“Automatización en la toma de decisiones de compras de productos para reposición de stock a través de un modelo 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 importante que tienen 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 un producto específico y en qué cantidades. En este trabajo se propone un modelo que ayuda a tomar las decisiones de volumen de compra de productos disponibilizado para la venta, a traves de un modelo basado en business intelligence y machine learning.**

***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.

En el proceso de gestión de compras en empresas como las retail, se utilizan técnicas de pronósticos para determinar las cantidades de las órdenes de compra y que pueden estar basadas en pronósticos cuantitativos o en pronósticos 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.

En el presente trabajo se elabora un 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 modelo se utilizan herramientas de Inteligencia de Negocios (Business Intelligence) y de Aprendizaje Automático (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 del aprendizaje automático 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.

Por ultimo se realiza el análisis de los resultados obtenidos de los algorimos utilizados y se realiza la evaluación del desempeño de los mismos.

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. Este trabajo intenta contribuir lograr decisiones eficientes de compras basadas en estimaciones eficientes de ventas. Se considera que es una decisión importantísima estimar o

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

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:

- 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.[7]

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:

- 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:

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 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.
- Promedio móvil ponderado: Ajusta el método de promedio simple asignando mayor peso a los datos más recientes. Sirve para reflejar el nivel de importancia de unos datos sobre otros como resultado de las fluctuaciones.
- 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.
- Suavización exponencial doble: Esta técnica es una modificación del suavizamiento exponencial simple. Agrega una constante de suavización delta (DELTA), cuya fun-

ción es reducir el error que ocurre entre la demanda real y el pronóstico.

- Estacional multiplicativo

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.

III. Business Intelligence

III-A. Concepto

Business intelligence abarca un conjunto de conceptos, técnicas y herramientas que se utiliza para la transformación de simples datos en información útil y significativa para el análisis de negocios.

La definición[**?**] 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. BI abarca las tecnologías de datawarehousing, los procesos en el 'back end'*[1]*, consultas, informes, análisis y las herramientas para mostrar información (estas son las herramientas de BI) 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:

- 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.
- Disminuir sustancialmente la incertidumbre que existe ante la toma de decisiones respecto a un plan estratégico.

III-B. Componentes de BI

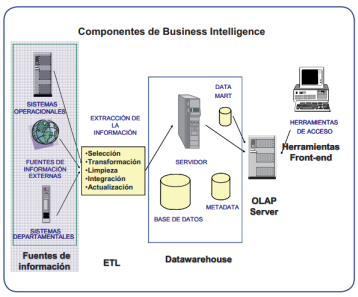

Figure 1. Componentes de Bussines Intelligence

[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*

III-B1. Fuentes de información: Las fuentes de información a las que podemos acceder son:

- Básicamente, de los sistemas operacionales o transaccionales, que incluyen aplicaciones desarrolladas a medida, ERP, CRM, SCM, etc.
- Sistemas de información departamentales: previsiones, presupuestos, hojas de cálculo, etcétera.
- Fuentes de información externa.

III-B2. ETL[2] *– Proceso de extracción, transformación y carga:* Antes de almacenar los datos en un datawarehouse, éstos deben ser transformados, limpiados, filtrados y redefinidos. Normalmente, la información que tenemos en los sistemas transaccionales no está preparada para la toma de decisiones. El proceso trata de recuperar los datos de las fuentes de información y alimentar el datawarehouse. El proceso de ETL[14] consume entre el 60% y el 80% del tiempo de un proyecto de business intelligence, por lo que es un proceso clave que requiere recursos, estrategia, habilidades y tecnologías.

La extracción, transformación y carga (el proceso ETL) es necesario para acceder a los datos de las fuentes de información al datawarehouse. El proceso ETL se divide en 5 subprocesos:

Extracción*:* Este proceso recupera los datos físicamente de las distintas fuentes de información.

Limpieza*:* Este proceso recupera los datos en bruto y comprueba su calidad, elimina los duplicados, corrige los valores erróneos y vacíos, es decir se transforman los datos para reducir los errores de carga.

