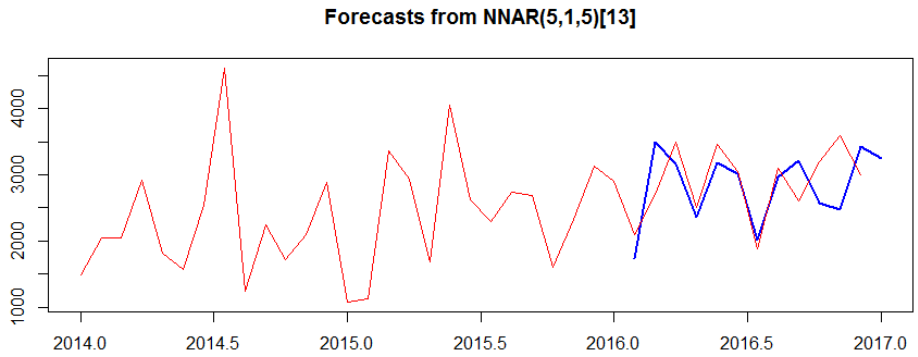


Figura 4.10: Predicción del modelo Promerco Pedigree 15Kg.

En ambos productos la predicción realizada y el resultado entregado por el $MAPE_p$ está muy por sobre el objetivo propuesta en el proyecto. Es muy importante destacar que este cliente en particular, está categorizado en un canal de venta distinto a los ejemplos anteriores: Promerco se considera parte del canal tradicional. Promerco es un distribuidor en la que una empresa familiar es dueña y que maneja todo el negocio. Esto hace que se tenga menor disponibilidad de la información, tanto de ventas como de clientes, dificultando el trabajo del día a día, buscando las oportunidades que permitan hacer crecer el negocio y no tener una visibilidad tan clara de la percepción de los clientes y consumidores finales de esta cadena. Dado lo anterior, desde un principio la incorporación de este cliente al proyecto, iba a entregar información relevante en datos de predicción. Al ser un cliente más informal, en términos de automatizado de pedidos

o ventas, y al ser quien distribuye a clientes pequeños como almaceneros, botillerías y locales de venta al por menor, la demanda es más bien incierta en todos los periodos del año.

Pese a lo anterior, los resultados de este cliente con el modelo creado, sobrepasan las expectativas, ya que la idea inicial era que probablemente el modelo no se ajustaría a los objetivos impuestos para predecir su demanda. Hay que comentar además, que a partir de los gráficos 4.9 y 4.10 la demanda de ambos productos tiene altos y bajos pero son, en cierta medida, naturales dentro de un periodo. Estos altos y bajos tienen directa relación con la forma en que se generan las compras de Promerco, ya que hace un pedido grande y al periodo siguiente uno más reducido. Este comportamiento, hace que las curvas estén siempre moviéndose en un mismo rango sin afectar en gran medida al modelo, donde el motivo principal, es que este es capaz de captar estos movimientos sin grandes dispersiones y datos fuera de lo normal.

4.2.7. Evaluación del modelo Total Snickers $NNAR(5,1,6)_{[13]}$.

Modelo	**ME**	**RMSE**	**MAE**	**MPE**	**MAPE**
Total Snickers	24.158	276.349	200.948	6.930 %	**27.172 %**

Tabla 4.14: Resultados de las métricas para evaluar el modelo Total Snickers

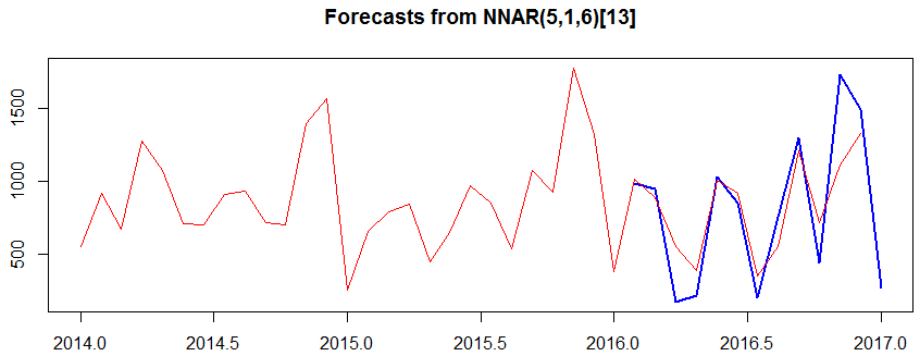


Figura 4.11: Predicción del modelo Total Snickers Single.

