Toma de decisiones para reposición de stock basado en Business Intelligence y Machine Learning

A. Garcete, R. Benítez , D. Pinto-Roa, A. Vazquez

***Abstract*—La gestión de compras en empresas retail supone uno de los procesos más importantes que tiene un impacto económico en toda la organización. Dentro de la gestión de compras, una decisión que regularmente debe tomarse es acerca del volumen a adquirir de un producto determinado para así reponer el stock, o si realmente hay que seguir adquiriendo dicho producto. En este trabajo se propone un modelo que ayuda a tomar las decisiones del volumen de compra de productos disponibilizados para la venta, a través de técnicas de Business Intelligence y Machine Learning integradas. Los experimentos indican que se puede automatizar la compra de productos para reposición de stock a través del modelo propuesto en este trabajo.**

***Index Terms*—retail, compras, business intelligence, machine learning.**

I. Introducción

En las empresas retail o de ventas minoristas uno de los principales problemas con que se enfrentan es el manejo eficiente del stock de manera a evitar tener los productos en exceso en los depósitos que incurran en sobrecostos, o en el otro extremo la falta de dichos productos o ruptura de stock lo cual conlleva a pérdidas de oportunidades de ventas por no disponer del producto que puede generar insatisfacción de los clientes y a su vez repercute en las utilidades de la empresa. Uno de los mayores desafíos de las empresas es la de estimar o predecir la cantidad de ventas para el próximo periodo de tiempo.

Actualmente en el proceso de gestión de compras en empresas como las retail, se utilizan algunas de las técnicas de pronósticos para determinar las cantidades de las órdenes de compra, las cuales pueden estar basadas en pronósticos cuantitativos o cualitativos. Los modelos de cantidad fija y los modelos de periodo fijo son ampliamente utilizados. Independientemente de la técnica elegida, el problema real de los pronósticos es su falta de confiabilidad, ya que por lo general no son precisos, entonces, la interrogante que siempre surge es si serán superiores o inferiores a la demanda real y en qué medida.

Con el presente trabajo se elabora un nuevo modelo de estimación de cantidades eficientes en las órdenes de compra de productos para la reposición de stock del siguiente periodo de venta. En este nuevo modelo se integran técnicas de Business Intelligence y Machine Learning.

En la etapa de Business Intelligence el objetivo principal es calcular los Indicadores Claves de Rendimiento (KPI - Key Performance Indicators) de los productos en base a los datos históricos obtenidos de la base de datos transaccional. Luego cada serie de KPI obtenidos pasan por un proceso de etiquetado, donde el experto en compras los analiza y determina qué nivel de compra conviene para cada serie de KPI.

En la etapa de Machine Learning se utiliza como entrada las series de KPI obtenidos en la etapa de Business Intelligence y que constituyen las instancias que alimentan los distintos algoritmos de clasificación de Machine Learning supervisado. Luego tienen lugar los procesos propios de esta etapa que son el entrenamiento y testeo para finalmente evaluar los distintos desempeños a fin de determinar los algoritmos más adecuados que serán utilizados para estimar las cantidades de las ordenes de compra por cada producto.

En cuanto a las limitaciones de este nuevo modelo se puede mencionar que no toma en cuenta los costes relacionados al inventario: como costes de mantenimiento, de personal, seguros, etc. El modelo planteado se aplica a empresas retail dedicadas a las venta de productos terminados.

En la segunda sección se abordan los conceptos de Administración de Compras. En la tercera sección se definen los conceptos de Business Intelligence y su modelado del problema. La cuarta sección aborda los conceptos de Machine Learning. En la quinta sección se realiza el modelado con Machine Learning y se analizan los resultados experimentales. En la última sección se realiza la conclusión general.

II. Administración de las compras

II-A. Introducción a las compras

Los términos compras, adquisiciones, administración de materiales, logística, abastecimiento, administración del suministro y administración de la cadena de suministro se utilizan de manera indistinta ya que no existe un concenso general sobre la terminología. El proceso de adquisición es el eje central de la actividad empresarial de administración de compras y del suministro. Cualquier organización requiere de proveedores por lo que es muy importante acoplarlos con efectividad al entorno organizacional, y que las decisiones de compras no contradigan las estrategias de la empresa.

Las empresas centran sus esfuerzos en aumentar sus ingresos, desminuir sus costos, o una combinación de ambos a fin de obtener ganacias de la forma más eficiente posible.

A. Garcete Facultad Politécnica, UNA, Paraguay, e-mail:albertogarcetepy@gmail.com

R. Benítez Facultad Politécnica, UNA, Paraguay, e-mail: raulkvd@gmail.com

D. P. Pinto-Roa Facultad Politécnica, UNA, Paraguay, e-mail: dpinto@pol.una.py

A. Vazquez Facultad Politécnica, UNA, Paraguay, e-mail: vazquez.aditardo@gmail.com

Este trabajo intenta contribuir lograr decisiones eficientes de compras basadas en estimaciones eficientes de ventas. Se considera que es una decisión importantísima estimar o predecir eficientemente la cantidad o volumen de productos a comprar para reposición de stock y que sirvan para el periodo de ventas que está por llegar.

El stock o existencia de una empresa es el conjunto de materiales y artículos que se almacenan, tanto aquellos que son necesarios para el proceso productivo como los destinados a la venta. La función que desempeña el stock o existencia en una empresa son [5]:

- Evitar la escasez, ante la incertidumbre de la demanda o ante un posible retraso en la reposición o suministro de los pedidos.
- Aprovechar la disminución de los costes a medida que aumenta el volumen de compras o de fabricación.
- Lograr un equilibrio entre las compras y las ventas para alcanzar la máxima competitividad.

En el proceso de compras el caso ideal por supuesto sería poder adivinar la cantidad que se va a vender en el siguiente periodo de venta (puede ser para el siguiente período semanal, quincenal, mensual, trimestral o semestral, etc.), y esto por cada producto que disponibilizamos para la venta. En este caso al término de cada periodo de venta se dispondría de stock cero, con lo cual se llega a una máxima eficiencia en compras. Adivinar es imposible, pero lo que si se puede hacer es estimar eficientemente la cantidad o volumen a vender.