Transformación*:* Este proceso recupera los datos limpios y de alta calidad y los estructura y sumariza en los distintos modelos de análisis.

Integración*:* Este proceso valida que los datos que cargamos en el datawarehouseson consistentes con las definiciones y formatos del datawarehouse; los integra en los distintos modelos de las distintas áreas de negocio que hemos definido en el mismo. Estos procesos pueden ser complejos

Actualización*:* Este proceso es el que nos permite añadir los nuevos datos al datawarehouse, determina la periodicidad con el que haremos nuevas cargas de datos al datawarehouse

III-B3. Datawarehouse o almacén de datos: Un datawarehouse es un repositorio de datos que proporciona una visión global, común e integrada de los datos de la organización con las siguientes propiedades: estable, coherente, fiable y con información histórica[8].

El profesor Hugh J. Watson [13] lo define como:

"Un datawarehouse es una colección de información creada para soportar las aplicaciones de toma de decisiones. Datawarehousing es el proceso completo de extraer información, transformarla y cargarla en un datawarehouse y el acceso a esta información por los usuarios finales y las aplicaciones."

Bill Inmon[5] fue el que definió las características que debe cumplir un datawarehouse:

- Orientado a un área: cada parte del datawarehouse está construida para resolver un problema de negocio.

[2]ETL corresponde a las siglas del inglés Extract, Transform and Load (Extracción, transformación y carga)

- Integrado: la información debe ser transformada en medidas comunes, códigos comunes y formatos comunes para ser útil.
- Indexado en el tiempo: se mantiene la información histórica.
- No volátil: los usuarios no la mantienen como lo harían en los entornos transaccionales. La información se almacena para la toma de decisiones

Ralph Kimbal[9] define los objetivos que debería cumplir un datawarehouse:

- El alcance de un datawarehouse puede ser bien un departamento o bien corporativo.
- El datawarehouse no es sólo información sino también las herramientas de consulta, análisis y presentación de la información.
- La información del datawarehouse es consistente.
- La calidad de información en el datawarehouse es el motor de business reengineering.

Existen otros elementos en el contexto de un datawarehouse :

- Datawarehousing: es el proceso de extraer y filtrar datos para transformarlos, integrarlos y almacenarlos en un almacén de datos.
- Data Mart: es un subconjunto de los datos del datawarehouse.
- Operational Data Store (ODS): es un tipo de almacén de datos que proporciona sólo los últimos valores de los datos.
- Staging Area: es el sistema que permanece entre las fuentes de datos y el datawarehouse.
- Procesos ETL: tecnología de integración de datos.
- Metadatos: datos estructurados y codificados que describen características de instancias.

Elementos de una datawarehouse:

- **Tablas de hecho**: es la representación en el datawarehouse de los procesos de negocio de la organización.
- **Dimensión**: es la representación en el datawarehouse de una vista para un cierto proceso de negocio.
- **Métrica**: son los indicadores de negocio de un proceso de negocio. Aquellos conceptos cuantificables que permiten medir nuestro proceso de negocio.

Tipos de esquemas para estructurar los datos en un datawarehouse:

- **Esquema en estrella:** consiste en estructurar la información en procesos, vistas y métricas recordando a una estrella.
- **Esquema en copo de nieve:** es un esquema de representación derivado del esquema en estrella, en el que las tablas de dimensión se normalizan en múltiples tablas.

III-B4. Herramientas de Business Intelligence: Existen distintas tecnologías que nos permiten analizar y visualizar la información que reside en un datawarehouse, pero la más extendida es la OLAP (Online Analytical Processing). Los usuarios[6] necesitan analizar información a distintos niveles de agregación y sobre múltiples dimensiones.

Las principales herramientas[3] de Business Intelligence son: :

- Generadores de informes: Utilizadas por desarrolladores profesionales para crear informes estándar para grupos, departamentos o la organización.
- Herramientas de usuario final de consultas e informes: Empleadas por usuarios finales para crear informes para ellos mismos o para otros; no requieren programación.
- Herramientas OLAP: Permiten a los usuarios finales tratar la información de forma multidimensional para explorarla desde distintas perspectivas y periodos de tiempo.
- Herramientas de Dashboard y Scorecard: Permiten a los usuarios finales ver información crítica para el rendimiento con un simple vistazo utilizando iconos gráficos y con la posibilidad de ver más detalle para analizar información detallada e informes, si lo desean.

