

“Automatización en la toma de decisiones de compra de productos 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 por ejemplo, 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 se debe tomar es acerca del volumen a adquirir de un producto determinado, 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 la decisión del volumen de compra de cualquier producto en stock, un modelo basado en business intelligence y machine learning.

Index Terms—retail, business intelligence, machine learning.

I. INTRODUCCIÓN

Gestionar la información en las empresas es, hoy en día, una herramienta clave para poder sobrevivir en un mercado cambiante, dinámico y global. Aprender a competir con esta información es fundamental para la toma de decisiones, el crecimiento y la gestión de nuestra empresa. La disciplina denominada como Business Intelligence nos acerca a los sistemas de información que nos ayudan a la toma de decisiones en nuestra organización[3].

En las empresas retail o de ventas minoristas uno de los principales problemas con los que se enfrentan es el manejo eficiente del stock de manera a no tener los productos en exceso en los depósitos que incurran en sobrecostos o en el otro extremo la falta de dichos productos, lo cual conlleva a pérdidas de oportunidades de ventas por no disponer del producto que a su vez repercute directamente en las utilidades. El presente trabajo se enfoca en establecer un modelo de estimación de Cantidad de Compra Óptima de productos. Para alcanzar el objetivo se utilizan una plataforma de Business Intelligence en la cual se modelan los Indicadores Claves de Rendimiento (KPI) que sirven como datos de entrada en los distintos algoritmos de aprendizaje automático, donde se evalúa el desempeño a fin de determinar los más adecuados para la creación de un modelo que permita estimar la Cantidad de Compra Óptima.

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

II. BUSINESS INTELLIGENCE

II-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. Las tecnologías de business intelligence son capaces de manejar grandes volúmenes de datos que ayudan a identificar, desarrollar y crear nuevas estrategias de negocios.

El primero que acuñó el término fue Howard Dresner, quién cuando era consultor de Gartner Group lo utilizó para describir un conjunto de conceptos y métodos que mejoran la toma de decisiones, partiendo de la información disponible acerca de los hechos. Entonces, partiendo de la definición del glosario de términos de Gartner [5]:

“Business Intelligence es un proceso interactivo para explorar y analizar información estructurada sobre un área (normalmente almacenada en un datawarehouse), para descubrir tendencias o patrones, a partir de las cuales derivar ideas y extraer conclusiones. Las áreas incluyen clientes, proveedores, productos, servicios y competidores. El proceso de business intelligence incluye la comunicación de los descubrimientos y efectuar los cambios”.

Indica que business intelligence es un proceso que se prolonga en el tiempo, que no es sólo para un momento puntual de la gestión empresarial, en el que podremos ver tendencias, patrones, cambios, variables, etc. Al explorar iremos descubriendo nuevas relaciones que hasta el momento desconocíamos. Al analizar de lo nuevo que hemos descubierto, veremos relaciones entre variables, tendencias, patrones y cuál puede ser la evolución de los mismos.

La información estructurada está en tablas relacionadas, dichas tablas a la vez están en un datawarehouse o almacén de datos. Nos podemos centrar en áreas específicas del negocio y en objetivos concretos como por ejemplo: reducir costes, incrementar ventas, aumentar la participación en el mercado, cumplir los objetivos de ventas presupuestados. Finalmente, lo descubierto y analizado se debe comunicar a aquellas personas en la organización que realizarán los cambios apropiados para mejorar la competitividad

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'".

II-B. Objetivos

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.
- Apoyar de forma sostenible y continuada a las organizaciones para mejorar su competitividad, ante el entorno de negocios cambiante de forma que puedan adaptarse a él.
- Ante la cantidad de información que va creciendo, disponer de más tiempo en analizarla, en vez de gastar mucho tiempo en prepararla, organizarla y estructurarla.
- 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.

II-C. Componentes de BI

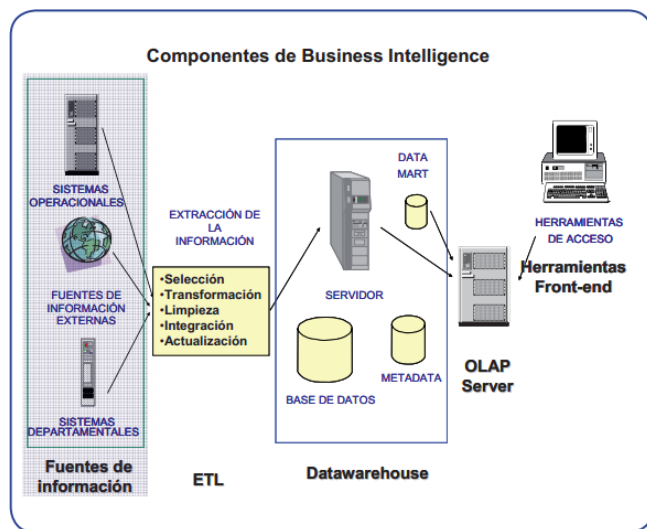


Figure 1. Fuentes de información

II-C1. Fuentes de información: Las fuentes de información a las que podemos acceder son:

¹Los 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".

- 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, en algunos casos comprada a terceros, como por ejemplo estudios de mercado (Nielsen en distribución de gran consumo, IMS de la industria farmacéutica). Las fuentes de información externas son fundamentales para enriquecer la información que tenemos de nuestros clientes. En algunos casos es interesante incorporar información referente, por ejemplo, a población, número de habitantes, etc. Podemos acceder a información de este tipo en la web del Instituto Nacional de Estadística (www.ine.es).

Existen muchos factores que contribuyen a la complejidad de cargar la información en un datawarehouse. Uno de los principales es el número de fuentes de información distintas de las que cargamos la información. Además, el número de fuentes de información varía de una organización a otra: en grandes corporaciones se habla de una media de 8 bases de datos, y en algunos casos puede llegar a 50.

Cada vez más la tecnología nos permite trabajar con información no estructurada, y se espera que este tipo de información sea cada vez más importante. Una encuesta² ha indicado que el 60% de los directores de Sistemas de Información y los de Tecnología consideran que la información semiestructurada es crítica para mejorar las operaciones y para la creación de nuevas oportunidades de negocio.