4.2.8. Evaluación del modelo Total Pedigree $NNAR(3,1,5)_{[13]}$.

Modelo	ME	RMSE	MAE	MPE	MAPE
Total Pedigree	-849.949	3319.143	2834.823	-8.209 %	**21.176 %**

Tabla 4.15: Resultados de las métricas para evaluar el modelo Total Snickers

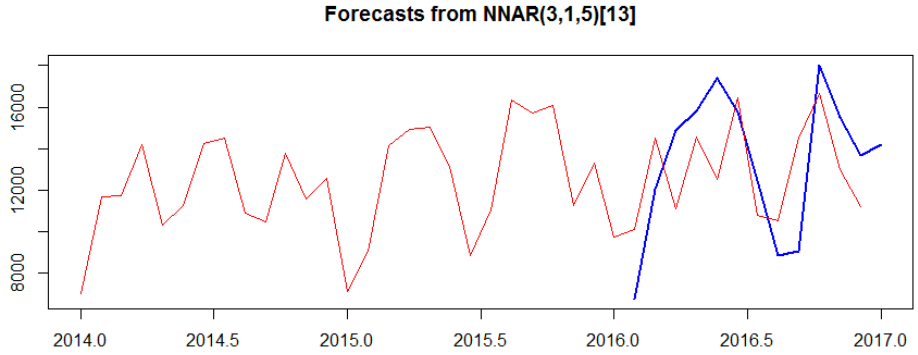


Figura 4.12: Predicción del modelo Total Pedigree 15Kg.

Finalmente, dentro del estudio está también el incluir en esta experiencia a todos los clientes de la compañía, con el fin de realizar predicciones a gran escala, que impacten

en las decisiones importantes de la compañía en relación a los objetivos monetarios y las metas impuestas desde países externos. Estas decisiones tienen relación como por ejemplo: qué productos son los que se venderán más en periodos posteriores, cuáles son los clientes o canales de comercialización que se deben potenciar o qué en términos financieros es lo que más conviene.

Con la suma de todos los clientes se obtiene un modelo para ambos productos, para el total de clientes, este modelo presente en los gráficos 4.11 y 4.12 muestran cómo estos productos tienen ciclos definidos a simple vista, en donde hay un aumento de demanda que en general se ve reflejado en los periodos finales del año. Hay que mencionar que el *forecast accuracy* para ambos productos están sobre los valores de los modelos anteriores pero de todas maneras se encuentran por sobre el objetivo del proyecto. Este valor tiene directa relación a la forma en que se desarrollan estos modelos, ya que como se sabe, los clientes como un todo, tienen distintas características, una de esas que ya se nombraron anteriormente como el canal de distribución (canal tradicional y supermercado) o la categoría en que se desenvuelve el producto. Esta diferencia entre los clientes da como resultado un aumento en el $MAPE_p$, comparando a varios de los clientes por si solos.

En general el problema principal visible en la compañía es alcanzar las metas impuestas por los altos mandos, en donde siempre se pide un porcentaje deseado por ellos, pero no con un estudio previo de las ventas, contexto y niveles de inventario. Esto se ve reflejado en el gráfico 4.5 en la etapa de evaluación del modelo en donde se aprecia claramente que los últimos periodos del 2016 hay un alza significativa de la demanda en relación a los demás datos, lo que conlleva a ser considerado como un *outlier* que podría llevar a un error al momento de la creación del modelo.