Del por qué la importancia de estimar de forma correcta esta cantidad o volumen, los expertos en negocios explican que los productos parados en stock mientras no se vendan es dinero en estantería, además que generan sobrecostos de mantenimiento como seguros, personal encargado, fecha de vencimiento de los productos, etc. Otro hecho no deseado es la ruptura de stock, es decir el no disponer de un producto en stock cuando haya clientes interesados en comprarlo, lo cual también es considerado pérdida para la empresa. Lo que se desea es mantener un nivel de stock óptimo, es decir, por una parte tener suficiente cantidad para satisfacer la demanda sin caer en roturas de stock y, por otra, evitar que haya un exceso inútil del mismo. Si bien el presente trabajo no está enfocado en medir los costos, lo que sí se busca es comprar de forma eficiente utilizando las herramientas de business intelligence y machine learning que ayudan a estimar lo que se va a vender en el siguiente periodo.

Una administración efectiva de las compras y del suministro contribuye de manera significativa al éxito organizacional. La función del suministro evoluciona a medida que la tecnología y el ambiente competitivo mundial requieren enfoques innovadores [5].

II-B. Cantidad de la orden de compra

El proceso de compra se trata mas bien de un conjunto de etapas: a) Detectar la necesidad, b) Traducir la necesidad en una especificación comercial, c) Buscar potenciales proveedores, d) Seleccionar el proveedor adecuado, e) Detallar la orden de compra y pactar el suministro, f) Recibir los productos, g) Pagar a los proveedores. En el detalle de la orden se ven reflejadas las estimaciones de las cantidades a comprar de los productos.

Antes de realizar una compra surgen las siguientes preguntas:

- £Cuándo debemos realizar un pedido?
- £Qué cantidad debemos solicitar en cada pedido?
- £Cuántas unidades de cada artículo debemos mantener en stock?

Para responder a estas preguntas actualmente se tienen las técnicas de pronósticos de demanda entre las que se destacan los Modelos de Pronóstico Cualitativo y los Modelos de Pronóstico Cuantitativo.

II-B1. Modelos de pronóstico cualitativo [9]:

- Jurado de opinión ejecutiva: Esta técnica se basa en la estimación por consenso entre un grupo de personas de alto mando en la empresa. Se apela a la experiencia y a los conocimientos técnicos de estos ejecutivos. Este método es utilizado cuando se requiere decidir con rapidez ante eventos inesperados, por ejemplo: lanzamiento de un nuevo producto.
- Consulta a la fuerza de ventas: Esta técnica se basa en la experiencia del personal más cercano al cliente que son los vendedores de la empresa. Cada vendedor realiza una estimación de la demanda en su zona de su influencia.
- Encuesta del mercado de consumo: Se encuesta a los clientes acerca de sus planes de compras o sus interes por determinados productos. La estimación se extrae de los resultados de las encuestas.
- Método Delphi: Esta técnica se basa en identificar un panel de expertos que pueden ser gerentes, empleados comunes, o expertos del sector. A cada uno de ellos se les solicita individualmente su estimación de la demanda. Se realiza un proceso iterativo hasta que los expertos alcancen un consenso.
- Analogía de productos similares: Esta técnica de predicción de la demanda se basa en el comportamiento de las ventas de un producto similar o modelo. Se puede realizar comparando con un producto sustituto o complementario.

II-B2. Modelos de pronóstico cuantitativo [9]: Los modelos cuantitativos a su vez se pueden clasificar en modelos de series de tiempo y en modelos causales.

II-B3. Modelos de series de tiempo: En el modelo de series de tiempo el pronóstico se basa solamente en datos anteriores y asume que los factores que influencian las ventas pasadas, presentes y futuras de sus productos continuarán.

- Promedio simplista: Se asume que la demanda del siguiente período será igual a la demanda del período inmediatamente anterior.
- Promedio móvil simple: Se aplica promedio sobre los datos históricos de ventas de una secuencia fija de períodos. Es útil cuando la demanda no presenta estacionalidad o tendencia.

$$PMS = \frac{\sum Demanda\_en\_n\_períodos\_previos}{n} \quad (1)$$

- Promedio móvil ponderado: Ajusta el método de promedio simple asignando mayor peso a los datos más re-

cientes. Sirve para reflejar el nivel de importancia de unos datos sobre otros como resultado de las fluctuaciones.

$$PMP = \frac{\sum (Peso\_período\_n)\,(Demanda\_período\_n)}{\sum Pesos} \tag{2}$$

- Suavización exponencial simple: Para calcular se requiere del pronóstico anterior, la demanda real del periodo de pronóstico y una constante de suavizamiento. Es útil cuando se cuenta con pocos datos históricos.

$$F_t = F_{t-1} + \alpha\,(A_{t-1} - F_{t-1}) \tag{3}$$

Donde F_t = pronóstico nuevo, F_{t-1} = el pronóstico anterior, $0 \leq \alpha \leq 1$ = constante de suavización, A_{t-1} = demanda real del período anterior.

- Suavización exponencial doble: Esta técnica es una modificación del suavizamiento exponencial simple. Agrega una constante de suavización delta (DELTA), cuya función es reducir el error que ocurre entre la demanda real y el pronóstico.

II-B4. Modelo causal: Utiliza una técnica matemática conocida como análisis de regresión, que relaciona una variable dependiente (por ejemplo, la demanda) con una variable independiente (por ejemplo, el precio, publicidad, etc.) en la forma de ecuación lineal.

- Regresión Lineal: Esta técnica permite obtener un estimado analizando el impacto de los factores causales con relación a la demanda del producto o servicio.

$$y = a + bx \tag{4}$$

Donde y = valor de la variable dependiente en este caso ventas, a = intersección en el eje y, b = pendiente de la línea de regresión, x = la variable independiente.

III. Business Intelligence

III-A. Concepto

Business Intelligence abarca un conjunto de conceptos, técnicas y herramientas que se utilizan para la transformación de simples datos en información útil y significativa para el análisis de negocios, ayudando a las organizaciones a mejorar la competitividad facilitando la toma de decisiones[2].

La definición[3] que propone The Datawarehouse Institute es:

"*Business Intelligence es un término paraguas que abarca los procesos, las herramientas y las tecnologías para convertir datos en información, información en conocimiento y planes para conducir de forma eficaz las actividades de los negocios. Business Intelligence abarca las tecnologías de datawarehousing, los procesos en el 'back end'*[1]*, consultas, informes, análisis y las herramientas para mostrar información (herramientas de Business Intelligence) y los procesos en el 'front end'*".