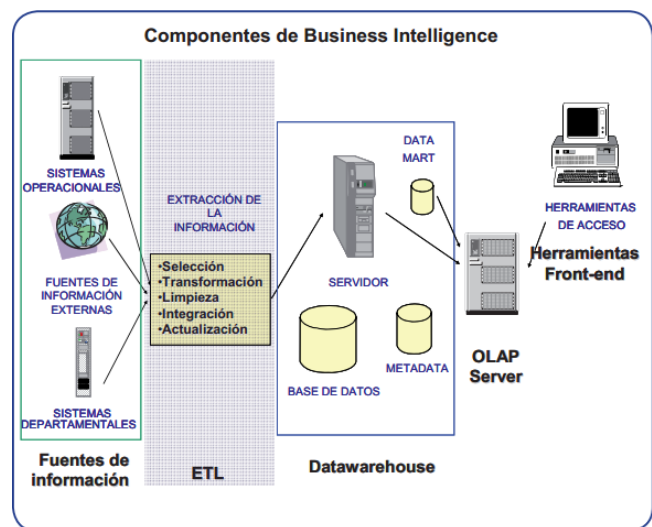


Figure 2. Proceso ETL

II-C2. ETL³ – 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

²Blumberg, R. y S. Atre "The Problem with Unstructured Data", DM Review, <http://dmreview.com/master.cfm?NavID=55&EdID=6287> (12 Septiembre 2003)

³ETL corresponde a las siglas del inglés Extract, Transform and Load (Extracción, transformación y carga)

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[15] 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. En este momento disponemos de los datos en bruto. El principal objetivo es extraer tan sólo aquellos datos de los sistemas transaccionales que son necesarios y prepararlos para el resto de los subprocesos de ETL. Para ello se deben determinar las mejores fuentes de información, las de mejor calidad. Con tal finalidad, deberemos analizar las fuentes disponibles y escoger aquellas que sean mejores.

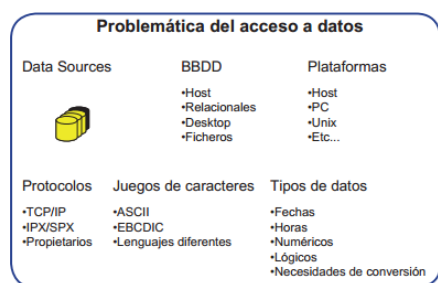


Figure 3. Problemas del acceso a datos

En la figura 3 mostramos los principales problemas con los que nos podemos encontrar al acceder a los datos para extraerlos: básicamente se refieren a que provienen de distintas fuentes, BBDD, plataformas tecnológicas, protocolos de comunicaciones, juegos de caracteres, y tipos de datos

Limpieza: Este proceso recupera los datos en bruto y comprueba su calidad, elimina los duplicados y, cuando es posible, corrige los valores erróneos y completa los valores vacíos, es decir se transforman los datos -siempre que sea posible- para reducir los errores de carga. En este momento disponemos de datos limpios y de alta calidad.

Los sistemas transaccionales contienen datos que no han sido depurados y que deben ser limpiados. Algunas causas que provocan que los datos estén “sucios” son:

- Valores por defecto.
- Ausencia de valor.
- Campos que tienen distintas utilidades.
- Valores contradictorios.
- Uso inapropiado de los campos.
- Reutilización de claves primarias.
- Selección del primer valor de una lista.
- Problemas de carga de antiguos sistemas o de integración entre sistemas.

La limpieza de datos se divide en distintas etapas:

- **Depurar los valores (parsing):** localiza e identifica los elementos individuales de información en las fuentes

de datos. Por ejemplo: separa el nombre completo en: nombre, primer apellido, segundo apellido; o la dirección en: calle, número, etc.

- **Corregir (correcting):** corrige los valores individuales de los atributos usando algoritmos de corrección y fuentes de datos externas. Por ejemplo: comprueba la dirección y su código postal correspondiente.
- **Estandarizar (standardizing):** aplica rutinas de conversión para transformar valores en formatos definidos y consistentes. Por ejemplo: trato de Sra. o Sr. cambiar a sus correspondientes nombres completos.
- **Relacionar (matching):** busca y relaciona los valores de registros, corrigiéndolos y estandarizándolos para eliminar duplicados. Por ejemplo: identificando nombres y direcciones similares.

Transformación: Este proceso recupera los datos limpios y de alta calidad y los estructura y resume en los distintos modelos de análisis. El resultado de este proceso es la obtención de datos limpios, consistentes, resumidos y útiles.

La transformación incluye:

- Cambios de formato.
- Sustitución de códigos.
- Valores derivados y agregados.

Los agregados como las sumas de las ventas normalmente se precalculan y se almacenan para conseguir mayores rendimientos. En este proceso también ajustamos el nivel de granularidad o detalle, por ejemplo: podemos tener detalles a nivel de líneas de factura en los datos extraídos, pero en el datawarehouse lo que almacenamos son las ventas semanales o mensuales. La diferencia del nivel de detalle en el análisis es lo que denominamos granularidad.

Integración: Este proceso valida que los datos que cargamos en el datawarehouse son 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

II-C3. 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[9].

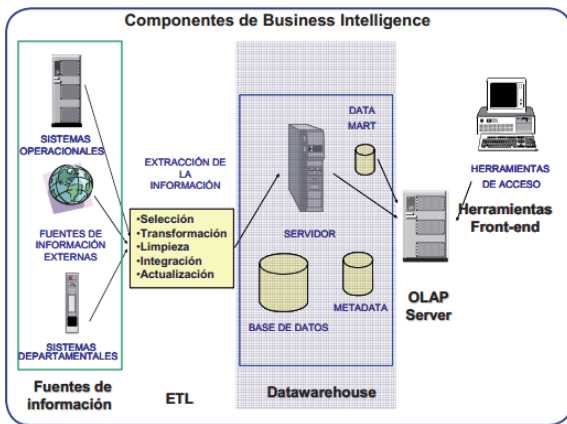


Figure 4. Datawarehouse

La aparición de los datawarehouses son la respuesta a las necesidades de los usuarios que necesitan información consistente, integrada, histórica y preparada para ser analizada y poder tomar decisiones. Si el datawarehouse está construido adecuadamente proporciona un entorno de información que nos permitirá encontrar nuevo conocimiento y generar valor.

El profesor Hugh J. Watson [14] 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[7] 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. Por ejemplo: entender los hábitos de compra de nuestros clientes, analizar la calidad de nuestros productos, analizar la productividad de una línea de fabricación.
- Integrado: la información debe ser transformada en medidas comunes, códigos comunes y formatos comunes para ser útil. Por ejemplo: la moneda en que están expresadas los importes es común.
- Indexado en el tiempo: se mantiene la información histórica. Ello nos permite por ejemplo: analizar la evolución de las ventas en los periodos que queramos.
- No volátil: los usuarios no la mantienen como lo harían en los entornos transaccionales. No se ve actualizado continuamente, sino periódicamente de forma preestablecida. La información se almacena para la toma de decisiones

Ralph Kimbal[10] 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.

