Técnica de pronóstico de la demanda basada en Business Intelligence y Machine Learning

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

***Abstract*—Pronosticar ciertos eventos constituye una actividad por la cual el ser humano siempre sintió fascinación y necesidad de realizarlo. En la actualidad, uno de esos eventos se relacionan con las empresas y consiste en pronosticar la demanda de ventas para un período futuro, a su vez representa uno de los más importantes retos con que se enfrenta una organización. Este pronóstico de demanda disminuirá la incertidumbre del Gerente de Compras en el momento de tomar decisiones acerca del volumen de productos a adquirir para la reposición de stock. Este trabajo propone una nueva técnica de pronósticos basada en la integración de herramientas de Business Intelligence y Machine Learning. Los experimentos indican que el modelo propuesto alcanza resultados prometedores y esta nueva técnica puede transformarse en una sólida herramienta de apoyo para la toma de decisiones.**

***Index Terms*—Pronóstico, Retail, Compras, Business Intelligence, KPI, 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 en empresas de este sector es pronosticar las ventas para el próximo periodo comercial.

Actualmente en el proceso de gestión de compras se utilizan ciertas técnicas de pronósticos para determinar las cantidades de las órdenes de compra, dichas técnicas pueden ser cuantitativas o cualitativas. 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 plantea una nueva técnica para estimar los volúmenes de ventas del siguiente periodo comercial, luego esta estimación servirá de apoyo en la toma de decisión de la cantidad establecida en las órdenes de compras para la reposición de stock. En este nuevo modelo se integran técnicas de Business Intelligence y Machine Learning.

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

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 utilizan como entrada las series de KPI obtenidas en la etapa anterior de Business Intelligence y constituyen las instancias que alimentan los distintos algoritmos de clasificación del aprendizaje supervisado implementado. 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 pronosticar las ventas futuras.

En la segunda sección se abordan los conceptos de Pronóstico de la Demanda. En la tercera sección se definen conceptos de Business Intelligence y la primera fase de modelado del problema. En la cuarta sección se abordan conceptos de Machine Learning, en la quinta sección se implementa la segunda fase del modelado del problema y se analizan los resultados experimentales. En la última sección se realiza la conclusión general.

II. Pronóstico de la demanda

La elaboración de pronósticos de ventas precisos es uno de los retos más importantes en empresas del tipo retail. En un escenario inicial se tienen los depósitos llenos de productos listos para ser llevados a los mostradores. A medida que pasa el tiempo la cantidad en depósito va decreciendo por la demanda de los clientes y llegado un momento crítico hay que tomar la decisión de reponer el stock. Si bien la reposición de stock se lleva a cabo dentro de un proceso empresarial llamado *Administración de Compras*, hay un componente vital dentro de este proceso que es estimar la cantidad o volumen de productos a adquirir para reponer el stock. Es ahí donde entra en juego el pronóstico de la demanda.

A continuación se explicará sintéticamente el proceso de *Administración de Compras*, para luego analizar las principales técnicas de pronósticos de demanda que están vigentes en el mundo empresarial.

II-A. Administración de compras [8]

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. Este trabajo intenta contribuir a lograr decisiones eficientes de compras basadas en pronósticos de demanda precisos. Se considera que es una decisión importantísima estimar o predecir eficientemente la cantidad o volumen de productos para reponer el 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:

- 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.

El proceso de compras o adquisiciones se trata 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 *Métodos de Pronósticos Cualitativos* y los *Métodos de Pronósticos Cuantitativo*s.

II-B. Métodos de pronósticos cualitativos [15][6]

- Opinión del Gerente: El pronóstico se basa en la opinión, experiencia y conocimiento de un solo gerente.
- Junta de opinión ejecutiva: Similar al método anterior, la diferencia está en que se basa en un grupo de ejecutivos que colaboran para emitir colectivamente el pronóstico, compartiendo de este modo la responsabilidad.
- Consulta a la fuerza de ventas: Esta técnica se basa en la experiencia del personal más cercano al cliente que es el cuerpo de vendedores de la empresa. Cada vendedor realiza una estimación de la demanda en su zona de su influencia. Luego las estimaciones son revisadas por los mandos superiores, para obtener un pronóstico corporativo final.
- Encuesta en el 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. Son útiles para elaborar planes de marketing, lazamiento de nuevos productos, etc.
- Método Delphi: Se basa en identificar un panel de expertos que pueden ser gerentes, empleados comunes, o expertos del sector. Se tiene un cuestionario donde cada uno de ellos lo completa de forma aislada. Se integran todas las respuestas, luego cada experto tiene acceso al set de respuestas y puede ajustar su respuesta conforme le parezca conveniente. Este proceso se realiza iterativamentte hasta que los expertos alcancen un consenso.
- Analogía de productos similares: 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-C. Métodos de pronósticos cuantitativos

Estos modelos se basan en métodos de pronóstico estadísticos que a partir de los datos históricos de ventas y suponiendo que las tendencias históricas continuarán, son capaces de anticipar la demanda futura [6]. En general, para modelar cuantitativamente se debe disponer de información sobre la variable a pronosticar, la información debe ser cuantificable y el patrón histórico de cierto modo se debe repetir en el futuro [2].

El pronóstico de la demanda de productos es sólo una aplicación importante de estos métodos. En otros casos, los pronósticos se podrían utilizar para evaluar los requerimientos de cantidades tan diversas como las partes de repuestos, el rendimiento de la producción y las necesidades de personal. Las técnicas de pronóstico se usan también frecuentemente para anticipar las tendencias económicas a nivel regional, nacional o incluso internacional [6].