Al apreciar ambos productos para el total de clientes, se muestra que Snickers tiene una predicción más cercana al valor mínimo esperado, a diferencia de Pedigree. Lo anterior, se debe a factores que impactan directamente a la categoría en que se desenvuelve el producto:

- Varios periodos de estacionalidad que impactan en la tendencia de demanda.
- Puesta en marcha de alguna legislación en el país, como el caso de hoy en día de la Ley de etiquetado 20,606, que regulariza el marketing y comercialización de ciertos productos aplicándoles algunos sellos a los empaques y haciendo desaparecer ciertos íconos de la marca en la publicidad.
- Problemas al momento de trabajar con la calidad de los datos. Los *outliers* pueden ser una razón para que estos valores no sean los deseados o algún error en la calidad de los datos que influya en gran medida a estos valores deficientes. Esta es una consecuencia de los puntos nombrados anteriormente.
- Problemas de quiebre de *stock*, por la lejanía de las plantas de producción, lo que repercute a un disminución y luego un aumento significativo de la demanda en periodos posteriores.

Para los modelos generados gracias al *dataset* de la tabla 4.7, se tienen un resumen de *forecast accuracy* calculados gracias a la resta del 100 % menos el valor $MAPE_p$. La tabla 4.16 muestra la evaluación de los modelos en relación al *forecast accuracy*, que es en el fondo el objetivo propuesto en este trabajo.

Cliente	**Producto**	***Forecast Accuracy***
Walmart	**Snickers Single**	18,337 %
	Pedigree 15Kg	82,018 %
Cencosud	**Snickers Single**	84,333 %
	Pedigree15Kg	82,378 %
Promerco	**Snickers Single**	89,430 %
	Pedigree 15Kg	85,778 %
Total	**Snickers Single**	72,828 %
	Pedigree 15Kg	78,824 %

Tabla 4.16: Evaluación respecto forecast accuracy.

4.2.9. Evaluación global de los modelos

Para evaluar estos modelos en general, se toman en cuenta dos criterios para poder valorarlos:

- Objetivos: precisión del modelo.
- Subjetivos: interpretación de resultados.

En general, todos los resultados cumplen con un nivel de predicción aceptables y que están por sobre el 70 % cumpliendo con el objetivo inicial de la compañía.

Es importante destacar que de las pruebas realizadas, sólo hay un caso (Walmart Snickers Single), donde el *forecast accuracy* no alcanza a cumplir con el mínimo valor objetivo. Sin embargo, el valor está muy cercano al deseado y por lo mismo, se tomará como un modelo correctamente generado pero con resultados insuficientes.

Para interpretar los resultados obtenidos, primero hay que agrupar los diferentes resultados de los modelos. A simple vista, la agrupación ideal viene dada por cliente, ya que los resultados de *forecast accuracy* por un cliente y sus productos, tienden a ser parecidos como valor. No obstante lo anterior, aparece otra vez reflejado el valor del cliente Walmart Snickers Single que sale de los parámetros de agrupación ya que no coincide con el valor obtenido por el mismo cliente con el producto Pedigree 15Kg. Estos resultados, serán estudiados con profundidad en la evaluación de los resultados que vienen en las páginas siguientes.

Otra interpretación es generada al agrupar los tres clientes y por otra parte, el valor en el total de clientes. En el primero, se establece un valor de *forecast accuracy* alrededor de 85 %, en cambio, en el segundo, el valor es reducido a un promedio de 75 %. Esto es relevante para estudiar en las páginas siguientes, ya que hipótesis inicial es que en total de clientes el *forecast accuracy* debería ser muy similar al valor de los clientes por separado.

Gracias a los resultados obtenidos en la sección anterior se puede confirmar que la mejor forma de evaluar los resultados de estos modelos, es con el *forecast accuracy*, y

lograr los resultados esperados por sobre el 70 %. Este será el valor a comparar con los resultados siguientes.

Es importante destacar que la evaluación de los resultados van en directa relación con los valores obtenidos en la sección anterior. Esto se debe a que al tener los resultados de las predicciones hechas por el modelo y compararlas con los valores reales, se obtienen resultados muy cercanos entre sí. Con esto, es posible integrar los modelos como un proyecto real a corto plazo en la empresa.

Para explicar los resultados, esta evaluación se va a separar en distintas partes, intentando evaluar los resultados por cada modelo creado, teniendo en cuenta las características de cada uno de esos clientes y productos que influyen en que los modelos entreguen o no buenos resultados, delimitadas por el contexto en el que se ha desarrollado la empresa a lo largo de los años, su estrategia y su comportamiento.

Como se menciona, este modelo es un modelo orientado a la toma de decisiones tanto globales como particulares y puede ser aplicable a cualquier ámbito en que la compañía lo requiera.