Según lo expuesto en la definición del término Business Intelligence podemos decir que tiene los siguientes objetivos principales[2]:

- Convertir datos en información, información en conocimiento y conocimiento en planes operativos o estratégicos.
- Facilitar la disponibilidad de información a los usuarios de negocios, que les ayude a tomar decisiones más rápidamente.
- Permitir a las organizaciones dirigir de mejor forma, decidir, medir, gestionar y optimizar el alcance de la eficiencia y los resultados financieros.
- Disminuir sustancialmente la incertidumbre que existe ante la toma de decisiones respecto a un plan estratégico.

III-B. Componentes de Business Intelligence

En el siguiente gráfico vemos los componentes que forman parte del Business Intelligence.

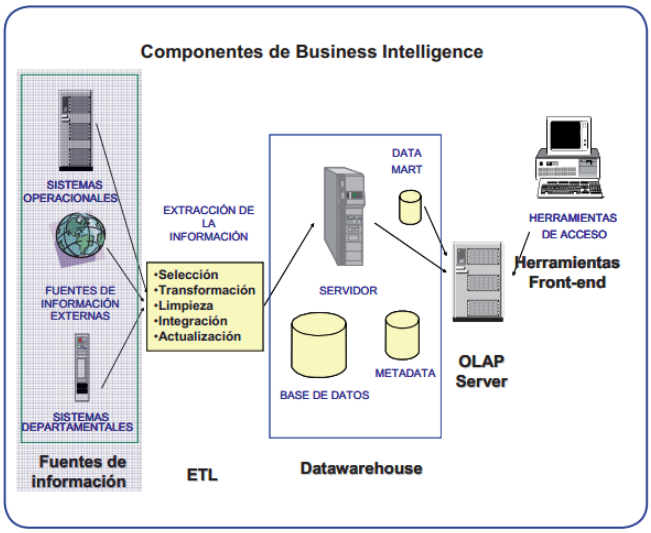

Figure 1. Componentes de Business Intelligence

Los componentes son[2]:

- Fuentes de información, de los cuales se obtienen los datos que se almacenan en el datawarehouse.
- Proceso de Extracción, Transformación y Carga, de los datos en el datawarehouse. Antes de almacenar los datos en el datawarehouse, estos deben ser transformados, limpiados, filtrados y redefinidos.
- El Datawarehouse, donde se almacenan los datos de manera a maximizar la flexibilidad, facilidad de acceso y administración.
- El motor OLAP, que nos provee capacidad de cálculo, consultas, pronósticos, análisis de escenarios en grandes volúmenes de datos.
- La herramientas de visualización, que nos permiten el análisis y la navegación a través de los mismos.

III-C. Indicadores Clave de Rendimiento

Los KPI *(Key Performance Indicators)* o Indicadores Clave de Rendimiento se tratan de indicadores que son decisivos para analizar de forma rápida la situación del negocio y que también facilitan la toma de decisiones. Todos los KPI son indicadores, pero no todos los indicadores son KPI[1].

[1]*Los términos "back end" y "front end" comúnmente usados en Sistemas de Información significan, respectivamente, la parte más cercana al área tecnológica y la más cercana a los usuarios. Si hiciéramos un paralelismo con una tienda, serían la "trastienda" y el "mostrador*

Un cuadro de gestión o de mando no debe excederse en la cantidad de KPI, porque puede darse el problema de "la parálisis por el análisis" que ocurre cuando se pasa de no tener ninguna información a contar con decenas de indicadores y una de las caracteristicas del entorno competitivo actual es que se deben tomar decisiones de forma rápida y antes de que lo hagan los demás competidores[2].

III-C1. Modelado del problema mediante la utilización de KPI: Esta sección se enfoca y analiza uno de los principales problemas con los que se enfrentan las empresas retail[2], la cual trata acerca de la reposición de stock[3], es decir, determinar la cantidad de productos que se deben comprar para satisfacer la demanda de los clientes. Utilizando conceptos y herramientas de Bussines Intelligence, se define y se diseña el datawarehouse donde son almacenados los datos históricos que servirán para el análisis, posteriormente se definen los indicadores claves de rendimiento como métricas que sirven como datos de entrada de las herramientas de aprendizaje automático para crear un modelo de predicción de la cantidad a comprar para la reposición del stock.

III-C2. Base de datos: Para el presente trabajo, contamos con una base de datos real con los registros de productos, proveedores, movimientos de compras, ventas y registro de stock realizados por una empresa retail. Los datos corresponden a movimientos realizados entre el año 2013 al 2016, el cual será el punto de partida para diseñar el datawarehouse.

- **Tabla de Productos**: Lista articulos registrados disponibles para la venta.
- **Tabla Proveedores**: Lista de proveedores.
- **Tabla de Ventas Cabecera**: La tabla de ventas cabecera es una de las tablas principales donde se registran los movimientos de ventas de la empresa retail. Contiene datos de la fecha, número de factura, cliente, montos totales entre otros datos.
- **Tabla de Ventas Detalle**: La tabla de ventas detalle contiene los registros de los productos que fueron comercializados, cada detalle esta relacionado a un registro cabecera. Contiene datos de la fecha, el producto, precio de costo, precio de venta, cantidad y otros datos.

III-C3. Datawarehouse: El datawarehouse para el modelado se diseña a partir de la definición de las tablas de hechos y dimensiones utilizando el esquema en estrella.

Tablas de hechos: De las tablas transaccionales definimos 3 tablas de hechos que nos servirá para definir los KPI que utilizaremos para el análisis.

- **Tabla de hechos Cabecera:** almacena los datos históricos de las ventas, cada registro guarda datos de: fecha, cliente, caja, número de factura y montos totales. Las métricas de la tabla de hechos son monto total, monto exento, monto gravado, monto gravado 5% y monto gravado 10%.
- **Tabla de hechos Detalles:** almacena los datos históricos del detalle, cada registro guarda información del número de comprobante, fecha, proveedor, cliente, cantidad y precio. Las métricas asociadas a la tabla de hechos son, cantidad, precio unitario, precio unitario neto, impuesto, costo y el importe total.
- **Tabla de hechos Stock:** almacena los datos históricos de cada movimiento de compra y de venta. Las métricas utilizadas para la tabla de hechos son: cantidad, precio unitario y costo unitario.