En general, los métodos cuantitativos se clasifican en técnicas de series de tiempo y en pronósticos causales.

II-C1. Métodos de series de tiempo: Una serie de tiempo es un conjunto de observaciones de la variable a pronosticar, medidas en puntos o períodos sucesivos del tiempo pasado [6]. El histórico de ventas de un producto, donde se observan los valores de las cantidades vendidas mensualmente, constituye un ejemplo de serie de tiempo. Los datos históricos de la variable a predecir están limitados a sus valores pasados.

El objetivo del método es obtener una buena predicción del valor futuro de la variable a pronosticar, enmarcado por supuesto en la serie de tiempo. Para lograr el objetivo, el modelo debe descubrir el patrón dentro de la serie y luego ser capaz de extrapolarlo hacia el futuro [2]. De cierta manera, hay una suposición intrínseca al modelo de que los factores que influyen en las ventas pasadas y presentes continuarán a futuro.

Si bien el volumen de ventas es un buen indicador de la historia de la demanda, no toma en cuenta muchos aspectos del proceso entero de ventas, como pueden ser la ruptura de stock, plazos de reposición de stock, precio del producto, la incidencia del marketing u otros. De igual modo se pueden descubrir tendencias, estacionalidad, ciclos, etc en la historia de la demanda para luego extrapolarlo a un tiempo futuro. También hay que destacar que el intervalo del muestreo tiene mucha influencia en el pronóstico y por ende en los resultados obtenidos [11].

En general, los métodos de series de tiempo se clasifican en: método de pronóstico del último valor, método de pronóstico por promedios, método de pronóstico de promedio móvil, método de pronóstico por suavizamiento exponencial, método de suavizamiento exponencial con tendencia y el método ARIMA (AutoRegressive Integrated Moving Average) [6].

- El método de pronóstico del último valor: Este método utiliza solo el último valor de la serie de tiempo como pronóstico del valor futuro. También es conocido como método ingenuo, porque sin mucho análisis aparentemente resulta ingenuo elegir un solo valor de toda la serie. Pero en ocasiones sí es una buena aproximación, como por ejemplo cuando hay demasiada fluctuación en la serie y entonces el último valor se convierte en el más fiable. Recomendable para series de tiempo inestables.

$$Pronóstico = último\ valor \tag{1}$$

- El método de pronóstico por promedios: En este caso se utilizan todos los valores de la serie y luego se promedia para obtener el valor de pronóstico de la serie. Recomendable para series de tiempo estables, razón por la cual todos los valores tienen el mismo peso y son considerados relevantes.

$$Pronóstico = promedio\ de\ todos\ los\ datos\ hasta\ la\ fecha \tag{2}$$

- El método de pronóstico de promedio móvil: Consiste en considerar solamente los últimos n períodos y luego promediarlo para así obtener el valor de pronóstico de la serie. Recomendable para series de tiempo medianamente estables, razón por la cual se toman en cuenta únicamente n valores que tienen el mismo peso y que son considerados relevantes.

$$Pronóstico = promedio\ de\ los\ últimos\ n\ valores \tag{3}$$

donde $n = número\ de\ periodos\ más\ recientes$

- El método de pronóstico por suavizamiento exponencial: Con este método se asignan pesos diferentes a los valores de la serie. El último período es el de mayor peso y así paulatinamente se van asignando pesos cada vez menores a los valores mas antiguos de la serie. Este resultado se puede obtener de forma simple y sintética mediante una combinación del último valor de la serie y del valor pronosticado para este mencionado último valor.

$$Pronóstico = \alpha\,(último\ valor) + (1\text{-}\alpha)\,(último\ pronóstico) \tag{4}$$

donde α (alfa) es una constante entre 0 y 1 llamada "constante de suavizamiento". Para series de tiempo estables es recomendable valores de α pequeños como 0.1, y para series de tiempo inestables valores mayores. En general en aplicaciones de hoy en día se utilizan valores entre 0.1 y 0.3.

- El método de suavizamiento exponencial con tendencia:

$$Pronóstico = \alpha\,(último\ valor) + (1\text{-}\alpha)\,(último\ pronóstico) + Tendencia\ estimada \tag{5}$$

$$Tendencia\ estimada = \beta\,(última\ tendencia)\ (1\text{-}\beta)\,(estimación\ anterior) \tag{6}$$

$$última\ tendencia = \alpha\,(último\ valor - penúltimo\ valor) + (1\text{-}\alpha)\,(último\ pronóstico - penúltimo\ pronóstico) \tag{7}$$

donde β (beta) es una constante de suavizamiento de tendencia entre 0 y 1. La elección del valor y rango de β tienen igual significado que α.

II-C2. Pronósticos causales [6]: Ciertamente las series de tiempo se basan en un solo indicador clave, como lo es por ejemplo la variable ventas. Siguiendo el ejemplo, el objetivo de la serie de tiempo es encontrar un valor de pronóstico de la variable ventas a partir de valores pasados de la misma variable ventas. Ahora bien, si tenemos dos variables en relación causa-efecto las series de tiempo no nos sirven.

El pronóstico causal obtiene una proyección de la cantidad de interés (la variable dependiente) relacionándola directamente con una o más cantidades (las variables independientes) que impulsan la cantidad de interés. Por ejemplo, las promociones sobre uno o varios productos pueden ser la causa de una mayor cantidad de ventas en dichos artículos, como tal tenemos una relación causa (promociones)-efecto (mayores ventas).