III-C. KPI

Los KPI *(Key Performance Indicators)* se trata de indicadores que son determinantes para analizar de forma rápida la marcha del negocio y que nos permiten tomar decisiones. Todos los KPI son indicadores, pero no todos los indicadores son KPI[2].

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" y una de las caracteristicas de nuestro entorno competitivo actual es que tenemos que tomar decisiones de forma rápida y antes de que lo hagan los demás.

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[3], la cual trata acerca de la reposición de stock[4], 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 reposision 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

[3]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.

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

datos de la fecha, numero 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. Dataware:

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 numero 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:

- **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 dias, 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.
- **Dimesión 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.
- **Dimension Cajas:** La tabla de dimensión Cajasa 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 (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.

TICKET MEDIO.: Es el importe medio por cada transacción de compra que se realiza de un determinado producto. El indicador viene determinado por dos variables: El importe total vendido del producto y el total de tickets en las que fue vendido el producto. Aplicando la siguiente fórmula obtenemos el valor de importe medio de venta para cada producto.

$$x = \frac{\sum(\text{Cantidad})}{\text{TotalTickes del periodo}}$$

Figure 2. Fórmula de Ticket Medio

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.

$$x = \sum(\text{Precio} * \text{cantidad})$$

Figure 3. Fórmula Cifra de Ventas

MARGEN COMERCIAL: Es la diferencia entre el precio de venta y precio de costo del producto, es un indicador que permite conocer la rentabilidad del producto. Se obtiene de la siguiente fórmula.

$$x = \frac{\sum((\text{PrecioVenta} - \text{Costo}) * \text{cantidad})}{\sum(\text{Precio} * \text{cantidad})} * 100$$

Figure 4. Fórmula Margen Comercial

ROTACIÓN DE STOCK: El KPI mide la cantidad de veces que el stock del producto se renueva durante un determinado ciclo comercial. Se obtiene de la siguiente fórmula.

$$x = \frac{\sum(\text{Total de ventas del periodo})}{\textit{Stock Promedio del periodo}}$$

Total de ventas del periodo = Cantidad total vendida del producto.

Stock Promedio = (Stock inicial – Stock final)/2

Figure 5. Fórmula Rotación Stock

COEFICIENTE DE RENTABILIDAD: El KPI 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.

$$x = (\sum(\text{PrecioVenta} - \text{Costo}) * \text{Cantidad}) * \text{Rotacion de stock}$$

Rotación de Stock = Valor Obtenido por la formula anterior

Figure 6. Fórmula Coeficiente de Rentabilidad

COBERTURA DE STOCK: Indica 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 = Stock actual / Promedio cantidad venta último 3 periodos

Figure 7. Fórmula Cobertura de Stock

III-C5. CÁLCULO DE LOS VALORES DE LOS KPI: Definidos los KPI a ser utilizados, obtenemos los valores para cada KPI de los datos almacenados en el datawarehouse, para ello codificamos a sentencias SQL las fórmulas detalladas mas arriba y almacenamos la información de los resultados en la estructura 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:

Se toma como base los siguientes KPI.

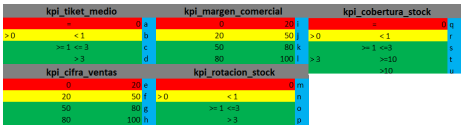

Figure 8. Tabla de rangos de KPI

Una vez asignado los valores “a”, “b”, “c”, “d”, “e” hasta “u” en la tabla de etiquetado del experto de la figura 9, se busca y se asigna el valor de la etiqueta correspondiente.

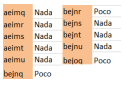

Figure 9. Tabla de etiquetado por el experto

Una vez finalizado el etiquetado, 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 ID 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. Aprendizaje Automático o 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”* [11]. Lo que nos lleva a pensar que uno pioneros del aprendizaje automático 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. Samuel define al aprendizaje automático como sigue: *“El aprendizaje automático es un campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programadas”*.