Se deben tener en cuenta que existen otros elementos en el contexto de un datawarehouse :

- **Datawarehousing:** es el proceso de extraer y filtrar datos de las operaciones procedentes de los distintos sistemas de información operacionales y sistemas externos, para transformarlos, integrarlos y almacenarlos en un almacén de datos con el fin de acceder a ellos para dar soporte en el proceso de toma de decisiones.
- **Data Mart:** es un subconjunto de los datos del datawarehouse cuyo objetivo es responder a un determinado análisis.
- **Operational Data Store (ODS):** es un tipo de almacén de datos que proporciona sólo los últimos valores de los datos y no su historial.
- **Staging Area:** es el sistema que permanece entre las fuentes de datos y el datawarehouse con el objetivo de:
 - Facilitar la extracción de datos desde fuentes de origen con una heterogeneidad y complejidad grande.
 - Mejorar la calidad de los datos.
 - Ser usado como caché de datos operacionales con el que posteriormente se realiza el proceso de datawarehousing.
 - Uso de la misma para acceder en detalle a información no contenida en el datawarehouse.
- **Procesos ETL:** tecnología de integración de datos basada en la consolidación de datos que se usa tradicionalmente para alimentar el datawarehouse, data mart, staging area y ODS.
- **Metadatos:** datos estructurados y codificados que describen características de instancias; aportan informaciones para ayudar a identificar, descubrir, valorar y administrar las instancias descritas.

Elementos de una datawarehouse: La estructura relacional de una base de datos operacional sigue las formas normales en su diseño. Un datawarehouse no debe seguir ese patrón de diseño. La idea principal es que la información sea presentada desnormalizada para optimizar las consultas. Para ello debemos identificar, en la organización, los procesos de negocio, las vistas para el proceso de negocio y las medidas cuantificables asociadas a los mismos. De esta manera hablaremos de:

- **Tablas de hecho:** es la representación en el datawarehouse de los procesos de negocio de la organización. A nivel de diseño es una tabla que permite guardar dos tipos de atributos diferenciados:
 - Medidas del proceso de trabajo que se pretende modelizar
 - Claves foráneas hacia registros de una tabla de dimensión
- **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. A nivel de diseño, consiste en una tabla de hechos en el centro para el hecho objeto de análisis y una o varias tablas de dimensión por cada punto de vista del análisis que participan en la descripción de ese hecho.

- **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. Por esta razón la tabla de hechos deja de ser la única tabla del esquema que se relaciona con otras tablas.

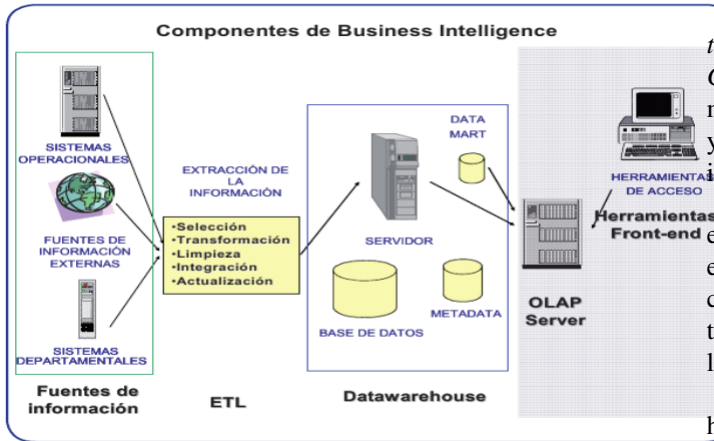


Figure 5. Herramientas de Bussines Intelligence

II-C4. 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[8] necesitan analizar información a distintos niveles de agregación y sobre múltiples dimensiones. Por ejemplo: ventas de productos por clientes o tipo de cliente, por zona de venta y por fecha. OLAP provee de estas funcionalidades y algunas más.

Las principales herramientas[4] 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.
- **Herramientas de planificación, modelización y consolidación:** Permite a los analistas y a los usuarios finales crear planes de negocio y simulaciones con la información de Business Intelligence. Pueden ser para elaborar la planificación, los presupuestos, las previsiones. Estas

herramientas proveen a los dashboards y los scorecards con los objetivos y los umbrales de las métricas.

- **Herramientas datamining:** Permiten a estadísticos o analistas de negocio crear modelos estadísticos de las actividades de los negocios. Datamining es el proceso para descubrir e interpretar patrones desconocidos en la información mediante los cuales resolver problemas de negocio. Los usos más habituales del datamining son: segmentación, venta cruzada, sendas de consumo, clasificación, previsiones, optimizaciones, etc.

II-C5. KPI: Son la iniciales de *Key Performance Indicators* que traducido al español vendrían a ser lo *Indicadores Claves de Desempeño*. 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 indicadores claves, porque puede darse el problema de “la parálisis por el análisis” y una de las características de nuestro entorno competitivo actual es que tenemos que tomar decisiones de forma rápida y antes de que lo hagan los demás.

Otra característica que define a los KPI es que cada empresa ha de definir cuáles son aquellos indicadores que quiere tener siempre presentes para manejar su rumbo.

El tercer elemento definitorio de los KPI es que no tienen que estar referidos exclusivamente a resultados de tipo financiero.

III. 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*” [12]. 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*”. Donde se nos indica que el aprendizaje en las máquinas deberá ser parecido al aprendizaje en los humanos, por ejemplo cuando una criatura comienza a hablar a través de la experiencia de pronunciar las palabras y de su interacción con otras personas, entonces sucede que su capacidad de hablar se va perfeccionando o mejorando.

III-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*”[11].

La dificultad radica en que debemos construir modelos que nos acerquen a una buena predicción sobre datos aún no conocidos o imprevistos.

Machine Learning

- Data comes as...
 - A set of examples $\{(\mathbf{x}_i, y_i) | 0 \leq i < n_{\text{samples}}\}$, with
 - Feature vector $\mathbf{x} \in \mathbb{R}^{n_{\text{features}}}$, 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 6. Definición presentada por Peter Prettenhofer y Gille Louppe.

III-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 para aprender desde los mismos.

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

III-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.:

III-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.