Una de las técnicas para resolver problemas de pronósticos causales es la *regresión lineal*. El objetivo de este método es encontrar la linea recta que más se aproxime a la relación entre la variable dependiente e independiente. La forma de la ecuación es la de la recta:

$$y = a + bx \tag{8}$$

Donde:

$$y = variable\ dependiente$$

$$x = variable\ independiente$$

$$a = intersección\ de\ la\ línea\ con\ el\ eje\ y$$

$$b = pendiente\ de\ la\ línea$$

Para obtener a y b se utiliza el método llamado de *mínimos cuadrados*, que encuentra el par de valores a y b tal que la suma del cuadrado de los errores de estimación sea la menor posible. Para problemas donde se consideran varios indicadores clave como variables independiente, la ecuación presenta la siguiente forma:

$$y = a + b_1x_1 + b_2x_{2+\ldots+}b_nx_n \tag{9}$$

Donde el proceso de obtención de a y $b_1, b_2, ..., b_n$ es también por el método de *mínimos cuadrados*.

II-C3. Precisión de los métodos de pronóstico: Para medir la desviación que hay entre el pronóstico y el valor real posterior se utiliza normalmente el valor de *error promedio del pronóstico*, conocido como *Desviación Absoluta Media (MAD)* y que se calcula con la siguiente fórmula:

$$MAD = \frac{suma\ de\ los\ errores\ de\ pronóstico}{número\ de\ pronósticos} \tag{10}$$

Otra forma significativa de medir la precisión es a través del *Error Cuadrático Promedio (MSE)* y que se calcula de la siguiente forma:

$$MSE = \frac{suma\ de\ los\ cuadrados\ de\ los\ errores\ de\ pronóstico}{número\ de\ pronósticos} \tag{11}$$

El resultado de esta fórmula pone de relieve los errores grandes de pronóstico, así como también destaca si el método de pronóstico es preciso. Se utiliza como complemento informativo a MAD.

II-D. Relación con el problema de estudio

Esta sección pretendió dar a conocer las técnicas de pronósticos de demanda ampliamente conocidas y estudiadas en los diferentes textos consultados [8][15][6] [2]. Si bien el presente trabajo también busca encontrar una estimación de pronóstico como lo hacen las técnicas de esta sección, debe quedar claro que la solución propuesta es una alternativa distinta, es decir, no se trata de un método cuantitativo o cualitativo propiamente dicho, sin embargo toma ciertos aspectos de ambos y se implementa con conceptos, tecnologías y herramientas de solución totalmente diferentes.

En las secciones tres y cuatro se combinan los conceptos, herramientas y tecnologías que modelan la solución de esta nueva técnica de pronóstico. En la sección quinta se realizan las experimentaciones en base al caso de estudio y se evaluan los resultados obtenidos.

III. Business Intelligence

En la actualidad Business Intelligence esta siendo cada vez más adoptado por las organizaciones debido a la necesidad de los mandos superiores de contar con información y su importancia a nivel estratégico en la toma de decisiones, en este capítulo se presenta los componentes de Business Intelligence, la definición de los indicadores clave de rendimiento y la asignación de etiquetas a cada conjunto de valores de indicadores que son el punto de entrada para el aprendizaje automático en el modelado de la solución del pronóstico de la demanda.

III-A. Definición

Business Intelligence engloba un conjunto de conceptos, técnicas basadas en computadoras y herramientas para analizar y transformar los datos empresariales en información significativa y útil que permite a las organizaciones una visión y toma de decisiones estratégicas, tácticas y operativas más efectivas. Las tecnologías de Business Intelligence ofrecen vistas históricas, actuales y predictivas de las operaciones, son procesos que se extienden en el tiempo, capaces de manejar grandes volúmenes de datos que ayudan a identificar, crear y desarrollar nuevas estrategias de negocios para mejorar la competitvidad. La era actual de las tecnologías de la información ha llevado a la necesidad de tener mejores, mas rápidas y más eficientes métodos de extraer lo datos de una organización, transformarlo en información y distribuirlo a las cadenas de mando. Business Intelligence responde a esa necesidad [3][9].

Una definición más formal que propone The Datawarehouse Institute es[4]:

"*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'*".

III-A1. Objetivos: Según lo expuesto en la definición del término Business Intelligence podemos decir que tiene los siguientes objetivos principales[3]:

- 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-A2. Componentes de Business Intelligence: Implementar un proyecto de Business Intelligence en una organización es un proceso que sigue una serie de pasos, cada paso puede verse como un componente. En la siguiente gráfica observamos los distintos componentes que forman parte de Business Intelligence.[3]

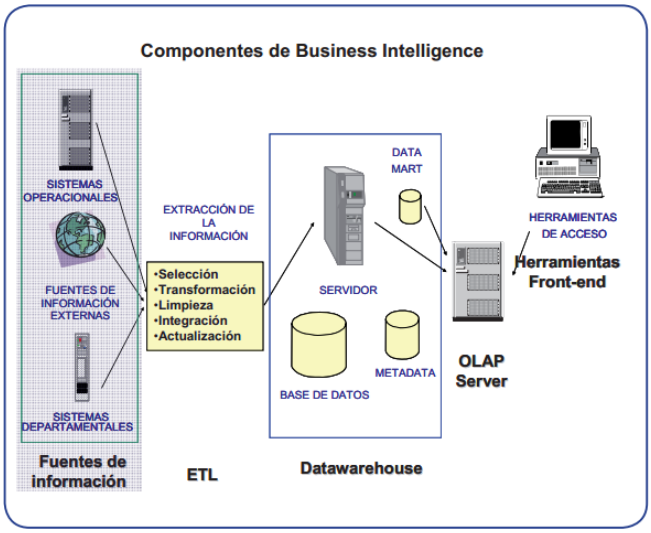

Figure 1. Componentes de Business Intelligence

A continuación, una breve descripción de los componentes de Business Intelligence:

- Fuentes de información, de los cuales se obtienen los datos que se almacenan en el datawarehouse. Básicamente son los sistemas operacionales o transaccionales de la organización.
- 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.