Otro investigador de aprendizaje automático 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”*.

IV-A. Definición

“The purpose of machine learning is to learn from training data in order to make as good as possible predictions on new, unseen, data”[10]. La dificultad radica en que debemos construir modelos que nos acerquen a una buan predicción sobre datos aún no conocidos o imprevistos.

Machine Learning

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

Figura 10. Definición presentada por Peter Prettenhofer y Gille Louppe.

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

IV-D1. Ejemplos o instancias (examples): La entrada de un esquema de aprendizaje automático es un conjunto de instancias. Estas instancias son las cosas que deben ser clasificadas, asociadas o agrupadas. En el escenario estándar, cada instancia es un ejemplo individual e independiente del concepto que se debe aprender.:

IV-D2. Características o atributos (features): Las instancias son caracterizadas mediante los valores de un conjunto predeterminado de atributos. Cada instancia proporciona una entrada al aprendizaje automático es caracterizado por los valores en un conjunto fijo y predefinido de características o atributos [15]. :

IV-D3. Etiquetas (labels): Las cantidades nominales tienen valores que son símbolos distintos. Los valores mismos sirven como etiquetas o nombres, de ahí el término nominal, que viene de la palabra latina para nombre. Los atributos nominales a veces se llaman categorizados, enumerados o discretos.:

IV-D4. Conjunto de entrenamiento (training set): son los ejemplos o instancias utilizadas en el proceso de entrenamiento del algoritmo de aprendizaje automático.:

IV-D5. Algoritmos de aprendizaje (learning algorithms): Hipótesis, Parámetros, Función de costo, Objetivo.:

IV-D6. Conjunto de prueba (test 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.:

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 $y \in \{\pm 1\}$ [12].:

IV-E2. Clasificación multiclase (multiclass classification): Es la extensión lógica de la clasificación binaria. La principal diferencia es que ahora $y \in \{1, \ldots, N\}$ puede asumir un rango de valores diferentes [12].:

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 [1]: 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.

IV-G1. Validación Cruzada (cross-validation): En la validación cruzada, usted decide sobre un número fijo de pliegues, o particiones, de los datos. Supongamos que usamos tres, luego los datos se dividen en tres particiones aproximadamente iguales; cada uno a su vez se utiliza para las pruebas y el resto se utiliza para el entrenamiento. Es decir, utilizar dos tercios de los datos para el entrenamiento y un tercio para las pruebas, y repetir el procedimiento tres veces para que al final, cada instancia se haya utilizado exactamente una vez para la prueba. Esto se denomina triple validación cruzada, y si la estratificación se adopta también, lo que es a menudo, triple validación cruzada estratificada.:

IV-G2. Validación Cruzada K-fold Estratificado (stratified k-fold cross validation): La manera estándar de predecir la tasa de error de una técnica de aprendizaje dada una única muestra fija de datos es usar la validación cruzada diez veces estratificada. Los datos se dividen aleatoriamente en 10 partes en las que la clase se representa en aproximadamente las mismas proporciones que en el conjunto de datos completo. Cada parte se extiende a su vez y el esquema de aprendizaje entrenado en los restantes nueve décimos; entonces su tasa de error se calcula en el conjunto de retención. Así, el procedimiento de aprendizaje se ejecuta un total de 10 veces en diferentes conjuntos de entrenamiento (cada conjunto tiene mucho en común con los demás). Finalmente, las 10 estimaciones de error se promedian para obtener una estimación del error global. Pruebas extensas en numerosos conjuntos de datos diferentes, con diferentes técnicas de aprendizaje, han demostrado que 10 es sobre el número correcto de pliegues para obtener la mejor estimación de error, y también hay algunas pruebas teóricas que apoyan esto.:

IV-G3. 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. Resultados de la evaluación

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. En WEKA el método correct() obtiene el número de instancias correctamente clasificadas (es decir, para las que se realizó una predicción correcta), en realidad es la suma de los pesos de estas instancias.:

IV-H2. Porcentaje de Aciertos: Porcentaje de instancias correctamente clasificadas. En WEKA el método pctCorrect() obtiene el porcentaje de instancias correctamente clasificadas (es decir, para las que se realizó una predicción correcta).:

IV-H3. Estadística Kappa (Kappa Statistic): Ejemplos en los que los errores cuestan diferentes cantidades incluyen las decisiones de préstamo: El costo de prestar a un deudor es mucho mayor que el costo de negocio perdido de rechazar un préstamo a un no deudor. Los verdaderos positivos (TP) y verdaderos negativos (TN) son clasificaciones correctas. Un falso positivo (PF) es cuando el resultado se predice incorrectamente como sí (o positivo) cuando es realmente no (o negativo). Un falso negativo (FN) es cuando el resultado se predice incorrectamente como negativo cuando es realmente positivo. La tasa de éxito global es el número de clasificaciones correctas dividido por el número total de clasificaciones: TP +TN/ TP+ TN +FP +FN, por último, la tasa de error es 1 menos esto. En la predicción multiclase, cada elemento de matriz muestra el número de ejemplos de prueba para los que la clase real es la fila y la clase prevista es la columna. Los buenos resultados corresponden a grandes números en la diagonal principal y pequeños, idealmente cero, fuera de los elementos diagonales. Una medida denominada Kappa Statistic tiene en cuenta este factor previsto deduciéndolo de los éxitos del predictor y expresando el resultado como una proporción del total para un predictor perfecto. El valor máximo de Kappa es 100 %, y el valor esperado para un predictor aleatorio con los mismos totales de columna es 0. En WEKA el método kappa() devuelve el valor de la estadística kappa si la clase es nominal.:

IV-H4. Sensibilidad (Recall) y Precisión (Precision): La gente ha lidiado con la compensación fundamental ilustrada por los gráficos de elevación y las curvas ROC en una amplia variedad de dominios. Comparar un sistema que localiza 100 documentos, 40 de los cuales son relevantes, con otro que localiza 400 documentos, 80 de los cuales son relevantes. £Cual es mejor? La respuesta ahora debe ser obvia: depende del costo relativo de falsos positivos, documentos devueltos que no sean relevantes, y falsos negativos, documentos que son relevantes pero que no se devuelven. Los investigadores de recuperación de información definen parámetros llamados recall y precisión. Recall = number of documents retrieved that are relevant / total number of documents that are relevant. Precision: number of documents retrieved that are relevant / total number of documents that are retrieved. En WEKA el método recall(int classIndex) calcula la sensibilidad con respecto a una clase en particular, esto se define como: positivos correctamente clasificados / positivos totales. En WEKA el método precision(int classIndex) calcula la precisión con respecto a una clase en particular, esto se define como: positivos correctamente clasificados / total predicho como positivo.:

*IV-H5. Medida-F o Puntuación-F (F-Measure o F-Score): En el análisis estadístico de clasificación binaria, 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. En WEKA el método fMeasure(int classIndex) calcula la Puntuación-F con respecto a una clase en particular, esto se define como: 2 * recall * precision / recall + precision.:*

V. 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 WEKA 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-A. Datos proveídos por business intelligence

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

La Figura 11 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 la tabla de la Figura 11 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.

KPI TICKET MEDIO	KPI CIFRA VENTAS	KPI MARGEN COMERCIAL	KPI ROTACION STOCK	KPI COEF RENTABILIDAD	KPI COBERTURA STOCK	CANTIDAD	ANHO	MES	RESULTADO
1,143	16000	33,131	2	10602	1,513	8	2013	12	Poco
1,143	32000	35,909	3,2	36771	3	16	2014	1	Medio
1	10000	35,909	0,833	2992	0,25	5	2014	2	Medio
1	14000	35,909	0,824	4140	1,034	7	2014	3	Poco
1,091	24000	35,909	3	25855	0,75	12	2014	4	Medio
1,2	12000	35,909	1	4309	0,125	6	2014	5	Medio
1	16000	35,909	0,8	4596	1,32	8	2014	6	Poco
1,083	26000	35,909	1,733	16183	1,038	13	2014	7	Medio
1,4	14000	35,909	1,4	7038	0,667	7	2014	8	Medio

Figura 11. Tabla de 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 229 productos diferentes analizados para períodos quincenales, lo que equivale a 229 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 de la tabla de la Figura 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 de la tabla de la Figura 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-B. 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. Hacia el final del capítulo se muestran ejemplos de estas tablas y se analizan el significado de los resultados que contienen.