Dimensiones: Las tablas de dimensiones diseñadas para el modelado del datawarehouse son:

- **Dimensión Fecha:** La tabla de dimensión fecha esta ligada a todas las tablas de hechos, sirve para limitar o agrupar los datos de las tablas de hechos al momento de realizar consultas sobre estas en el tiempo. Con la dimesión fecha se pueden establecer niveles jerárquicos en días, semanas, meses, trimestres, semestres y años.
- **Dimensión Productos:** La tabla de dimensión producto esta relacionada a las tablas de hechos Detalles y Stock, contiene los atributos o campos por la cual se pueden filtrar o agrupar datos al realizar consultas sobre la tabla de hechos.
- **Dimensión Proveedores:** La tabla de dimensión proveedores esta relacionada a la tabla de hechos Detalles, contiene los atributos o campos por la cual se puede filtrar o agrupar datos al realizar consultas sobre la tabla de hechos.
- **Dimensión Clientes:** La tabla de dimensión Clientes esta relacionada a las tablas de hechos Cabecera y Detalles, contiene los atributos o campos por la cual se pueden filtrar o agrupar datos al realizar consultas sobre la tabla de hechos.
- **Dimensión Cajas:** La tabla de dimensión Cajas esta relacionada a la tabla de hechos Cabecera, contiene los atributos o campos por la cual se pueden filtrar o agrupar datos al realizar consultas sobre la tabla de hechos.

III-C4. Definición de los KPI: En el marco de esta tesis, en esta sección se definirán los KPI que se utilizarán en el modelado para la estimación de cantidades eficientes en las órdenes de compra de productos para la reposición de stock del siguiente periodo de tiempo[1] (Ej.: cantidad a comprar la satisfacer la demanda de la siguiente semana, quincena, o mes).

Cada KPI mide un valor obtenido de los datos históricos almacenados en el datawarehouse. El cálculo de cada valor se realiza para cada producto y en un periodo de tiempo (semanal, quincenal o mensual), es decir, cada producto tendrá un valor distinto para cada uno de los KPI citados a continuación.

III-C4a. Ticket Medio.: Es la cantidad media por cada transacción de compra que se realiza de un determinado producto. El indicador viene determinado por dos variables: La cantidad total vendida del producto y el total de tickets en las que fue vendido el producto. Aplicando la siguiente fórmula obtenemos el valor de la cantidad media de venta para cada producto.

[2]Una empresa retail es cualquier comercio que vende sus productos al consumidor final, desde un supermercado a una tienda de barrio, desde un negocio de electrodomésticos a una franquicia textil, ya sea con cientos de puntos de venta o con un solo establecimiento.

[3]Stock o existencia es la cantidad de un determinado producto almacenado o disponible para la venta.

$$X = \frac{\sum (Cantidad)}{Total\,Tickets\,Periodo} \tag{5}$$

III-C4b. Cifra de Ventas: La cifra de ventas es un KPI que sirve para explicar el importe total de ventas que se ha obtenido para un producto. Se obtiene de la siguiente fórmula.

$$X = \sum (Precio * Cantidad) \tag{6}$$

III-C4c. Margen Comercial: Es la razón entre el precio de venta y precio de costo del producto, es un indicador que permite conocer el porcentaje de rentabilidad del producto. Se obtiene de la siguiente fórmula.

$$X = \frac{\sum ((Precio - Costo) * Cantidad)}{\sum (Precio * Cantidad)} * 100 \tag{7}$$

III-C4d. Rotación de Stock: Este indicador mide la cantidad de veces que el stock del producto se renueva durante un determinado ciclo comercial, es decir, la cantidad de veces que se recupera la inversión. Se obtiene de la siguiente fórmula.

$$X = \frac{\sum (Total\,Ventas\,Periodo)}{\left(\frac{Stock\,Inicial - Stock\,Final}{2}\right)} \tag{8}$$

III-C4e. Coeficiente de Rentabilidad: El indicador mide la rentabilidad obtenida por la empresa basada en el margen y la rotación, el objetivo de toda empresa retail es aumentar los niveles de rotación. El coeficiente se obtiene de la siguiente fórmula.

$$X = \left(\sum (Precio - Costo) * Cantidad\right) * Rotacion\,Stock \tag{9}$$

III-C4f. Cobertura de Stock: Este indicador muestra el periodo de tiempo (habitualmente se expresa en días o semanas) que el negocio puede continuar vendiendo con el stock de que dispone en el momento, sin incorporar nuevas cantidades de ese producto.

$$X = \frac{Stock\,Actual}{Promedio\,Cantidad\,Venta\,Ultimos\,3\,Periodos} \tag{10}$$

III-C5. Cálculo de valores para los Indicadores Clave de Desempeño.: Definidos los KPI a ser utilizados, obtenemos los valores de cada KPI para cada producto y periodo de los datos almacenados en el datawarehouse, para ello codificamos a sentencias SQL las fórmulas detalladas en la sección anterior y almacenamos la información de los resultados en la base de datos. Además de los valores de los KPI, en cada registro adicionalmente se guarda la información de la cantidad, fecha, año, mes, quincena y semana.

Agrupamos el conjunto de los valores de los KPI de cada producto en 3 periodos de tiempo: semanal, quincenal y mensual.

III-C6. Asignación de etiquetas: A cada tupla de valores KPI obtenidos para cada producto se le debe asignar una etiqueta, el cual es uno de los valores importantes para el modelado mediante el aprendizaje automático. Esta asignación de las etiquetas esta basada en la opinión de un experto (que habitualmente podría ser el gerente de compras). La estrategia utilizada para el etiquedato es de la siguiente manera:

Para cada KPI se definen un rango de valores y se asigna una letra (a, b, c ,d, e, f, g, h, i, ..., u) de acuerdo al valor obtenido.