Gracias a los resultados obtenidos es una buena idea utilizar estos modelos para poder tener un mejor *forecast accuracy*, ya que actualmente no existe una herramienta sólida en la empresa, que permita obtener mejores resultados. Los errores que presentan estos modelos son generalmente buenos y sobrepasan las expectativas que se tenían en el desarrollo del modelo.

4.3. Despliegue

4.3.1. Plan de implementación

Para implementar este proyecto en la compañía el primer paso es tener acceso siempre que se requiera a la base de datos que contiene toda la información relativa a la demanda de los clientes. A partir de ahí, los pasos a seguir serán los mismos que se han descrito anteriormente, donde probablemente el tiempo empleado en las primeras fases

de recolección, comprensión y preparación de los datos sea considerablemente mayor, dado un número muy alto de registros de demanda, y donde se encontrarán demandas irregulares, *outliers* y posiblemente problemas con la calidad de los datos.

Como primera etapa de implementación se presentarán los modelos propuestos en este trabajo y se trabajará con estos para ir obteniendo las predicciones futuras. Logrando un buen desempeño con estos modelos en la empresa, viene la segunda etapa que es decidir cuáles productos vale la pena estudiar para realizar nuevos modelos, haciendo que esta sea la herramienta utilizada por la compañía para predecir la demanda de la mayor cantidad de productos teniendo en cuenta el tiempo y recursos que lleva realizar estas nuevas predicciones. Este análisis va de la mano con los productos más importantes de la compañía y de aquellos que se quiera potenciar, donde muy probablemente sean aquellos que significan la mayor cantidad de ventas.

4.3.2. Monitoreo y mantenimiento

Es la etapa más importante para aquellos casos en que los resultados entregados son parte del negocio y van a servir para las predicciones futuras. Es por esto, que como plan de monitoreo y mantenimiento se definen algunas acciones a realizar:

- Extracción y almacenamiento de los datos de demanda de la compañía, en un archivo local único para la predicción, con los formatos establecidos y los datos necesarios para realizar este proyecto.
- Dado que para la predicción se consideró la demanda por periodos de los años 2014 a 2016, esta base debe ir siendo actualizada mensualmente para ir generando las predicciones posteriores.
- Comparación mensual de la demanda real con la predicción generada por el modelo.

Con este plan de monitoreo se logrará que la demanda se ajuste de mejor forma gracias a que se utilizarán los modelos de predicción propuestos, logrando que en el futuro los datos obtenidos sean un mejor reflejo del comportamiento de la demanda y

que las decisiones humanas tomen menor peso, ya que los modelos mostrarán valores acotados a la realidad. Esto tiene como consecuencia, que la empresa trabajará con valores para poder asignar metas más ajustadas a los vendedores o metas de crecimiento anual.

El mantenimiento es crucial, ya que estos modelos tienen que ir actualizándose a medida que crezca la cantidad de datos históricos, donde sería ideal ir probando los modelos propuestos cada año, así se tendrá una mejor aproximación de cómo va cambiando la demanda a través del tiempo, lo que ayudaría a hacer los modelos más exactos y actualizados a la realidad futura.

5. Conclusiones

Antes de comenzar con este proyecto, Mars hacía predicciones de demanda revisando periodos y años anteriores, sin evaluar detenidamente la calidad y veracidad de los datos entregados y no se consideraban eventualidades particulares que impactaban en las predicciones que se hacían. Ahora bien, el nuevo modelo de predicción permite realizar una estimación más real y cercana que permita verdaderamente a la compañía manejar tendencias reales y desafiantes, que permitan confiar en la calidad de los datos y automatizar cada vez más los procesos de predicción y de análisis, donde la minería de datos toma un gran protagonismo.

La minería de datos es indispensable al momento de poder crear buenos modelos, debido a que es imposible llegar y tomar los datos e introducirlos en alguna herramienta que genere predicciones, sin evaluarlos y trabajarlos antes. En general, hay muchas herramientas en el mercado para poder hacer predicciones pero es muy difícil que estas entreguen un resultado esperado, dado los procesos que se manejan en estas herramientas en relación a los datos que llevan a encaminar de buena forma estas herramientas y modelos. La comprensión de los datos es la etapa indispensable, ya que al tener un estudio sobre estos, es más fácil agruparlos, darles algún significado e incluso entender su comportamiento. Esto hace que esta etapa influya a lo largo de todo el proyecto, siendo la más importante, contribuyendo a obtener datos de óptima calidad junto con poder desarrollar modelos que generen valor distintivo.