V-C. 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.

En la Figura 12 se muestra un fragmento del algoritmo que realiza el entrenamiento y la evaluación con la API de WEKA. Los “modelos” constituyen los algoritmos de clasificación, así que cada modelo se puede evaluar por Percentage Split o por Stratified K-fold Cross Validation. Para esta tesis evaluamos por ambos métodos.

```
//Por cada modelo, se construye su clasificador y se evalua.
for (int j=0; j<modelos.length; j++){

    //Se evalua el modelo.
    if (tipoEvaluacion == "split"){

        modelo.buildClassifier(this.train);
        evaluar = new Evaluation(this.test);
        evaluar.evaluateModel(modelo, this.test);

    }else if (tipoEvaluacion == "crossEstratificado"){


        modelo.buildClassifier(this.instancias);
        evaluar = new Evaluation(this.instancias);
        evaluar.crossValidateModel(modelo, this.instancias, this.folds, new Random(this.seed));
    }

    //INDICADORES RESUMEN:

    .....

    //INDICADORES DE PRECISION POR CLASE:

    .....

}
```

Figura 12. Esquema de entrenamiento y evaluación en WEKA.

V-D. Métricas de los resultados de la evaluación

En la Figura 13 se muestra qué métricas de evaluación son obtenidas de la API WEKA. Entonces 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).

```
//Por cada modelo, se construye su clasificador y se evalua.
for (int j=0; j<modelos.length; j++){

    //BUILD CLASSIFIER. EVALUATION.

    .....

    //INDICADORES RESUMEN:

    //Aciertos.
    cantidad_aciertos = evaluar.correct();
    porcentaje_aciertos = evaluar.pctCorrect();

    //Kappa.
    kappa_statistic = evaluar.kappa();

    //INDICADORES DE PRECISION POR CLASE:
    for (int q=0; q < this.etiquetas.size(); q++){

        //Area under ROC.
        roc_area = evaluar.areaUnderROC(indice_clase);

        //Recall.
        recall = evaluar.recall(indice_clase);

        //Precision.
        precision = evaluar.precision(indice_clase);

        //F-Measure.
        f_measure = evaluar.fMeasure(indice_clase);

        //Area under precision-recall curve (AUPRC).
        prc_area = evaluar.areaUnderPRC(indice_clase);
    }
}
```

Figura 13. Esquema de obtención de las métricas de evaluación en WEKA.

V-E. Tablas de resultado de las métricas de evaluación

Se muestran tablas de ejemplo que contienen los resultados de las métricas de evaluación. Como se mencionó en la sub sección 1.2, hay dos tablas que se generan al concluir el proceso de aprendizaje automático.

En la Figura 14 se muestra una pequeña parte de la tabla con los resultados de evaluación del aprendizaje automático con el método Percentage Split para los periodos mensuales, quincenales y semanales.