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

III-C2. Clasificación: Un problema de aprendizaje supervisado donde la respuesta a aprender es un valor de un conjunto finito de posibles valores discretos. *Classification learning is sometimes called supervised, because, in a sense, the scheme operates under supervision by being provided with the actual outcome for each of the training examples:*

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

III-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.:

III-D. Componentes esenciales

III-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.:

III-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 [16]. :

III-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.:

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

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

III-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.:

III-E. El problema de la clasificación

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

III-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 $y \in \{\pm 1\}$ [13].:

III-E2. Clasificación multiclase (multiclass classification): Es la extensión lógica de la clasificación binaria. La principal diferencia es que ahora $y \in \{1, \dots, N\}$ puede asumir un rango de valores diferentes [13].:

III-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 [6]. Los algoritmos de clasificación de Weka que se utilizarán es el siguiente [1]:

III-F1. BayesNet: Bayes Network learning using various search algorithms and quality measures. Base class for a Bayes Network classifier. Provides datastructures (network structure, conditional probability distributions, etc.) and facilities common to Bayes Network learning algorithms like K2 and B.:

III-F2. NaiveBayes: Class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an UpdatableClassifier (which in typical usage are initialized with zero training instances).:

III-F3. NaiveBayesUpdateable: Class for a Naive Bayes classifier using estimator classes. This is the updateable version of NaiveBayes. This classifier will use a default precision of 0.1 for numeric attributes when buildClassifier is called with zero training instances.:

III-F4. Logistic: Class for building and using a multinomial logistic regression model with a ridge estimator. If there are k classes for n instances with m attributes, the parameter matrix B to be calculated will be an $m \times (k-1)$ matrix.:

III-F5. MultilayerPerceptron: A Classifier that uses back-propagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric in which case the the output nodes become unthresholded linear units).:

III-F6. SimpleLogistic: Classifier for building linear logistic regression models. LogitBoost with simple regression functions as base learners is used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection.:

III-F7. SMO: Implements John Platt's sequential minimal optimization algorithm for training a support vector classifier. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. (In that case the coefficients in the output are based on the normalized data, not the original data --- this is important for interpreting the classifier). Multi-class problems are solved using pairwise classification (aka 1-vs-1). To obtain proper probability estimates, use the option that fits calibration models to the outputs of the support vector machine. In the multi-class case, the predicted probabilities are coupled using Hastie and Tibshirani's pairwise coupling method. :

III-F8. OneR: Class for building and using a 1R classifier; in other words, uses the minimum-error attribute for prediction, discretizing numeric attributes.:

III-F9. DecisionTable: Class for building and using a simple decision table majority classifier.:

III-F10. JRip: This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP. :

III-F11. PART: Class for generating a PART decision list. Uses separate-and-conquer. Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule.:

III-F12. ZeroR: Class for building and using a 0-R classifier. Predicts the mean (for a numeric class) or the mode (for a nominal class).:

III-F13. DecisionStump: Class for building and using a decision stump. Usually used in conjunction with a boosting algorithm. Does regression (based on mean-squared error) or classification (based on entropy). Missing is treated as a separate value.:

III-F14. J48: Class for generating a pruned or unpruned C4.5 decision tree.:

III-F15. LMT: Classifier for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values.:

III-F16. RandomForest: Class for constructing a forest of random trees.:

III-F17. RandomTree: Class for constructing a tree that considers K randomly chosen attributes at each node. Performs no pruning. Also has an option to allow estimation of class probabilities (or target mean in the regression case) based on a hold-out set (backfitting). :

III-F18. REPTree: Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).:

III-G. Evaluación de lo aprendido

La evaluación es la clave para lograr avances reales en el aprendizaje automático.

III-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.:

III-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 apoya esto. Aunque estos argumentos no son en absoluto concluyentes, y el debate continúa enfurecido en los círculos de aprendizaje automático y de minería de datos sobre cuál es el mejor esquema de evaluación, la validación cruzada diez veces se ha convertido en el método estándar en términos

prácticos. Las pruebas también han demostrado que el uso de la estratificación mejora ligeramente los resultados. Por lo tanto, la técnica de evaluación estándar en situaciones en las que sólo se dispone de datos limitados es la validación cruzada diez veces estratificada. La estratificación reduce la variación, ciertamente no la elimina completamente. Cuando se busca una estimación exacta del error, es un procedimiento estándar repetir el proceso de validación cruzada 10 veces, es decir, diez veces la validación cruzada diez veces, y el promedio de los resultados. Esto implica invocar el algoritmo de aprendizaje 100 veces en conjuntos de datos que son todas las nueve décimas del tamaño del original. Obtener una buena medida de rendimiento es una empresa de computación intensiva.:

III-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.:

III-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. Ya sabemos las clasificaciones de cada instancia en el conjunto de entrenamiento, que después de todo es por qué podemos usarlo para el entrenamiento. La tasa de error en el conjunto de entrenamiento no es probable que sea un buen indicador de rendimiento futuro debido a que el clasificador se ha aprendido de los mismos datos de entrenamiento, cualquier estimación de rendimiento basada en esos datos será optimista, incluso excesivamente optimista.

La tasa de error en los datos de entrenamiento se llama error de resustitución porque se calcula resustituyendo las instancias de entrenamiento en un clasificador que se construyó a partir de ellas. 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.

Generalmente, cuanto mayor es la muestra de entrenamiento, mejor es el clasificador, aunque los retornos comienzan a disminuir una vez que se sobrepasa un cierto volumen de datos de entrenamiento. Y cuanto mayor es la muestra de prueba, más precisa es la estimación del error. La precisión de la estimación del error puede ser cuantificada estadísticamente.

El verdadero problema ocurre cuando no hay una gran cantidad de datos disponibles.

En términos prácticos, es común tener un tercio de los datos para la prueba y utilizar los dos tercios restantes para el entrenamiento.

En general, no se puede saber si una muestra es representativa o no. Pero hay una comprobación simple que puede ser útil: Cada clase en el conjunto de datos completo debe estar representada en la proporción correcta en los conjuntos de entrenamiento y prueba. Deberían estar representados en la proporción adecuada en los conjuntos de entrenamiento y prueba. Si, por mala suerte, todos los ejemplos con una cierta clase se omitieran en el conjunto de entrenamiento, difícilmente podría esperar que una clase aprendiera de esos datos para funcionar bien en los ejemplos de esa clase.

Debe asegurarse de que el muestreo aleatorio se realiza de una manera que garantiza que cada clase esté representada correctamente en los conjuntos de entrenamiento y prueba. Este procedimiento se denomina estratificación, y hablamos de la retención estratificada (stratified holdout). Aunque en general vale la pena hacerlo, la estratificación sólo proporciona una salvaguardia primitiva contra la representación desigual en los conjuntos de entrenamiento y pruebas.