Tabla I
RANGO KPI TICKET MEDIO

(=) igual a 0	a
> (mayor) a 0 y < (menor) a 1	b
>= (mayor o igual) a 1 y <= (menor o igual) a 3	c
> (mayor) a 3	d

Tabla II
RANGO KPI CIFRA VENTAS (%)

>= (mayor o igual) a 0 y <= (menor o igual) a 20	e
> (mayor) a 20 y <= (menor o igual) a 50	f
> (mayor) a 50 y <= (menor o igual) a 80	g
> (mayor) a 80 y <= (menor o igual) a 100	h

Tabla III
RANGO KPI MARGEN COMERCIAL (%)

>= (mayor o igual) a 0 y <= (menor o igual) a 20	i
> (mayor) a 20 y <= (menor o igual) a 50	j
> (mayor) a 50 y <= (menor o igual) a 80	k
> (mayor) a 80 y <= (menor o igual) a 100	l

Tabla IV
RANGO KPI ROTACIÓN STOCK

(=) igual a 0	m
> (mayor) a 0 y < (menor) a 1	n
>= (mayor o igual) a 1 y <= (menor o igual) a 3	o
> (mayor) a 3	p

Tabla V
RANGO KPI COBERTURA STOCK

(=) igual a 0	q
> (mayor) a 0 y < (menor) a 1	r
>= (mayor o igual) a 1 y <= (menor o igual) a 3	s
> (mayor) a 3 y <= (menor o igual) a 10	t
> (mayor) a 10	u

Una vez asignado las letras “a”, “b”, “c”, “d”, “e” hasta “u” se busca la combinación de letras correspondientes en la tabla de etiquetado realizado por el experto y se asigna el valor de la etiqueta correspondiente.

Tabla VI
TABLA DE ETIQUETADO POR EL EXPERTO

aeimq	Nada	bejnq	Poco	bejoq	Poco
aeimr	Nada	bejnr	Poco	bejpq	Medio
aeims	Nada	bejns	Nada	beknq	Poco
aeimt	Nada	bejnt	Nada	beknr	Poco
aeimu	Nada	bejnu	Nada	bekns	Nada
...	...	...	...	...	...

Una vez finalizado el etiquetado de la totalidad de las tuplas de KPI por cada producto y periodo, lo resultados son exportados a archivos con extensión csv, para cada producto se crea 3 archivos, uno por cada periodo (semanal, quincenal, mensual) que tiene como nombre el Identificador del producto y que contiene los valores de los resultados para los KPI. Estos archivos son los datos que sirven como entrada para crear el modelo de estimación de compra eficiente para la reposicion de stock mediante algoritmos de aprendizaje automático

IV. Machine Learning

En 1959 Arthur Samuel en una publicación escribió: *"Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort"* [7]. Lo que nos lleva a pensar que uno pioneros de machine learning ya dejaba visualizar que los programas, a partir del aprendizaje sobre los datos históricos (la experiencia), podrían efectuar tareas de toma de decisiones sin ser programadas explícitamente dichas decisiones.

IV-A. Definición

Samuel define machine learning como sigue: *"Machine Learning es un campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programadas"*. Otro investigador de machine learning Tom Mitchell propuso en 1998 la siguiente definición: *"Well posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E"*.

"The purpose of machine learning is to learn from training data in order to make as good as possible predictions on new, unseen, data"[6]. La dificultad radica en que debemos construir modelos que nos acerquen a una buena predicción sobre datos aún no conocidos o imprevistos. Peter Prettenhofer y Gille Louppe presentan la siguiente definición:

Data comes as...

- A set of examples $\{(x_i, y_i) \mid 0 \leq i < n\_samples\}$, with
  - Feature vector $x \in \mathbb{R}^{n\_features}$, and
  - Response $y \in \mathbb{R}$(regression) or $y \in \{-1, 1\}$ (classification)
- Goal is to...
  - Find a function $\hat{y} = f(x)$
  - Such that error $L(y, \hat{y})$on new (unseen) x is minimal

IV-B. Categoría de algoritmos

Los algoritmos de aprendizaje automático se pueden categorizar según la forma en que se realiza el aprendizaje, pero teniendo en cuenta que todos reciben un conjunto de ejemplos del que aprender.

IV-B1. Aprendizaje supervisado (supervised learning): El algoritmo recibe datos de entrenamiento que contienen la respuesta correcta para cada ejemplo.:

IV-B2. Aprendizaje no supervisado (unsupervised learning): El algoritmo busca estructuras en los datos de entrenamiento, como encontrar qué ejemplos son similares entre sí, y agruparlos en clusters.:

IV-C. Tipos de problemas

Teniendo en cuenta las clases de problemas que los algoritmos de aprendizaje pueden resolver, los tipos de problemas se pueden agrupar como sigue.

IV-C1. Regresión: Un problema de aprendizaje supervisado donde la respuesta a aprender es un valor continuo.:

IV-C2. Clasificación: Un problema de aprendizaje supervisado donde la respuesta a aprender es un valor de un conjunto finito de posibles valores discretos.:

IV-C3. Segmentación: Un problema de aprendizaje no supervisado donde la estructura a aprender es un conjunto de clusters donde cada cluster tiene similares ejemplos.:

IV-C4. Análisis de red: Un problema de aprendizaje no supervisado donde la estructura a aprender es información acerca de la importancia y el rol de los nodos en una red.:

IV-D. Componentes esenciales [10].

La entrada de un esquema de aprendizaje automático es un conjunto de instancias o ejemplos (examples). Estas instancias son las cosas que deben ser clasificadas, asociadas o agrupadas. Las instancias son caracterizadas mediante los valores de un conjunto predeterminado de atributos (features). Los ejemplos o instancias utilizadas en el proceso de entrenamiento del algoritmo de aprendizaje automático constituye el conjunto de entrenamiento (training set). Para predecir el rendimiento de un clasificador sobre nuevos datos, necesitamos evaluar su tasa de error en un conjunto de datos que no desempeñó ningún papel en la formación del clasificador, este conjunto de datos independiente se denomina conjunto de prueba (test set).

IV-E. El problema de la clasificación

En los problemas de clasificación el modelo creado debe predecir la clase, tipo o categoría de la salida.