PRODUCTO	PERÍODO	TIPO CLASIFICADOR	CLASIFICADOR	TIPO EVALUACIÓN	% ACIERTOS	KAPPA	Nada PREC	Nada RCALL	Nada FMEA	Nada ROCA	Nada PRCA	Poco PREC	Poco RCALL	Poco FMEA	Poco ROCA	Poco PRCA	Medio PREC	Medio RCALL	Medio FMEA	Medio ROCA	Medio PRCA	Mucho PREC	Mucho RCALL	Mucho FMEA	Mucho ROCA	Mucho PRCA
100	MENSUAL	bayes	BayesNet	split	98.33	0.96	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00	1.00	0.91	1.00	0.95	1.00	0.96
100	MENSUAL	bayes	NaiveBayes	split	81.67	0.63	0.00	0.00	0.00	0.00	0.00	0.60	1.00	0.75	0.98	0.87	0.95	0.80	0.86	0.93	0.98	0.62	0.80	0.70	0.96	0.85
100	MENSUAL	bayes	NaiveBayesUpdateable	split	81.67	0.63	0.00	0.00	0.00	0.00	0.00	0.60	1.00	0.75	0.98	0.87	0.95	0.80	0.86	0.93	0.98	0.62	0.80	0.70	0.96	0.85
100	MENSUAL	functions	Logistic	split	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	MENSUAL	functions	MultilayerPerceptron	split	96.67	0.92	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.98	0.92	0.96	1.00	0.80	0.89	0.94	0.88
100	MENSUAL	functions	SimpleLogistic	split	96.67	0.92	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.98	0.99	1.00	1.00	0.80	0.89	0.99	0.97
100	MENSUAL	functions	SMO	split	95.53	0.83	0.00	0.00	0.00	0.00	0.00	1.00	0.67	0.80	0.92	0.75	0.92	1.00	0.96	0.88	0.92	1.00	0.80	0.89	0.96	0.87
100	MENSUAL	rules	OneR	split	85.33	0.49	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.81	1.00	0.90	0.69	0.81	0.00	0.00	0.00	0.50	0.17
100	MENSUAL	rules	DecisionTable	split	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	MENSUAL	rules	JRip	split	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	MENSUAL	rules	PART	split	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	MENSUAL	rules	ZeroR	split	73.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.10	0.73	1.00	0.85	0.50	0.73	0.00	0.00	0.00	0.50	0.17
100	MENSUAL	trees	DecisionStump	split	83.33	0.49	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.81	1.00	0.90	0.69	0.81	0.00	0.00	0.00	0.56	0.19
100	MENSUAL	trees	J48	split	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	MENSUAL	trees	LMT	split	96.67	0.92	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.98	0.99	1.00	1.00	0.80	0.89	0.99	0.97
100	MENSUAL	trees	RandomForest	split	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	MENSUAL	trees	RandomTree	split	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
100	MENSUAL	trees	REPTree	split	83.33	0.49	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.81	1.00	0.90	0.69	0.81	0.00	0.00	0.00	0.56	0.19

Figura 14. Resultados de la evaluación del aprendizaje automático con el método Percentage Split.

En la Figura 15 se muestra una pequeña parte de la tabla con los resultados de evaluación del aprendizaje automático con el método Stratified K-fold Cross Validation para los periodos mensuales, quincenales y semanales.