[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*

- Las herramientas de visualización, que nos permiten el análisis y la navegación a través de los mismos.

III-B. 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. Una característica de los KPI es que todos los KPI son indicadores, pero no todos los indicadores son KPI, otra característica es que cada organización debe definir sus propios KPI que desean tener siempre presente para manejar su rumbo, estas varían de acuerdo según la actividad realizada, el tipo de producto o la estrategia de negocios, por dicho motivo los KPI no pueden copiarse de una organización a otra ya que cada organización es diferente y requiere de una reflexión estratégica de los cuales saldrán los correspondientes KPI. 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 características del entorno competitivo actual es que se deben tomar decisiones de forma rápida y antes de que lo hagan los demás competidores[1].

En la siguiente sección haremos uso de los conceptos, herramientas y tecnologías que nos provee Business Intelligence para obtener los datos que ayudarán a modelar una solución al problema de estudio sobre pronóstico de la demanda. Iniciaremos con una breve descripción de la fuente de información, el proceso ETL (Extracción, Transformación y Carga), la especificación del datawarehouse, la definición de los indicadores clave de rendimiento y finalmente el etiquetado a cada tupla de indicadores que se transformarán en datos de entrada para el proceso aprendizaje automático.

III-C. Planteamiento del problema

Esta sección se centra en los tres primeros componentes de Business Intelligence siguiendo el proceso de modelado dimensional[10]. Uno de los principales problemas con los que se enfrentan las empresas retail[2], trata acerca de pronosticar la demanda, es decir, determinar la cantidad de productos que se deben disponer para satisfacer la demanda de los clientes para el siguiente periodo. Utilizando los conceptos y herramientas de Business Intelligence, se parte con el análisis de una fuente de información auténtica obtenida de una empresa retail; se diseña el datawarehouse que será poblado con información de las transacciones operacionales diarias del negocio; posteriormente se definen los indicadores claves de rendimiento como métricas, se calculan los valores de los indicadores con los datos históricos almacenados en el dataware para cada periodo de tiempo establecido y se asigna una etiqueta a cada tupla de valores KPI que finalmente servirán como parámetros de entrada de las herramientas de aprendizaje automático para diseñar una solución que pronostique la demanda.

[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.

III-C1. Fuente de Información: Para el presente trabajo, se cuenta con una base de datos relacional Oracle 10g con las operaciones transaccionales de una empresa retail dedicada a la venta de productos alimenticios y artículos de limpieza, algunas de las líneas de productos con que cuenta la empresa son: aceites corporales, acondicionadores, aromatizantes, cuidado corporal, desodorantes, limpiadores, salud e higiene, salud y belleza, aguas, gaseosas, cervezas, vinos, chocolates, galletitas, enlatados, lácteos, yerbas y varias líneas de productos más. La base de datos almacena datos de la operaciones comprendidas entre los periodos de noviembre del 2013 a octubre de 2016. A continuación una reseña de las principales tablas tenidas en cuenta para el diseño del dataware.

- **Tabla de Productos**: almacena información como descripción, costo, precio de venta, categoría de los artículos disponibles para la venta, cuenta con 13.200 artículos registrados.
- **Tabla Proveedores**: almacena información como descripción, dirección, ruc de los proveedores de la empresa, cuenta con 1.623 proveedores registrados.
- **Tabla de Ventas Cabecera**: es una de las tablas principales donde se registran los movimientos de ventas de la empresa. Contiene información como número de factura, fecha, cliente, montos totales, cuenta con 301.316 registros de ventas
- **Tabla de Ventas Detalle**: contiene los registros de los productos que fueron comercializados, cada detalle está relacionado a un registro de venta cabecera. Contiene datos de la fecha, el producto, precio de costo, precio de venta, cantidad y otros datos, cuenta con 981.402 registros.
- **Tabla de Movimientos de Stock:** contiene los registros de movimientos de stock de ventas y compras detallado. Contiene datos de fecha, producto, cantidad, tipo de movimiento, costo, cuenta con 1.062.440 registros.

III-C2. Proceso ETL.: Durante el proceso ETL se realizó la limpieza y transformación de los datos de origen, en la tabla de Productos se detectaron artículos con datos de proveedor nulo al cual se asignó un proveedor por defecto de la tabla dimensional de proveedores, artículos que tenían costo cero, en tal caso dichos valores eran asignados con un costo promedio tomados de la tabla de Ventas Detalle en caso de ocurrencia, de igual manera en la tabla de Ventas Detalle se detectaron registros donde los valores de costo eran iguales a cero, los cuales eran modificados por el costo promedio del producto.

III-C3. Datawarehouse: El datawarehouse se diseña a partir de la definición de las tablas de hechos y dimensiones utilizando el modelo de esquema en estrella [10].