Una manera más general de mitigar cualquier sesgo causado por la muestra particular escogida para retener, es repetir todo el proceso, entrenamiento y prueba, varias veces con diferentes muestras aleatorias. En cada iteración, una cierta proporción, digamos dos tercios, de los datos se selecciona al azar para el entrenamiento, posiblemente con estratificación, y el resto se utiliza para la prueba. Las tasas de error en las diferentes iteraciones se promedian para obtener una tasa de error global. Este es el método de retención repetida de la estimación de la tasa de error.

III-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. :

III-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).:

III-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. Y la detección de mancha de aceite: El costo de no detectar una mancha real que amenaza el medio

ambiente es mucho mayor que el costo de una falsa alarma. El costo de identificar malos problemas con una máquina que resulta estar libre de fallas es menor que el costo de pasar por alto los problemas con uno que está a punto de fallar. Los verdaderos positivos (TP) y negativos verdaderos (TN) son clasificaciones correctas. Un falso positivo (PF) es cuando el resultado se predice incorrectamente como sí (o positivo) cuando es realmente no (negativo). Un falso negativo (FN) es cuando el resultado se predice incorrectamente como negativo cuando es realmente positivo. La tasa positiva verdadera es TP dividida por el número total de positivos, que es $TP + FN$; la tasa de falsos positivos es FP dividida por el número total de negativos, que es $FP + TN$. 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 resumen, la estadística Kappa se utiliza para medir el acuerdo entre categorizaciones predichas y observadas de un conjunto de datos, mientras se corrige un acuerdo que se produce por casualidad. Sin embargo, al igual que la tasa de éxito simple, no toma en cuenta los costos. En WEKA el método `kappa()` devuelve el valor de la estadística kappa si la clase es nominal.:

III-H4. Curvas ROC (ROC Curves): Están estrechamente relacionados con una técnica gráfica para evaluar los esquemas de minería de datos conocidos como curvas ROC, que se utilizan en la misma situación, donde el alumno está tratando de seleccionar muestras de instancias de prueba que tienen una alta proporción de positivos. El acrónimo significa la característica de funcionamiento del receptor (Receiver Operating Characteristic), un término usado en la detección de señal para caracterizar la compensación entre la tasa de acierto y la tasa de falsa alarma sobre un canal ruidoso. Representan la tasa de positivos verdaderos en el eje vertical con respecto a la tasa de verdaderos negativos en el eje horizontal. El primero es el número de positivos incluidos en la muestra, expresado como un porcentaje del número total de positivos ($TP\ Rate = 100 \times TP / (TP + FN)$); el último es el número de negativos incluidos en la muestra, expresado como porcentaje del número total de negativos ($FP\ Rate = 100 \times FP / (FP + TN)$). El eje vertical es el mismo que el gráfico de elevación excepto que se expresa como un porcentaje. El eje horizontal es ligeramente diferente: es el número de negativos en lugar del tamaño de la muestra. Each point corresponds to drawing a line at a certain position on the ranked list, counting the yes's and no's above it, and plotting them vertically and horizontally, respectively. Esto es sólo una forma de usar la validación cruzada para generar curvas

ROC. Un enfoque más simple es recopilar las probabilidades predichas para todos los conjuntos de prueba diferentes (de los cuales hay 10 en un 10-fold validación cruzada), junto con las etiquetas de clase verdadera de las instancias correspondientes, y generar una única lista rankeada basada en estos datos. Si el esquema de aprendizaje no permite ordenar las instancias, primero puede hacer que sea sensible al costo como se describió anteriormente. Para cada fold de una 10-fold validación cruzada, ponderar las instancias para una selección de diferentes ratios de coste, entrenar el esquema en cada conjunto ponderado, contar los verdaderos positivos y falsos positivos en el conjunto de pruebas y trazar el punto resultante en los ejes ROC. En WEKA el método `areaUnderROC(int classIndex)` devuelve el área bajo ROC para las predicciones que se han recogido en el método `evaluateClassifier(Classifier, Instances)`.:

III-H5. 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. ¿Cuál 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.:

III-H6. 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: $2 * recall * precision / recall + precision$.:

III-H7. Curvas Sensibilidad-Precisión (Recall-Precision Curves): Los expertos en recuperación de información usan curvas recall-precision que trazan una contra la otra, para diferentes números de documentos recuperados de la misma manera que las curvas ROC y los gráficos de elevación, excepto que debido a que los ejes son diferentes, las curvas son de forma hiperbólica y el punto de operación deseado está hacia la parte superior derecha. Table 5.7 summarizes the three different ways introduced for evaluating the same basic tradeoff; TP, FP, TN, and FN are the numbers of true positives, false positives, true negatives, and false negatives, respec-

tively. Diferentes técnicas dan diferentes compensaciones, y se pueden trazar como líneas diferentes en cualquiera de estas gráficas. En WEKA el método `areaUnderPRC(int classIndex)` devuelve el área bajo la curva sensibilidad-precisión para las predicciones que se han recogido en el método `evaluateClassifier (Classifier, Instances):`.

IV. MODELADO MEDIANTE LA UTILIZACIÓN DE KPI

Esta sección se enfoca y analiza uno de los principales problemas con los que se enfrentan las empresas retail⁴, la cual trata acerca de la reposición de stock⁵, es decir, determinar la cantidad de productos que se deben comprar para satisfacer la demanda de los clientes por un determinado periodo de tiempo. Utilizando los conceptos y herramientas de bussines intelligence mencionados en el capítulo 2, se define y se diseña el datawarehouse donde son almacenados los datos históricos de las compras y ventas de la empresa retail que servirán para el análisis, posteriormente se definen los indicadores claves de rendimiento como métricas que son utilizados como datos de entrada en las herramientas de aprendizaje automático para crear un modelo para la predicción de la cantidad óptima que debe comprarse para la reposición del stock y así cubrir la demanda de productos de los clientes.

IV-A. Problema de la reposición de stock

LLamamos stocks o existencias de una empresa al conjunto de materiales y artículos que se almacenan, tanto aquellos que son necesarios para el proceso productivo como los destinados a la venta.

Es importante mantener un stock adecuado de los productos para garantizar la demanda de los clientes pero cuidando que su almacenamiento resulte rentable. Tener un stock insuficiente puede acarrear consigo un serie de inconvenientes como: pérdida de ventas, pérdida de imagen y pérdida de la confianza de los clientes. Por otra parte, se debe tener cuidado con tener un stock excesivo, ya que podría incurrir en mayores costes de almacenamiento y repercutir en el precio de venta final.

De esto surgen las siguientes interrogantes::

- ¿Cuándo debemos realizar un pedido?
- ¿Qué cantidad debemos hacer el pedido?

En nuestro medio en numerosas ocasiones el determinar cuando se debe realizar un pedido y la cantidad que se precisa comprar se realiza en forma empírica mediante una revisión de las cantidad vendida en el último periodo y un conteo rápido de la cantidad existente actualmente.