IV-E1. Clasificación binaria (binary classification): En su forma más simple se reduce a la pregunta: dado un patrón x extraído de un dominio X, estimar qué valor asumirá una variable aleatoria binaria asociada y ∈ {± 1} [8].:

IV-E2. Clasificación multiclase (multiclass classification): Es la extensión lógica de la clasificación binaria. La principal diferencia es que ahora y ∈ {1, ... , N} puede asumir un rango de valores diferentes [8].:

IV-F. Algoritmos de clasificación en WEKA

Weka es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Los algoritmos pueden ser aplicados directamente a un conjunto de datos o llamados desde código Java. Weka contiene herramientas para preprocesamiento de datos, clasificación, regresión, clustering,

reglas de asociación y visualización. También es adecuado para desarrollar nuevos esquemas de aprendizaje automático [4]. Los algoritmos de clasificación de Weka que se utilizarán son los siguientes [11]: BayesNet, NaiveBayes, NaiveBayesUpdateable, Logistic, MultilayerPerceptron, SimpleLogistic, SMO, OneR, DecisionTable, JRip, PART, ZeroR, DecisionStump, J48, LMT, RandomForest, RandomTree, REPTree.

IV-G. Evaluación de lo aprendido

La evaluación es la clave para lograr avances reales en el aprendizaje automático. Entre las técnicas se destaca la Validación Cruzada (Cross-Validation) que consiste dividir los datos en un número de pliegues o particiones. Si por ejemplo elegimos cuatro, entonces cada partición se utiliza para las pruebas y las demás para el entrenamiento. Al repetir este proceso 4 veces se consigue que cada partición se haya utilizado una vez como conjunto de pruebas. La técnica estándar para predecir la tasa de error es la Validación Cruzada Estratificada (Stratified k-fold Cross-Validation). La estratificación se refiere al proceso de reorganizar los datos de tal manera a asegurar que cada pliegue sea una buena representación del conjunto. Comúnmente se acepta que 10 es el número de pliegues con el que se obtiene la mejor estimación de error, idea basada en diversas pruebas sobre conjuntos de datos diferentes y para distintas técnicas de aprendizaje [10].

Otra técnica es el Porcentaje de División (Percentage Split) con el que puede retener para la prueba un determinado porcentaje de los datos. Es una alternativa utilizar un conjunto de pruebas separado o una división porcentual de los datos de entrenamiento. Si elegimos 60 % como porcentaje de división, entonces el conjunto de prueba se constituirá con el 40 % de las instancias y el conjunto de entrenamiento con el 60 % de las instancias.

IV-H. Métricas de desempeño

Para los problemas de clasificación, es natural medir el rendimiento de un clasificador en términos de la tasa de error (error rate). El clasificador predice la clase de cada instancia: si es correcta, se cuenta como un éxito; sino, es un error. La tasa de error es sólo la proporción de errores cometidos sobre un conjunto de instancias, y mide el rendimiento general del clasificador. Por supuesto, lo que nos interesa es el probable desempeño futuro en nuevos datos, no el rendimiento pasado en datos antiguos.

Para predecir el rendimiento de un clasificador en nuevos datos, necesitamos evaluar su tasa de error en un conjunto de datos que no desempeñó ningún papel en la formación del clasificador. Este conjunto de datos independiente se denomina conjunto de pruebas. En tales situaciones, la gente suele hablar de tres conjuntos de datos: los datos de entrenamiento, los datos de validación y los datos de prueba. Los datos de entrenamiento son utilizados por uno o más esquemas de aprendizaje para conocer clasificadores. Los datos de validación se utilizan para optimizar los parámetros de los clasificadores, o para seleccionar uno determinado. A continuación, los datos de prueba se utilizan para calcular la tasa de error del método final optimizado. Cada uno de los tres conjuntos debe ser independiente: El conjunto de validación debe ser diferente del conjunto de entrenamiento para obtener un buen desempeño en la etapa de optimización o selección y el conjunto de pruebas debe ser diferente de ambos para obtener una estimación confiable de la tasa de error real.

IV-H1. Aciertos: Número de instancias correctamente clasificadas.:

IV-H2. Porcentaje de Aciertos: Porcentaje de instancias correctamente clasificadas.:

IV-H3. Estadística Kappa (Kappa Statistic): En problemas de clasificación para aplicaciones reales normalmente los errores cuestan diferentes cantidades. Por ejemplo en bancos y financieras el costo de prestar a una persona que no paga sus deudas es mayor que el costo de rechazar un préstamo a una persona que es pagadora. Los Verdaderos Positivos (True Positive - TP) y Verdaderos Negativos (True Negative - TN) son clasificaciones correctas. Un Falso Positivo (False Positive - FP) es cuando el resultado se predice incorrectamente como sí (o positivo) cuando es realmente no (o negativo). Un Falso Negativo (False Negative - FN) es cuando el resultado se predice incorrectamente como negativo cuando es realmente positivo. En la predicción multiclase, cada elemento de la matriz de confusión muestra el número de ejemplos de prueba para los que la clase real es la fila y la clase prevista es la columna. Son buenos resultados los grandes números en la diagonal principal e idealmente cero fuera de la diagonal principal. "Kappa se utiliza para medir el acuerdo entre la predicción y la observación de las categorizaciones de un conjunto de datos, mientras que se corrige para un acuerdo que ocurre por casualidad". [10]. Si los evaluadores están totalmente de acuerdo, entonces Kappa alcanza su valor máximo y es igual a 1. Si no hay total acuerdo entre los evaluadores, entonces Kappa tiene un valor < 1*.:*

$$\kappa = \frac{Pr\,(a) - Pr\,(e)}{1 - Pr\,(e)} \tag{11}$$

Donde: $Pr\,(a)$ es el acuerdo observado relativo entre los observadores y $Pr\,(e)$ es la probabilidad hipotética de acuerdo al azar utilizando los datos observados para calcular las probabilidades de que cada observador clasifique aleatoriamente cada categoría.