Para el caso particular de este trabajo la comprensión de los datos fue la etapa más difícil de desarrollar, ya que había un gran número de datos dispersos en todas las áreas de trabajo. Por esto, optó por rescatar toda la información desde SAP y trabajar como base e incluir de a poco, luego de estudios y reuniones con los encargados, los distintos archivos de datos que hacían referencia con la base de SAP.

Las muestras escogidas sobre qué clientes y productos aplicar en el modelo, están basadas en la importancia de venta de cada uno de ellos y del conocimiento que tenían las distintas áreas funcionales de la organización. La muestra con la que se trabajó

reflejaba la venta de más de la mitad de la compañía y por lo mismo, los resultados que se reflejarían serían decisivos a la hora de la aplicación del modelo y del futuro de éste. Separarlos a su vez por categoría y canal de distribución, ayudaría a comprender cuáles eran las áreas de oportunidad a trabajar y cómo impactaba directamente al modelo de predicción.

Para poder utilizar RNA existen un sin número de herramientas junto con bibliotecas y trabajos de todo tipo en donde se utiliza este modelo. En general, el trabajo con RNA utilizándolas para la predicción tienden a ser más simples de lo que parece, ya que como todos los modelos están creados y programados, solo se necesita cambiar parámetros de estos para ir viendo cómo estos ayudan a obtener las predicciones y soluciones a los problemas. Las RNA son muy complejas en su comprensión pero si sólo se utilizan como herramientas para la predicción, pueden ser entendibles superficialmente y aplicarlas fácilmente para poder elaborar diferentes actividades en la minera de datos. Las RNA fueron las escogidas para predecir la demanda, siendo que hay varios otros modelos que también entregan estos resultados como lo es ARIMA o VAR, estas lograron cumplir con las expectativas del proyecto, haciendo predicciones confiables con una curva de aprendizaje aceptable.

Para los modelos obtenidos se hicieron varios análisis y pruebas que se presentaron a lo largo de este informe, en el que se hace énfasis en la búsqueda del objetivo principal del proyecto y de cómo este resultado impactaría positivamente el funcionamiento de la compañía. Se consiguió un resultado satisfactorio, que sitúa a la exactitud de predicción de demanda o al principal indicador *forecast accuracy* por sobre el 80 % y por consiguiente un resultado de $MAPE_p$ por debajo del 20 %, ajustándose correctamente a la historia predictiva de la compañía, permitiendo entregarle una herramienta de fácil uso, que podrá integrar diversas áreas de la compañía y encontrar eficiencias en todos los procesos que se lleven a cabo.

Al ser este el objetivo, trabajar con los modelos y parámetros en un sentido más analítico, daría como resultado que los experimentos tomaran mucho tiempo y recursos

para buscar el modelo óptimo. Los modelos presentados son modelos de buena calidad pero puede suceder que si evaluamos el universo de modelos, los seleccionados no sean los más indicados para la predicción. Aún así, los objetivos son satisfechos en un tiempo razonable sin profundizar en el funcionamiento del modelo a través de la programación en R.

Este modelo y proyecto inicial es una pauta para las próximas tomas de decisiones de la compañía que permitirá entregar la planificación necesaria para poder agilizar la toma de decisiones y que por sobre todo, esas decisiones sean de calidad, que permitan entregar confianza y generar paso a paso la costumbre de utilizarla por todas las áreas que se ven involucradas día a día. Sin embargo, se puede adelantar que usando esta herramienta de predicción y cambiando poco a poco, algunas de las prácticas que normalmente se llevan a cabo, las curvas de demanda deberían normalizarse y disminuir cada vez más los *outliers* que se iban generando en las pruebas del modelo y que eran el principal problema en la calidad de los datos obtenidos. Una vez que la empresa pueda reducir la cantidad significativa de *outliers* que se vieron involucrados en el proceso de revisión de la calidad de los datos, estos últimos mejorarán a tal punto, que la herramienta de predicción podrán utilizarla como una herramienta casi exacta de la realidad.