IV-B. Base de datos

Para el presente trabajo, contamos con una base de datos auténtica con los registros de productos, proveedores, movimientos de compras, ventas y registro de stock realizados por una empresa retail. Los datos transaccionales contenidos

en la base de datos corresponden a movimientos realizados entre el año 2013 al 2016, el cual será el punto de partida para diseñar el datawarehouse.

La base de datos es una base de datos relacional gestionada y administrada por Oracle 11g Enterprise.

Tabla de Productos: La empresa retail cuenta con 13.200 artículos registrados disponibles para la venta.

IDENTIFICADOR	ID_EMPRESA	ID_UNIDAD	DESCRIPCION	ID_CATEGORIA	ACTIVO	USO_INTERNO	ITEM_INVENTARIO
790	1	1	AGUA SCHWEPES TONICA X 500 ML	...	1	N	S
4300	1	1	ESCOBILLA DURA	...	1	N	S
4311	1	1	DETERGENTE ACTIVO 100 DE 750	...	1	N	S
4314	1	1	ESCOBILLO GRANDE	...	1	N	S
4342	1	1	PELUCHE BARBERE	...	1	N	S
4343	1	1	PELUCHE OSTO ROSADO MUSICAL	...	1	N	S
4344	1	1	PELUCHE OSTO C-CORAZON (CANTA)	...	1	N	S
4353	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4354	1	1	REVISAL CERA LUST. POLISH P/AUTOMA	...	1	N	S
4355	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4356	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4357	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4358	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4359	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4360	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4361	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4362	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4363	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4364	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4365	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4366	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4367	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4368	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4369	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4370	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4371	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4372	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4373	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4374	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4375	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4376	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4377	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4378	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4379	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4380	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4381	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4382	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4383	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4384	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4385	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4386	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4387	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4388	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4389	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4390	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4391	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4392	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4393	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4394	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4395	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4396	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4397	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4398	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4399	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S
4400	1	1	REVISAL CERA CROMADA P/AUTOMA	...	1	N	S

Figure 7. Productos.

Tabla Proveedores: La empresa contiene una base de 1.623 proveedores registrados.

IDENTIFICADOR	ID_EMPRESA	ID_UNIDAD	DENOMINACION	ACTIVO	USR_CRE	REC_CRE	USR_MOD	REC_MOD
212	1	1	CARF FICHA S.R.L.	...	CAROL	05/11/2009 10:47:41	CAROL	05/11/2009 10:47:41
233	1	1	ALMACEN 30	...	CAROL	11/03/2009 14:19:51	CAROL	11/03/2009 14:19:51
231	1	1	DEMA S.R.L.	...	CAROL	11/03/2009 14:17:47	CAROL	11/03/2009 14:17:47
235	1	1	DECO CHIRRASQUERA	...	CAROL	11/03/2009 14:20:38	CAROL	11/03/2009 14:20:38
237	1	1	EL SOL	...	CAROL	11/03/2009 14:21:34	CAROL	11/03/2009 14:21:34
327	1	1	MAX SUPPLIER S.R.L.	...	CAROL	30/04/2009 09:14:54	CAROL	30/04/2009 09:14:54
331	1	1	LETICIA SHOP	...	CAROL	12/10/2009 07:47:27	CAROL	12/10/2009 07:47:27
354	1	1	HOTEL CONSAIR S.A.	...	CAROL	12/10/2009 07:50:02	CAROL	12/10/2009 07:50:02
800	1	1	ARTESANIA	...	APPUSS	05/01/2010 13:29:51	APPUSS	05/01/2010 13:29:51
811	1	1	DISTRIBUIDORA DUARTE	...	APPUSS	05/01/2010 13:37:45	APPUSS	05/01/2010 13:37:45
814	1	1	OIRO VERDE S.R.L.	...	APPUSS	05/01/2010 13:45:11	APPUSS	05/01/2010 13:45:11
817	1	1	SOL S.R.L.	...	APPUSS	05/01/2010 14:00:18	APPUSS	05/01/2010 14:00:18
820	1	1	DISTRIBUIDORA MAJIA S.R.L.	...	APPUSS	05/01/2010 14:03:12	APPUSS	05/01/2010 14:03:12
823	1	1	NESTLE PARAGUAY S.A.	...	APPUSS	05/01/2010 14:05:34	APPUSS	05/01/2010 14:05:34
826	1	1	DISTRIBUIDORA REPRESENTACIONES	...	APPUSS	05/01/2010 14:23:39	APPUSS	05/01/2010 14:23:39
831	1	1	JUSMEN	...	APPUSS	05/01/2010 15:30:08	APPUSS	05/01/2010 15:30:08
834	1	1	NA CREST S.A.	...	APPUSS	05/01/2010 15:36:16	APPUSS	05/01/2010 15:36:16
837	1	1	VICTOR RIOS S.A.	...	BLASIDA	05/01/2010 15:46:44	BLASIDA	05/01/2010 15:46:44
850	1	1	FERRETERIA ELECTRICIDAD "SAN JUAN"	...	BLASIDA	07/01/2010 07:48:25	BLASIDA	07/01/2010 07:48:25
863	1	1	TELECENTRO S.A.	...	ANTONIA	07/01/2010 08:31:39	ANTONIA	07/01/2010 08:31:39
866	1	1	DISTRIBUIDORA CENTRAL S.A.	...	TAMARA	07/01/2010 09:03:03	TAMARA	07/01/2010 09:03:03
869	1	1	HERP S.A.	...	ANTONIA	07/01/2010 09:39:30	ANTONIA	07/01/2010 09:39:30
872	1	1	DISTRIBUIDORA KAMEY	...	TAMARA	07/01/2010 10:10:00	TAMARA	07/01/2010 10:10:00
875	1	1	SUPER BOTTINO HINDS S.A.	...	TAMARA	07/01/2010 10:14:08	TAMARA	07/01/2010 10:14:08
886	1	1	LA ESPERANZA	...	TAMARA	07/01/2010 10:47:15	TAMARA	07/01/2010 10:47:15
890	1	1	NOVEK S.A.	...	TAMARA	07/01/2010 10:50:42	TAMARA	07/01/2010 10:50:42
892	1	1	TARUMA S.A.C.	...	TAMARA	07/01/2010 10:53:58	TAMARA	07/01/2010 10:53:58
895	1	1	DISTRIBUIDORA "LA FAMILIA"	...	TAMARA	07/01/2010 11:02:21	TAMARA	07/01/2010 11:02:21
896	1	1	LOS TRES REYES S.A.	...	BLASIDA	07/01/2010 14:00:29	BLASIDA	07/01/2010 14:00:29
900	1	1	ALEM S.R.L.	...	TAMARA	07/01/2010 16:14:20	TAMARA	07/01/2010 16:14:20
923	1	1	PARADERIA ALCOR RIAN	...	TAMARA	07/01/2010 16:31:22	TAMARA	07/01/2010 16:31:22
941	1	1	A.I.S.A. CALIDAD ANTE TODO	...	ANTONIA	04/01/2010 11:38:51	ANTONIA	04/01/2010 11:38:51
944	1	1	ALPATEX S.A.C.I	...	ANTONIA	04/01/2010 15:18:25	ANTONIA	04/01/2010 15:18:25
947	1	1	ALCOS S.R.L.	...	ANTONIA	04/01/2010 16:03:24	ANTONIA	04/01/2010 16:03:24
950	1	1	MULTIENAVES S.R.L.	...	BLASIDA	07/01/2010 15:09:51	BLASIDA	07/01/2010 15:09:51
951	1	1	DISTRIBUIDORA C.L.R.	...	ANTONIA	07/01/2010 16:11:42	ANTONIA	07/01/2010 16:11:42
952	1	1	COTILLON FLORIDA	...	TAMARA	07/01/2010 16:32:57	TAMARA	07/01/2010 16:32:57
959	1	1	POMPI	...	TAMARA	07/01/2010 16:37:31	TAMARA	07/01/2010 16:37:31
964	1	1	COMERCIAL SAN LUIS DEL SUR S.A.	...	TAMARA	07/01/2010 18:05:06	TAMARA	07/01/2010 18:05:06
987	1	1	ELVA COMERCIAL	...	BLASIDA	08/01/2010 09:07:38	BLASIDA	08/01/2010 09:07:38