PRODUCTO	PERÍODO	TIPO CLASIFICADOR	CLASIFICADOR	TIPO EVALUACIÓN	% ACIERTOS	KAPPA	Nada PREC	Nada RCALL	Nada FMEA	Nada ROCA	Nada PRCA	Poco PREC	Poco RCALL	Poco FMEA	Poco ROCA	Poco PRCA	Medio PREC	Medio RCALL	Medio FMEA	Medio ROCA	Medio PRCA	Mucho PREC	Mucho RCALL	Mucho FMEA	Mucho ROCA	Mucho PRCA
100	MENSUAL	bayes	BayesNet	crossEstratificad	98.67	0.97	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00	0.90	1.00	0.95	1.00	1.00		
100	MENSUAL	bayes	NaiveBayes	crossEstratificad	90.00	0.76	0.00	0.00	0.00	0.00	0.00	0.82	1.00	0.90	0.99	0.94	0.96	0.90	0.93	0.91	0.97	0.67	0.78	0.72	0.95	0.72
100	MENSUAL	bayes	NaiveBayesUpdateable	crossEstratificad	90.00	0.76	0.00	0.00	0.00	0.00	0.00	0.82	1.00	0.90	0.99	0.94	0.96	0.90	0.93	0.91	0.97	0.67	0.78	0.72	0.95	0.72
100	MENSUAL	functions	Logistic	crossEstratificad	99.33	0.98	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.95	1.00	0.97	1.00	0.98	
100	MENSUAL	functions	MultilayerPerceptron	crossEstratificad	97.33	0.93	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.98	0.94	0.93	1.00	0.78	0.88	0.90	0.88
100	MENSUAL	functions	SimpleLogistic	crossEstratificad	98.00	0.95	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.97	0.99	1.00	0.86	1.00	0.92	0.99	0.95		
100	MENSUAL	functions	SMO	crossEstratificad	90.00	0.71	0.00	0.00	0.00	0.00	0.00	1.00	0.50	0.67	0.86	0.60	0.90	0.98	0.94	0.81	0.89	0.88	0.78	0.82	0.96	0.75
100	MENSUAL	rules	OneR	crossEstratificad	91.33	0.78	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.95	0.94	0.94	0.89	0.94	0.63	0.67	0.65	0.81	0.46	
100	MENSUAL	rules	DecisionTable	crossEstratificad	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
100	MENSUAL	rules	JRip	crossEstratificad	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
100	MENSUAL	rules	PART	crossEstratificad	99.33	0.98	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.95	1.00	0.97	1.00	0.95	
100	MENSUAL	rules	ZeroR	crossEstratificad	76.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.11	0.76	1.00	0.86	0.46	0.74	0.00	0.00	0.00	0.43	0.11
100	MENSUAL	trees	DecisionStump	crossEstratificad	88.00	0.62	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.86	1.00	0.93	0.72	0.85	0.00	0.00	0.00	0.52	0.12
100	MENSUAL	trees	J48	crossEstratificad	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
100	MENSUAL	trees	LMT	crossEstratificad	98.00	0.95	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.97	0.99	1.00	0.86	1.00	0.92	0.99	0.95		
100	MENSUAL	trees	RandomForest	crossEstratificad	100.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
100	MENSUAL	trees	RandomTree	crossEstratificad	92.67	0.82	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.96	0.94	0.95	0.99	1.00	0.67	0.78	0.72	0.98	0.85	

Figura 15. Resultados de la evaluación del aprendizaje automático con el método Stratified K-fold Cross Validation.

VI. Análisis de los resultados experimentales

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

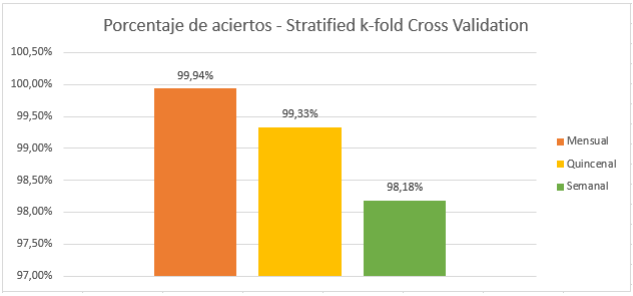

Figura 16. Promedios del porcentaje de aciertos utilizando Stratified k-fold Cross Validation.

En la Figura 17 se muestra un gráfico de barras con promedios del porcentaje de aciertos utilizando Percentage Split Validation como método de evaluación.

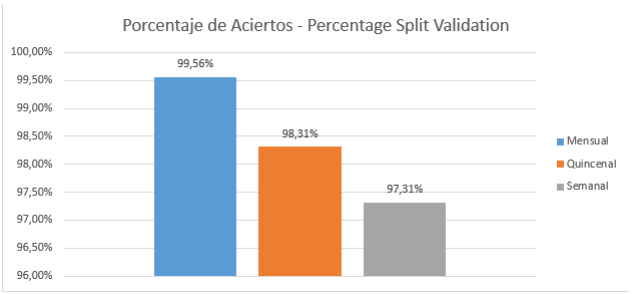

Figura 17. Promedios del porcentaje de aciertos utilizando Percentage Split Validation.

VII. Conclusiones

Aqui las conclusiones.

Referencias

[1] Interface classifier.

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

[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] W.H. Inmon. *Building the datawarehouse*. QED Press, 1992.

[6] W.H. Inmon. *Building the datawarehouse*. Willey, 1996.

[7] 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.

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

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

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

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

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

[13] Hugh James Watson. Recent developments in datawarehousing: A tutorial. 2006.

[14] Colin White Wayne Eckerson. Evaluating etl and data integration platforms. Technical report, TDWI Report Series, 2003.

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