Finalmente, como se dijo en páginas anteriores, el proyecto puede y es aconsejable expandirlo a todas las áreas de la empresa que tengan directamente relación con las ventas de la compañía. El proceso inicial será largo, puesto que habrá que evaluar cada modelo, pero en el largo plazo, se traerán beneficios importantes en la compañía, no solo en términos monetarios, sino que también en recursos de tiempo y personas. La cantidad de personas que se dedican hoy a procesar la cantidad de información y el tiempo que se dedicaba a generar modelos que permitieran predecir, afectaba directamente a la eficiencia de la compañía, exactitud y veracidad de la información. Este ejercicio anterior, se vería reducido además automatizando a través de Azure Machine Learning, donde la predicción sería casi inmediata luego de haber diseñado los mode-

los.

Dentro de las futuras fases de este trabajo, hay una idea que se estudió desde el inicio del proyecto, la cual es poder obtener una predicción semanal de la demanda. Esta idea se descartó al tener una visión de los datos, ya que muchos clientes realizan una sola compra grande en el periodo. Para los clientes es una vez al mes, para la compañía es en la tercera semana de cada periodo, por lo que cuando se hace la apertura a nivel de semana, los resultados arrojados están muy por debajo de lo que podría esperarse. Los valores de la demanda de los clientes, en gran medida, se mantienen en valores 0 en la mayoría de las semanas del periodo Mars, haciendo que los algoritmos, obtengan resultados erróneos. Al poder tener una predicción semanal, ayudaría a tratar de mejor manera la demanda de los clientes, ya que se podría ir abasteciendo parcialmente, logrando una mejor eficiencia a nivel de cliente y a nivel de bodega (menos productos en bodega es un ahorro de recursos para la compañía). Esta idea cobra fuerza al ver los resultados obtenidos en este trabajo, ya que mejorando el forecast accuracy, se puede dimensionar realmente el impacto de las ventas por periodos y las consecuencias de una mala predicción. Sin embargo, esto es un cambio de comportamiento que debe liderar la compañía, entrenando a los clientes y a la fuerza de ventas para poder poner en práctica esta nueva idea, y por consiguiente, tener como resultado el nivel de predicción esperado, con lo cual los datos futuros serían valores capaces de incluir en el algoritmo, haciendo que el modelo anterior utilizado, lograra obtener la predicción deseada.

6. Anexo

Walmart Snickers	**k**						
p		1	2	3	4	5	6
	1	30.634	18.512	10.016	6.277	4.147	2.975
	2	26.988	14.825	10.698	5.946	3.613	1.580
	3	26.784	13.657	4.752	0.821	1.868	0.115
	4	26.020	11.795	3.192	0.621	0.711	0.036
	5	24.064	11.274	4.044	0.417	0.036	**0.021**

Tabla 6.1: Rendimiento de la función $NNAR(p, 1, k)_{[13]}$ Walmart Snickers.

Walmart Pedigree	**k**						
p		1	2	3	4	5	6
	1	20.580	14.505	9.948	4.924	1.454	1.314
	2	16.689	7.949	2.944	0.483	0.291	0.078
	3	18.742	10.623	3.215	0.757	0.040	0.176
	4	11.116	2.947	1.652	0.294	**0.012**	0.024
	5	11.716	3.383	1.440	0.150	0.014	0.015

Tabla 6.2: Rendimiento de la función $NNAR(p, 1, k)_{[13]}$ Walmart Pedigree.

Cencosud Snickers	**k**						
		1	2	3	4	5	6
p	1	23.535	20.483	10.018	5.086	3.334	1.119
	2	16.691	7.949	4.097	0.792	0.026	0.020
	3	17.232	10.623	2.622	0.082	0.017	0.028
	4	14.691	3.730	0.979	0.137	**0.013**	0.081
	5	14.525	3.217	0.796	0.515	0.014	0.016

Tabla 6.3: Rendimiento de la función $NNAR(p,1,k)_{[13]}$ Cencosud Snickers.