Figure 8. 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, la tabla contiene 301.316 registros de ventas durante un periodo de 4 años, desde el 07/11/2013 hasta el 04/10/2016. Contiene datos de la fecha, numero de factura, cliente, montos totales entre otros datos.

⁴Una 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.

⁵Stock o existencia es la cantidad de un determinado producto almacenado o disponible para la venta.

select * from VTA_COMPROBANTES

ID	IDENTIFICADOR	ID_TRANACCION	NUMERO	FECHA	ESTADO	ID_SITIO	ID_MONEDA	MONTO_TOTAL	ID_TIPO_COMPROB	USUARIO
1	130214	4 005-001-000177	4076	20/11/2013 08:54:43	0	1243	1	4500	21	VANANNA
2	130215	4 005-001-000178	4076	20/11/2013 08:55:36	0	1243	1	23300	21	VANANNA
3	130216	4 005-001-000179	4076	20/11/2013 08:56:46	0	1243	1	165000	21	VANANNA
4	130217	4 005-001-000180	4076	20/11/2013 09:05:47	0	1243	1	7300	21	VANANNA
5	130218	4 005-001-000181	4076	20/11/2013 09:07:52	0	1243	1	22000	21	VANANNA
6	130219	4 005-001-000182	4076	20/11/2013 09:14:39	0	1243	1	4000	21	VANANNA
7	130266	4 005-001-000529	4076	20/11/2013 06:12:30	0	1243	1	2500	21	ROBERTO
8	130267	4 005-001-000530	4076	20/11/2013 06:20:00	0	1243	1	15500	21	ROBERTO
9	130268	4 005-001-000531	4076	20/11/2013 06:30:15	0	1243	1	18000	21	ROBERTO
10	130269	4 005-001-000532	4076	20/11/2013 06:34:38	0	1243	1	12500	21	ROBERTO
11	130270	4 005-001-000533	4076	20/11/2013 06:36:48	0	1243	1	2500	21	ROBERTO
12	130271	4 005-001-000534	4076	20/11/2013 07:18:27	0	1243	1	800	21	VANANNA
13	130272	4 005-001-000535	4076	20/11/2013 07:19:36	0	1243	1	13500	21	VANANNA
14	130273	4 005-001-000536	4076	20/11/2013 07:47:14	0	1243	1	7500	21	VANANNA
15	130274	4 005-001-000537	4076	20/11/2013 07:51:26	0	1243	1	12000	21	VANANNA
16	130275	4 005-001-000538	4076	20/11/2013 07:52:48	0	1243	1	28500	21	VANANNA
17	130276	4 005-001-000539	4076	20/11/2013 07:55:06	0	1243	1	4000	21	VANANNA
18	130277	4 005-001-000540	4076	20/11/2013 07:56:23	0	1243	1	8200	21	VANANNA
19	130278	4 005-001-000541	4076	20/11/2013 08:01:08	0	1243	1	4500	21	VANANNA
20	130279	4 005-001-000542	4076	20/11/2013 08:07:59	0	1243	1	3500	21	VANANNA
21	130280	4 005-001-000543	4076	20/11/2013 08:08:16	0	1243	1	1500	21	VANANNA
22	130281	4 005-001-000544	4076	20/11/2013 08:11:58	0	1243	1	7000	21	VANANNA
23	130289	4 005-001-000552	4076	20/11/2013 11:30:38	0	1243	1	54000	21	JAVER
24	123273	4 005-001-000075	4076	08/11/2013 19:33:25	0	1243	1	20000	21	JAVER
25	123274	4 005-001-000076	4076	08/11/2013 19:35:42	0	1243	1	13500	21	JAVER
26	123275	4 005-001-000077	4076	08/11/2013 19:38:18	0	1243	1	10500	21	JAVER
27	123276	4 005-001-000078	4076	08/11/2013 19:37:10	0	1243	1	5000	21	JAVER
28	123277	4 005-001-000079	4076	08/11/2013 19:37:10	0	1243	1	5000	21	JAVER
29	123278	4 005-001-000080	4076	08/11/2013 19:41:54	0	1243	1	306500	21	JAVER
30	123279	4 005-001-000081	4076	08/11/2013 19:43:38	0	1243	1	10000	21	JAVER
31	126073	4 005-001-0001138	4076	10/11/2013 06:52:52	0	1243	1	205000	21	ROBERTO
32	126074	4 005-001-0001139	4076	10/11/2013 06:29:37	0	1243	1	63000	21	ROBERTO
33	126075	4 005-001-0001160	4076	10/11/2013 06:33:26	0	1243	1	13000	21	ROBERTO
34	126076	4 005-001-0001161	4076	10/11/2013 06:34:01	0	1243	1	105000	21	ROBERTO
35	126077	4 005-001-0001162	4076	10/11/2013 06:42:54	0	1243	1	23000	21	ROBERTO
36	126078	4 005-001-0001163	4076	10/11/2013 06:51:15	0	1243	1	5000	21	ROBERTO
37	126079	4 005-001-0001164	4076	10/11/2013 07:04:40	0	1243	1	17500	21	ROBERTO
38	126080	4 005-001-0001165	4076	10/11/2013 07:11:49	0	1243	1	45000	21	ROBERTO
39	126081	4 005-001-0001166	4076	10/11/2013 07:15:48	0	1243	1	23000	21	ROBERTO
40	126082	4 005-001-0001167	4076	10/11/2013 07:16:50	0	1243	1	9000	21	ROBERTO

Figure 9. Tabla de Ventas Cabecera.