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 IVA.
- **Tabla de hechos Detalles:** almacena los datos históricos del detalle, cada registro guarda información del número

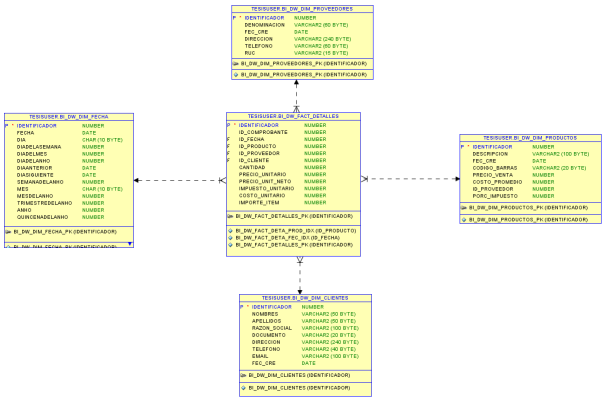

Figure 2. Tabla de hechos Detalles

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 dimensión 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.

En la siguiente figura observamos un ejemplo del modelo en esquema estrella de la tabla de hechos Detalles.

III-C4. Definición de los KPI: En el marco de esta tesis, en esta sección se definirán los KPI que se utilizarán como datos de entrada en las herramientas de aprendizaje automático para el modelado de una solución para pronosticar la demanda y como consecuencia la reposición de stock del siguiente periodo de tiempo[1] (Ej.: cantidad a comprar para satisfacer la demanda de la siguiente semana, quincena, o mes), estos KPI fueron adaptados a la solución planteada debido a que en los textos los KPI difinidos engloban a todas las áreas de la organización como compras, ventas, marketing, recursos humanos. 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.

$$X = \frac{\sum (Cantidad)}{Total\,Tickets\,Periodo} \tag{12}$$

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.

$$X = \sum (Precio * Cantidad) \tag{13}$$

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.

$$X = \frac{\sum ((Precio - Costo) * Cantidad)}{\sum (Precio * Cantidad)} * 100 \tag{14}$$

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.

$$X = \frac{\sum (Total\,Ventas\,Periodo)}{\left(\frac{Stock\,Inicial - Stock\,Final}{2}\right)} \tag{15}$$

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.

$$X = \left(\sum (Precio - Costo) * Cantidad\right) * Rotacion\,Stock \tag{16}$$

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.

$$X = \frac{Stock\ Actual}{Promedio\,Cantidad\,Venta\,Ultimos\,3\,Periodos} \quad (17)$$

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. Para el cálculo de los valores de los KPI se establecieron restricciones, que si un producto no fuese vendido por una cantidad consecutiva de periodos esta era descartada, ya que los valores de los cálculos para cada KPI daban cero el cual no tiene relevancia para el modelado.

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 puntos focales más importantes para el modelado mediante el aprendizaje automático. Para una mayor fiabilidad esta asignación de etiquetas debe ser realizada y revisada por el experto del área de compras (que podría ser el gerente de administración de compras u otra persona a cargo de la reposición de stock), sin embargo para el presente trabajo el etiquetado fue realizado en forma empírica, sin la intervención de un experto por la dificultad de contar con una persona especializada en el área. 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 pronóstico para la reposición de stock mediante algoritmos de aprendizaje automático.

En la figura a continuación un ejemplo del etiquetado para los valores de los KPI correspondientes al periodo semanal.

III-D. Resumen de la sección

Hemos visto como los componentes de Business Intelligence ayudaron a obtener el conjunto de valores para los KPI a partir de los datos históricos almacenados en el dataware y el etiquetado a cada tupla de valores KPI, estos datos

KPI TIKET MEDIO	KPI CIFRA VENTAS	KPI MARGEN COMERCIAL	KPI ROTACION STOCK	KPI COEF RENTABILIDAD	KPI COBERTURA STOCK	CANTIDAD	AÑO	MES	SEMANA	RESULTADO
4667	28000	12008	0.483	5797	2.571	7	2013	12	49	Mucho
4000	4000	1715	0.061	104	3.4	1	2013	12	50	Nada
4000	20000	8577	0.27	2318	4.364	5	2013	12	51	Nada
4000	16000	6862	0.211	1445	4.846	4	2013	12	52	Nada
4000	8000	3431	0.125	429	5.1	2	2013	12	53	Nada
4000	20000	8577	0.4	3431	4.091	5	2014	1	1	Medio
6000	12000	5146	0.353	1816	2.727	3	2014	1	2	Nada
4000	12000	5146	0.353	1816	2.1	3	2014	1	3	Nada
5600	28000	12008	1.077	12932	2.727	7	2014	1	4	Medio

Figure 3. Etiquetado de KPI del periodo semanal

serán utilizados como un conjunto de entrenamiento por los algoritmos de aprendizaje automático del cual se obtendrá un modelo de solución al problema de estudio planteado de pronóstico de la demanda.

IV. Machine Learning

En esta sección se realiza un breve repaso sobre conceptos básicos que envuelven a Machine Learning. El objetivo es visualizar qué aspectos de Machine Learning fueron tomados como componentes de solución al problema de estudio.

IV-A. Definición

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"* [13]. Este pionero de machine learning ya presagiaba que los programas, a partir del aprendizaje sobre datos históricos (la experiencia), podrían efectuar tareas de toma de decisiones sin ser programadas explícitamente dichas decisiones.