IV-H4. Sensibilidad (Recall): Calcula la sensibilidad con respecto a una clase en particular, esto se define como: positivos correctamente clasificados / positivos totales [10].:

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

IV-H5. Precisión (Precision): Calcula la precisión con respecto a una clase en particular, esto se define como: positivos correctamente clasificados / total predicho como positivo [10].:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

IV-H6. Puntuación-F (F-Measure): La Puntuación-F es una medida de la exactitud de una prueba. La Puntuación-F puede interpretarse como un promedio ponderado de la precisión y sensibilidad, donde alcanza su mejor valor en 1 y

*el peor en 0. Se define como: 2 * Recall * Precision / (Recall + Precision) [10].:*

$$F - Measure = \frac{2 * Recall * Precision}{(Recall + Precision)} \quad (14)$$

V. Experimentación

V-A. Modelado del Aprendizaje Automático

Se describirá cómo es la implementación del proceso de aprendizaje automático para este caso de estudio. Se mostrará primeramente cómo está constituída la salida del proceso de business intelligence, que en esencia proveen las instancias necesarias para la entrada del proceso de aprendizaje automático. También se verá qué clasificadores fueron utilizados, cómo se realizó el paso de entrenamiento y de evaluación, y cuáles son las métricas de evaluación consideradas para medir el rendimiento de los clasificadores.

V-B. Datos proveídos por business intelligence

La salida de business intelligence se constituye de archivos CSV que podemos representar como se muestran en el Cuadro 1.

El Cuadro 1 es una porción de un archivo CSV que contiene la salida de business intelligence calculada sobre las ventas mensuales de un determinado producto. Hay 309 productos diferentes analizados para períodos mensuales, lo que equivale a 309 archivos CSV. En realidad el Cuadro 1 tiene un máximo de 34 filas sin incluir el encabezado, lo que corresponde directamente a 34 instancias o ejemplos. Entonces, por cada producto analizado tenemos disponibles un máximo de 34 instancias. La última columna es la clase de cada instancia, una columna etiquetada de valores discretos. Todas las anteriores columnas constituyen el conjunto de características o atributos de las instancias. El listado completo de encabezados es: KPI TICKET MEDIO, KPI CIFRA VENTAS, KPI MARGEN COMERCIAL, KPI ROTACION STOCK, KPI COEF RENTABILIDAD, KPI COBERTURA STOCK, CANTIDAD, ANHO, MES, SEMANA y RESULTADO.

KPI TM	KPI CV	KPI MC	KPI RS	...	RESULT
1,143	16000	33,131	2	...	Poco
1,143	32000	35,909	3,2	...	Medio
1	10000	35,909	0,833	...	Medio
1	14000	35,909	0,824	...	Poco
1,091	24000	35,909	3	...	Medio
1,2	12000	35,909	1	...	Medio
1	16000	35,909	0,8	...	Poco
1,083	26000	35,909	1,733	...	Medio
1,4	14000	35,909	1,4	...	Medio

Tabla VII
Ejemplo que corresponde a métricas BI mensuales sobre las ventas de un producto.

De forma similar hay archivos CSV que contienen la salida de business intelligence calculada sobre las ventas quincenales de un determinado producto. Hay 228 productos diferentes analizados para períodos quincenales, lo que equivale a 228 archivos CSV. Dichos archivos CSV tienen un máximo de 68 filas, sin incluir el encabezado, lo que corresponde directamente a 68 instancias o ejemplos. Entonces, por cada producto analizado tenemos disponible un máximo de 68 instancias. Se diferencia del Cuadro 1 en que tiene una columna mas, que es el número de QUINCENA en el año. La última columna es la clase de cada instancia, una columna etiquetada de valores discretos. Todas las anteriores columnas constituyen el conjunto de características o atributos de las instancias.

Y también de forma similar hay archivos CSV que contienen la salida de business intelligence calculada sobre las ventas semanales de un determinado producto. Hay 127 productos diferentes analizados para períodos semanales, lo que equivale a 127 archivos CSV. Dichos archivos CSV tienen un máximo de 151 filas, sin incluir el encabezado, lo que corresponde directamente a 151 instancias o ejemplos. Entonces, por cada producto analizado tenemos disponible un máximo de 151 instancias. Se diferencia del Cuadro 1 en que tiene una columna mas, que es el número de SEMANA en el año. La última columna es la clase de cada instancia, una columna etiquetada de valores discretos. Todas las anteriores columnas constituyen el conjunto de características o atributos de las instancias.

V-C. Esquema del procesamiento de las instancias

Se debe recorrer todo el conjunto de archivos CSV, tanto los archivos que contienen instancias referentes a BI mensuales, los que contienen instancias referentes a BI quincenales y los que contienen instancias referentes a BI semanales.

Luego, cada archivo de instancias se entrena con todos los algoritmos de clasificación WEKA posibles y la evaluación se hace tanto por el método Percentage Split así como también por el método Stratified K-fold Cross Validation. Finalmente las métricas de evaluación se almacenan en dos tablas; una tabla con los resultados de evaluación del aprendizaje automático con el método Percentage Split para los periodos mensuales, quincenales y semanales; y otra tabla con los resultados de evaluación del aprendizaje automático con el método Stratified K-fold Cross Validation para los periodos mensuales, quincenales y semanales.

V-D. Entrenamiento y evaluación de las instancias

A continuación se lista el conjunto de clasificadores WEKA utilizados durante el procesamiento de cada archivo CSV. A su vez estos clasificadores se pueden sub dividir en basesianos, basados en funciones, reglas y árboles.

- Bayesianos: BayesNet, NaiveBayes, NaiveBayesUpdateable.
- Basados en funciones: Logistic, MultilayerPerceptron, SimpleLogistic, SMO.
- Basados en reglas: OneR, DecisionTable, JRip, PART, ZeroR.
- Basados en árboles: DecisionStump, J48, LMT, RandomForest, RandomTree, REPTree.

Con la API de WEKA se realiza el entrenamiento y la evaluación. Se utilizan los algoritmos de clasificación mencionados, así que por cada modelo se evalua por Percentage Split o por Stratified K-fold Cross Validation. Para esta tesis evaluamos por ambos métodos.

V-E. Métricas de los resultados de la evaluación

Por cada modelo procesado, luego de construir su clasificador y evaluarlo se obtienen las métricas Cantidad de Aciertos o el Porcentaje de Acierto, la estadística Kappa; y por cada clase (Nada, Medio, Mucho) se obtienen las métricas Area Under ROC (ROCA), Recall (RCALL), Precision (PREC), F-Measure (FMEA) y Area Under Precision-Recall Curve (PRCA).

V-F. Resultados numéricos

Hay dos tablas que se generan al concluir el proceso de aprendizaje automático.