Cencosud Pedigree	**k**						
		1	2	3	4	5	6
p	1	15.196	8.119	4.200	2.822	2.335	1.729
	2	10.650	4.681	1.685	0.564	0.322	0.283
	3	9.799	3.105	1.020	0.186	0.064	0.010
	4	6.231	3.170	0.446	0.042	0.025	**0.010**
	5	9.786	3.739	1.475	0.012	0.042	0.014

Tabla 6.4: Rendimiento de la función $NNAR(p,1,k)_{[13]}$ Cencosud Pedigree.

Promerco Snickers	**k**						
p		1	2	3	4	5	6
	1	17.042	9.288	5.787	4.176	2.895	2.755
	2	9.401	5.494	3.075	2.459	2.402	2.381
	3	8.460	5.118	3.222	2.525	2.380	2.381
	4	7.516	2.203	1.207	0.043	**0.017**	0.021
	5	7.593	2.521	0.745	0.054	0.030	0.024

Tabla 6.5: Rendimiento de la función $NNAR(p,1,k)_{[13]}$ Promerco Snickers.

Promerco Pedigree	**k**						
p		1	2	3	4	5	6
	1	25.752	17.359	5.787	10.068	4.850	2.984
	2	22.926	9.830	6.639	2.707	0.908	0.059
	3	17.402	11.046	4.338	1.385	0.128	0.016
	4	15.945	7.331	3.144	0.290	0.229	0.115
	5	14.207	7.980	2.401	0.705	**0.009**	0.013

Tabla 6.6: Rendimiento de la función $NNAR(p,1,k)_{[13]}$ Promerco Pedigree.

Referencias Bibliográficas

[1] George Canavos. *Probabilidad y estadística: Aplicaciones y métodos*. McGRAW-HILL / INTERAMERICANA DE MEXICO, 2017.

[2] José Alberto Gallardo Arancibia. Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. 2009.

[3] Juan Miguel Moine, Silvia Gordillo, and Ana silva Haedo. Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. *CACIC 2011 - XVII CONGRESO ARGENTINO DE CIENCIAS DE LA COMPUTACIÓN*, 2011.

[4] Damián Jorge Matich. Redes Neuronales: Conceptos Básicos y Aplicaciones. *Universidad Tecnológica Nacional*, 2001.

[5] Juan Camilo Santana. Predicción de series temporales con redes neuronales: una aplicación a la inflación colombiana. *Revista Colombiana de Estadística*, 29:77–92, 2006.

[6] Juan Miguel Jiménez Panda. Prónostico de demanda de llamadas en los call center, utilizando redes neuronales artificiales. *Universidad de Piura, Facultad de Ingeniería Industrial y de Sistemas*, 2003.

[7] Hector Tabares, John Branch, and Jaime Valencia. Generación dinámica de la topología de una red neuronal artificial del tipo perceptron multicapa. *Universidad Nacional de Colombia, Escuela de Sistemas*, 2006.

[8] Federico Peralta. Elementos para un mapa de actividades para proyectos de explotación de información. *Escuela de postgrado facultad regional Buenos Aires tecnológica nacional*, 2013.

[9] Documentación Microsoft Azure Machine Learning. https://docs.microsoft.com/es-es/azure/machine-learning/machine-learning-what-is-machine-learning. Consulta: 10 Marzo 2017.

[10] Thomas Rahlf. *Data Visualization with R*. Springer International Publishing, 2017.

[11] W.N.Venables and B.D. Ripley. *Modern applied Statistics with S*. Springer International Publishing, 2002.

[12] Raúl González Duque. *Python para todos*.

[13] Ruben Thoplan. Simple v/s Sophisticated Methods of Forecasting for Mauritius Monthly Tourist Arrival Data. *International Journal of Statistics and Applications*, 2014.

[14] Juan D. Velásquez, Yris Olaya, and Juan D. Velásquez. Predicción de series temporales usando máquinas de vectores de soporte. *Ingeniare*, 18:64–75, 2009.

[15] Prapanna Mondal, Labani Shit, and Saptarsi Goswami. Study of Effectiveness of Time Series Modeling (ARIMA) in Forecasting Stock Prices. *International Journal of Computer Science, Engineering and Applications (IJCSEA)*, 4, 2014.