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"[12]. 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. Categorías de los 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. El problema de estudio utiliza algoritmos de aprendizaje supervisado, donde el experto en compras da la respuesta correcta a 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. El problema de estudio se encara como un problema de clasificación, donde hay cuatro 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. Problemas de clasificación

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

IV-D1. Clasificación binaria (binary classification): En su forma más simple se reduce a la siguiente cuestión: dado un patrón x extraído de un dominio X, estimar qué valor asumirá una variable aleatoria binaria asociada $y \in \{\pm 1\}$ [14].

IV-D2. Clasificación multiclase (multiclass classification): Es la extensión lógica de la clasificación binaria. La principal diferencia es que ahora $y \in \{1, 2, 3.., N\}$ puede asumir un rango de valores diferentes [14]. El problema de estudio utiliza clasificación multiclase, donde $y \in \{Nada, Poco, Medio, Mucho\}$

IV-E. Instancias, conjuntos de entrenamiento y testeo [16]

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). Para el problema de estudio el proceso de Business Intelligence es quien provee las instancias.

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-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 [5]. En el problema de estudio se utiliza el conjunto de algoritmos de clasificación de Weka [17]. En la tabla VII se listan los algoritmos de clasificación en Weka asociados a los nombres de algoritmos o estrategias que implementan.

Tabla VII
CLASIFICADORES WEKA [7]

Clasificadores Weka	Algoritmo que implementa
BayesNet	Bayes Network learning algorithms.
NaiveBayes	Naive Bayes classifier.
NaiveBayesUpdateable	Updateable version of NaiveBayes.
Logistic	Using multinomial logistic regression model.
MultilayerPerceptron	Uses backpropagation to classify instances.
SimpleLogistic	LogitBoost with simple regression.
SMO	Sequential Minimal Optimization algorithm for training a support vector classifier.
OneR	Using a 1R classifier.
DecisionTable	Using a simple decision table majority classifier.
JRip	Implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER).
PART	PART decision list. Uses separate-and-conquer. Builds a partial C4.5 decision tree.
ZeroR	Using a 0-R classifier.
DecisionStump	Using a decision stump. Usually used in conjunction with a boosting algorithm.
J48	Pruned or unpruned C4.5 decision tree.
LMT	Logistic model trees. Classification trees with logistic regression.
RandomForest	Forest of random trees.
REPTree	Fast decision tree learner. Builds a decision/regression tree.

IV-G. Evaluación del aprendizaje

La evaluación es la clave para lograr avances reales en el aprendizaje automático. Entre las técnicas de evaluación se destacan la Validación Cruzada (Cross-Validation) y la Validación Cruzada k-pliegues Estratificado (Stratified k-fold Cross-Validation).

La técnica de Cross-Validation consiste en 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 Stratified k-fold Cross-Validation, donde 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 [16].

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 [16]

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 se cuenta como 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 se 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". Si los evaluadores están totalmente de acuerdo Kappa alcanza su valor máximo igual a 1. Si no hay total acuerdo entre los evaluadores, entonces Kappa tiene un valor < 1.

$$\kappa = \frac{Pr\,(a) - Pr\,(e)}{1 - Pr\,(e)} \tag{18}$$

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.

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

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.

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

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).

$$F - Measure = \frac{2 * Recall * Precision}{(Recall + Precision)} \tag{21}$$

V. Experimentación

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 proceso de entrenamiento y de evaluación, y cuáles son las métricas de evaluación consideradas para medir el rendimiento de los clasificadores.

Para resumir la técnica propuesta, en la siguiente figura se ilustra el mapa mental general.

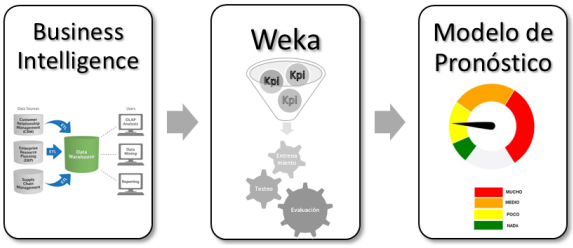

Figura 4. Mapa mental general

V-A. Datos proveídos por Business Intelligence

La salida de Business Intelligence provee tres conjuntos de datos independientes que se corresponden con los períodos de análisis: Mensuales, Quincenales y Semanales.

- Períodos Mensuales: Se analizaron 309 productos diferentes y por cada producto se tiene un máximo de 34 instancias. Cada instancia tiene los siguientes atributos: Ticket Medio, Cifra de Ventas, Margen Comercial, Rotación de Stock, Coeficiente de Rentabilidad, Cobertura de Stock, Cantidad, Año, Mes. La clase de cada instancia está definido por $y \in \{Nada, Poco, Medio, Mucho\}$.
- Períodos Quincenales: Se analizaron 228 productos diferentes y por cada producto se tiene un máximo de 68 instancias. Cada instancia tiene los siguientes atributos: Ticket Medio, Cifra de Ventas, Margen Comercial, Rotación de Stock, Coeficiente de Rentabilidad, Cobertura de Stock, Cantidad, Año, Quincena. La clase de cada instancia está definido por $y \in \{Nada, Poco, Medio, Mucho\}$.
- Períodos Semanales: Se analizaron 127 productos diferentes y por cada producto se tiene un máximo de 151 instancias. Cada instancia tiene los siguientes atributos: Ticket Medio, Cifra de Ventas, Margen Comercial, Rotación de Stock, Coeficiente de Rentabilidad, Cobertura de Stock, Cantidad, Año, Semana. La clase de cada instancia está definido por $y \in \{Nada, Poco, Medio, Mucho\}$.