Una con los resultados de evaluación del aprendizaje automático con el método Percentage Split para los periodos mensuales, quincenales y semanales que contienen los datos sobre: PRODUCTO, PERIODO, TIPO DE CLASIFICADOR, TIPO DE EVALUACIÓN, % DE ACIERTOS, KAPPA. También la PRECISIÓN, RECALL y F-MEASURE por cada clase (Nada, Poco, Medio, Mucho).

Otra con los resultados de evaluación del aprendizaje automático con el método Stratified K-fold Cross Validation para los periodos mensuales, quincenales y semanales que contienen los datos sobre: PRODUCTO, PERIODO, TIPO DE CLASIFICADOR, TIPO DE EVALUACIÓN, % DE ACIERTOS, KAPPA. También la PRECISIÓN, RECALL y F-MEASURE por cada clase (Nada, Poco, Medio, Mucho).

VI. Discusión

El análisis global de los resultados se basa en la métrica Kappa. Por cada producto analizado se elige como clasificador aquel que haya alcanzado el mayor valor de Kappa. Entonces luego se calculan los promedios de porcentaje de aciertos para periodos mensuales, quincenales y semanales.

En la Figura 2 se muestra un gráfico de barras con promedios del porcentaje de aciertos utilizando Stratified k-fold Cross Validation como método de evaluación.

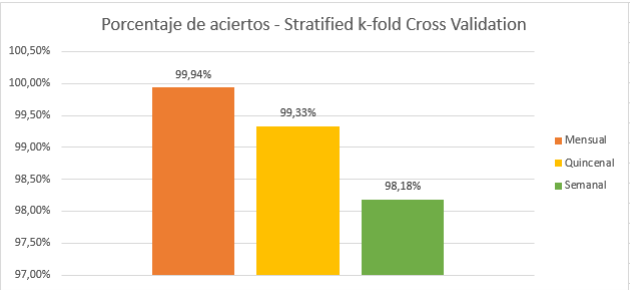

Figura 2. Promedios del porcentaje de aciertos utilizando Stratified k-fold Cross Validation.

En la Figura 3 se muestra un gráfico de barras con promedios del porcentaje de aciertos utilizando Percentage Split Validation como método de evaluación.

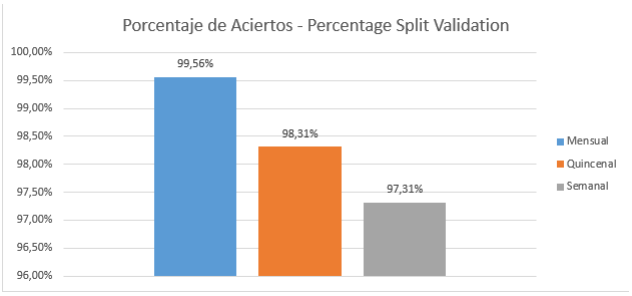

Figura 3. Promedios del porcentaje de aciertos utilizando Percentage Split Validation.

Se puede observar que se obtienen altos porcentajes de aciertos tanto para los periodos mensuales, quincenales como semanales. Como se trata de una prueba exhaustiva, por cada producto se intenta con todos los algoritmos de clasificación posible, además que se evalúan con dos métodos, entonces podemos deducir que se trata se resultados fiables. Obtener buenos resultados depende en gran medida de que los valores de KPI hayan sido obtenidos correctamente y también que el etiquetado haya sido realizado por un experto en compras o ventas.

VII. Conclusiones

Este trabajo se enfocó principalmente en elaborar una nueva técnica de estimación de compras de productos para la reposición de stock en las empresas retail. Como ya se había mencionado en el capítulo 2 la gestión de compras es uno de los ejes centrales de la actividad empresarial y las decisiones de compras son uno de los principales problemas con los que se enfrentan las empresas al momento de la reposición del stock. Partiendo de la premisa de la problemática y analizando las técnicas utilizadas en la actualidad basadas en pronósticos y con el creciente incremento en la incorporación de tecnologías de Business Intelligence dentro de las empresas para hacer uso de los datos almacenados, se vió la oportunidad de aprovechar esa información histórica. Con esta nueva técnica se utiliza los Indicadores Claves de Rendimiento y apoyados en la experiencia y el conocimiento de un experto en compras (gerente o encargado de compras) se realiza el modelado utilizando algoritmos clasificadores de Machine Learning.

De acuerdo a los resultados experimentales se obtuvieron altas tasas de acierto, haciendo pruebas exhaustivas con todos los algoritmos de clasificación posibles y evaluando a la vez con dos métodos. Apoyados en estos resultados se puede creer en la fiabilidad de este nuevo modelo.

Con esta técnica de estimación planteada se pretende que el modelo se convierta en una herramienta de apoyo en la toma de decisiones al gerente encargado de realizar las compras en la reposición de stock.

Referencias

[1] Marcos Alvarez. *Cuadro de Mando Retail*. Profit, 2013.

[2] Josep LluÃ■s Cano. *Busines Intelligence: Competir con informaciÃşn*. ESADE, Banesto, Banesto Pyme, 2007.
[3] Wayne W. Eckerson and Cindi Howson. Enterprise business intelligence: Strategies and technologies for deploying bi on an enterprise scale tdwi report series. 2005.
[4] Machine Learning Group. Weka 3: Data mining software in java.
[5] P. Fraser Johnson, Michiel R. Leenders, and Anna E. Flynn. *AdministraciÃşn de compras y abastecimientos*. McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V, 2012.
[6] Jean Francois Puget. What is machine learning?, May 2016.
[7] Arthur Samuel. Some studies in machine learning using the game of checker. *IBM Journal 3, 211-229*, 1959.
[8] Alex Smola and S.V.N. Vishwanathan. *Introduction to Machine Learning*. The Press Syndicate of The University of Cambridge, 2008.
[9] Naim Caba Villalobos, Oswaldo Chamorro Altahona, and TomÃąs JosÃľ Fontalvo Herrera. *GestiÃşn de la ProducciÃşn y Operaciones*. 2011.
[10] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining - Practical Machine Learning Tools and Techniques - Third Edition*. Copyright ÂĽ 2017 Elsevier Inc. All rights reserved, tercera edition, 2011.
[11] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining - Practical Machine Learning Tools and Techniques - Fourth Edition*. Cuarta edition, 2016.