V-B. Esquema general de procesamiento

Se implementa el siguiente esquema de procesamiento con el dataset:

- Instancias de Períodos Mensuales: Por cada producto, se realiza el entrenamiento y testeo de sus instancias con todos los algoritmos de clasificación posibles, luego se analizan las métricas de desempeño arrojadas por cada algoritmo y finalmente se elige el mejor clasificador

que será utilizado para la predicción de la demanda en períodos mensuales futuros.
- Instancias de Períodos Quincenales: Por cada producto, se realiza el entrenamiento y testeo de sus instancias con todos los algoritmos de clasificación posibles, luego se analizan las métricas de desempeño arrojadas por cada algoritmo y finalmente se elige el mejor clasificador que será utilizado para la predicción de la demanda en períodos quincenales futuros.
- Instancias de Períodos Semanales: Por cada producto, se realiza el entrenamiento y testeo de sus instancias con todos los algoritmos de clasificación posibles, luego se analizan las métricas de desempeño arrojadas por cada algoritmo y finalmente se elige el mejor clasificador que será utilizado para la predicción de la demanda en períodos semanales futuros.

V-C. Entrenamiento y evaluación

Como se mencionó en el esquema general de procesamiento, el entrenamiento y testeo se realiza con todos los algoritmos de clasificación posibles, para ello se utiliza la herramienta WEKA y los algoritmos de clasificación que implementa según la tabla VII. Otra forma de categorizar los clasificadores incluidos en WEKA es como sigue:

- 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.

En el siguiente pseudocódigo se presenta la estrategia de aprendizaje y selección de los clasificadores.

Algorithm 1 Pseudocódigo para el proceso de clasificación.

```
for cada período de análisis {mensual,quincenal,semanal}:
    for cada producto con sus instancias:
        establecer conjunto de entrenamiento;
        establecer conjunto de testeo;
        for cada algoritmo de clasificación:

            construir clasificador (conjunto de entrenamiento);
            evaluar clasificador (conjunto de testeo);
            obtener métricas de evaluación;
        endfor;
        criterios de línea de base (ZeroR, criterios del experto u otro);
        seleccionar mejor clasificador (max(Kappa));
        guardar clasificador;
    endfor;
endfor;
```

La evaluación se hace por el método Stratified k-fold Cross Validation para un valor de k igual a 10 y las métricas de desempeño consideradas son el *Porcentaje de Aciertos* y la *Estadística Kappa*. El criterio de línea de base utilizado fue el clasificador ZeroR, el cual es uno de los criterios de línea de base más representativos para problemas de clasificación. Se considera que también puede resultar conveniente que el experto en compras establezca su propio criterio de línea de base, como puede ser un umbral mínimo de porcentaje de aciertos aceptado.

Por cada producto y período de análisis se elige como clasificador aquel que haya alcanzado el mayor valor de *Kappa*. En la siguiente figura se muestra como quedó la distribución de clasificadores para períodos mensuales. Por ejemplo, el clasificador *Logistic* resultó una mejor solución para 149 productos.

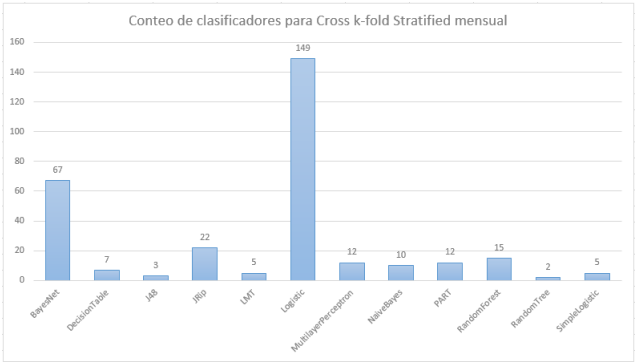

Figura 5. Conteo de clasificadores para período mensual.

Se puede observar en la siguiente gráfica de barras que la técnica propuesta alcanza altos porcentajes de aciertos en promedio, tanto para periodos mensuales, quincenales como semanales.

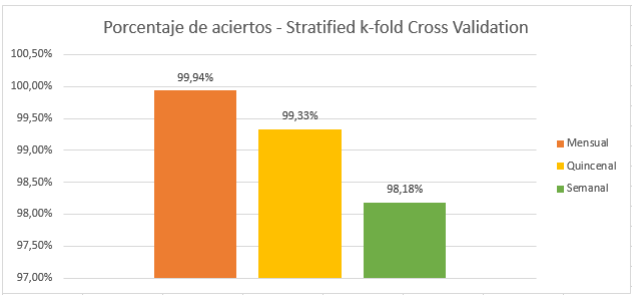

Figura 6. Promedio de porcentaje de aciertos para los tres períodos de análisis.

Como se trata de una prueba exhaustiva, por cada producto se intenta con todos los algoritmos de clasificación posible y se evalúa con el método Stratified 10-fold Cross Validation. Estas métricas de desempeño preliminares dan indicio de que la técnica propuesta en este trabajo puede alcanzar altos grados de confiabilidad. 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.

V-D. Cómo hacer las predicciones

En el siguiente pseudocódigo se muestra el mecanismo para obtener el pronóstico de la demanda en un ambiente de producción.

Algorithm 2 Pseudocódigo para el proceso de pronóstico de la demanda.

```
for cada próximo período a pronosticar {mensual,quincenal,semanal}:
    for cada producto:
        obtener KPIs del período actual finalizado;
        ejecutar su mejor clasificador (KPIs);
        obtener etiqueta {nada,poco,medio,mucho}
        extrapolar a valores continuos(criterio experto);
    endfor;
endfor;
```

Como se mencionó anteriormente, el modelo propuesto arroja como resultado un valor discreto $y \in \{Nada, Poco, Medio, Mucho\}$. Luego en función de la etiqueta resultante, del tipo de producto y del período seleccionado, el experto extrapola a un valor continuo que representa la cantidad en la orden de compra. El significado de las etiquetas varía

VI. Discusión

VI-A. Impacto del período de análisis

Una de las decisiones que se debe tomar es acerca del tiempo asignado al período de análisis. En este trabajo se analizaron tres períodos distintos: mensuales, quincenales y semanales con propósitos experimentales y por ser los más comunes en el ámbito comercial. En la práctica, la elección del período es una decisión estratégica a nivel gerencial que depende en gran medida del sector y tamaño de la empresa, tipos de productos, etc.

En el presente trabajo, por tratarse de períodos de tiempo muy cercanos (1, 2 y 4 semanas) no se observan diferencias significativas en el porcentaje de aciertos. Otro factor a tener en cuenta es que para períodos de tiempo muy extensos (6, 12 meses) existe mayor incertidumbre en el pronóstico.

VI-B. Impacto del etiquetado

La técnica propuesta se trata de un sistema parametrizado, donde las variables principales son el período comercial y las etiquetas seleccionadas para la clasificación. Por cuestiones de practicidad y generalidad se eligió para este trabajo un enfoque de problema de clasificación. El etiquetado proporciona mayor flexibilidad al sistema y un entorno más controlable, en comparación a un sistema de asignación de valores continuos. La flexibilidad del sistema permitió emular la opinión del experto en compras y encontrar una cantidad eficiente de etiquetas.

VII. Conclusiones

Este trabajo se enfocó en proponer una nueva técnica de estimación de la demanda de productos, para reposición de stock en empresas retail. Como se mencionó en la Sección 2, la gestión de compras es uno los ejes centrales en la actividad empresarial y la decisión del volumen de compras para cada producto es un desafío que enfrentan las empresas al momento de reponer el stock. Partiendo de esta premisa y analizando las técnicas de pronóstico de la demanda empleadas en la actualidad, y el creciente incremento del uso de tecnologías de Business Intelligence en las organizaciones, se encontró la oportunidad de desarrollar una nueva técnica de pronóstico. En esta nueva técnica se utilizan los Indicadores Claves de Rendimiento y apoyados en la experiencia de un experto en compras (gerente o encargado de compras) se realiza el modelado utilizando algoritmos de clasificación de Machine Learning.

De acuerdo a los resultados experimentales se obtuvieron altas tasas de aciertos, haciendo pruebas exhaustivas con varios algoritmos de clasificación y evaluando con un método ampliamente aceptado. La técnica propuesta pretende que este nuevo modelo se convierta en una herramienta de apoyo en la toma de decisiones del gerente de compras en el proceso de reposición de stock.

Referencias

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

[2] David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, and Kipp Martin. *MÃl'todos cuantitativos para los negocios*. Âl' D.R. 2011 por Cengage Learning Editores, S.A. de C.V., una compaÃśÃ■a de Cengage Learning, Inc, 11 edition, 2011.

[3] Josep LluÃ■s Cano. *Busines Intelligence: Competir con informaciÃşn*. ESADE, Banesto, Banesto Pyme, 2007.

[4] Wayne W. Eckerson and Cindi Howson. Enterprise business intelligence: Strategies and technologies for deploying bi on an enterprise scale tdwi report series. 2005.

[5] Machine Learning Group. Weka 3: Data mining software in java.

[6] Frederick S. Hillier and Mark S. Hillier. *MÃl'todos cuantitativos para administraciÃşn*. Tercera edition, 2008.

[7] http://weka.sourceforge.net/doc.dev/weka/classifiers/Classifier.html. Interface classifier.

[8] 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.

[9] Jordi Conesa Josep Curto. *IntroducciÃşn al Busines Intelligence*. Editorial UOC, 2010.

[10] Ralph Kimball. *The datawarehouse Toolkit*. John Wiley & Sons, Inc, 1992.

[11] Arley PÃl'rez, Alberto Medina, Pavel Alonso, and Nguyen RamÃ■rez. MÃl'todos y tÃl'cnicas para la previsiÃşn de la demanda. *Universidad de Matanzas Camilo Cienfuegos - Facultad Industrial-EconomÃ■a*, 2007.

[12] Jean Francois Puget. What is machine learning?, May 2016.

[13] Arthur Samuel. Some studies in machine learning using the game of checker. *IBM Journal 3, 211-229*, 1959.

[14] Alex Smola and S.V.N. Vishwanathan. *Introduction to Machine Learning*. The Press Syndicate of The University of Cambridge, 2008.

[15] Naim Caba Villalobos, Oswaldo Chamorro Altahona, and TomÃąs JosÃl' Fontalvo Herrera. *GestiÃşn de la ProducciÃşn y Operaciones*. 2011.

[16] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining - Practical Machine Learning Tools and Techniques - Third Edition*. Copyright Âl' 2017 Elsevier Inc. All rights reserved, tercera edition, 2011.

[17] 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.