Tabla de Ventas Detalle: La tabla de ventas detalle contiene los registros de los productos que fueron comercializados, cada detalle está relacionado a un registro cabecera. La tabla contiene 981.402 registros de productos que fueron vendidos en el mismo periodo indicado en el punto anterior. Contiene datos de la fecha, el producto vendido, precio de costo, precio de venta, cantidad, porcentaje de impuesto y otros datos.

select * from VTA_ITEM_COMPROB

ID	IDENTIFICADOR	ID_COMPROBANTE	ID_EMPRESA	NUMERO_ITEM	IMPORTE_ITEM	IMPORTE_ITEM	FECH_CNE	FECH_CNE	USUR_MCD	FECH_MCD	ID_PRODUCTO
1	405338	127919	1	4076	5	FR	3500	ROBERTO	14/11/2013 07:01:40	*	2388
2	405337	127919	1	4076	5	FR	3500	ROBERTO	14/11/2013 07:01:42	*	284
3	405338	127919	1	4076	6	FR	7500	ROBERTO	14/11/2013 07:02:02	*	3095
4	405076	127460	1	4076	8	FR	5000	ROBERTO	14/11/2013 07:01:01	*	25234
5	405077	127460	1	4076	1	FR	7500	ROBERTO	14/11/2013 07:01:17	*	847
6	405076	127460	1	4076	2	FR	8000	ROBERTO	14/11/2013 07:02:03	*	25233
7	405075	127460	1	4076	2	FR	2000	ROBERTO	14/11/2013 07:01:27	*	27146
8	405080	127460	1	4076	4	FR	4500	ROBERTO	14/11/2013 07:03:02	*	15407
9	405081	127460	1	4076	2	FR	5000	ROBERTO	14/11/2013 07:03:02	*	4084
10	405082	127460	1	4076	6	FR	4500	ROBERTO	14/11/2013 07:03:02	*	561
11	405083	127460	1	4076	8	FR	4500	ROBERTO	14/11/2013 07:03:02	*	211
12	405084	127460	1	4076	8	FR	4500	ROBERTO	14/11/2013 07:03:02	*	6235
13	405085	127460	1	4076	9	FR	5500	ROBERTO	14/11/2013 07:03:04	*	1741
14	405086	127460	1	4076	10	FR	7500	ROBERTO	14/11/2013 07:03:04	*	80
15	405105	127747	1	4076	3	FR	13500	JAVER	14/11/2013 19:40:16	*	8934
16	405106	127747	1	4076	4	FR	8500	JAVER	14/11/2013 19:40:23	*	312
17	405107	127747	1	4076	5	FR	4500	JAVER	14/11/2013 19:40:32	*	8934
18	405108	127747	1	4076	6	FR	4500	JAVER	14/11/2013 19:40:32	*	312
19	405182	127723	1	4076	10	FR	2000	JAVER	14/11/2013 19:34:57	*	516
20	405183	127723	1	4076	10	FR	7500	JAVER	14/11/2013 19:35:11	*	7918
21	405184	127723	1	4076	12	FR	8500	JAVER	14/11/2013 19:35:19	*	562
22	405185	127723	1	4076	12	FR	7500	JAVER	14/11/2013 19:35:19	*	562
23	405186	127724	1	4076	1	FR	10000	JAVER	14/11/2013 19:36:15	*	21688
24	405187	127725	1	4076	2	FR	8000	JAVER	14/11/2013 19:36:15	*	4457
25	405188	127725	1	4076	2	FR	8000	JAVER	14/11/2013 19:37:29	*	25123
26	405189	127725	1	4076	3	FR	4500	JAVER	14/11/2013 19:37:54	*	1008
27	405190	127725	1	4076	4	FR	8000	JAVER	14/11/2013 19:37:40	*	7225
28	405191	127725	1	4076	5	FR	9500	JAVER	14/11/2013 19:38:03	*	39950
29	405192	127725	1	4076	6	FR	10000	JAVER	14/11/2013 19:38:19	*	23101
30	405193	127725	1	4076	7	FR	4500	JAVER	14/11/2013 19:38:40	*	23986
31	405194	127726	1	4076	1	FR	1500	JAVER	14/11/2013 19:38:53	*	8623
32	405195	127726	1	4076	2	FR	2000	JAVER	14/11/2013 19:39:00	*	284
33	405196	127727	1	4076	1	FR	2000	JAVER	14/11/2013 19:40:04	*	575
34	405373	125420	1	4076	1	FR	12200	ROBERTO	06/11/2013 23:16:22	*	562
35	405374	125420	1	4076	2	FR	3500	ROBERTO	06/11/2013 23:16:22	*	17520
36	405375	125420	1	4076	3	FR	3500	ROBERTO	06/11/2013 23:17:11	*	790
37	405376	125420	1	4076	4	FR	5000	ROBERTO	06/11/2013 23:17:11	*	14461
38	405377	125420	1	4076	5	FR	3500	ROBERTO	06/11/2013 23:17:39	*	25145
39	405378	125420	1	4076	6	FR	4500	ROBERTO	06/11/2013 23:17:39	*	765
40	405379	125420	1	4076	2	FR	3500	ROBERTO	06/11/2013 23:18:45	*	13451

Figure 10. Tabla de Ventas Detalle.

IV-C. Dataware

Definimos el datawarehouse tomando como origen la base de datos explicada en el punto anterior. A partir del modelo transaccional creamos 3 cubos con sus tablas de hechos, métricas y dimensiones.

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 cabecera guarda datos tales como: fecha de la venta, el cliente que realizó la compra, la caja donde fue hecha la operación, el número de la factura y los montos totales de la venta. Las métricas establecidas en esta tabla de hechos son monto

total, monto exento, monto gravado, monto gravado 5% y monto gravado 10%.

TESISUSER.BI_DW_FACT_CABEC

P *	IDENTIFICADOR	NUMBER
F	ID_CLIENTE	NUMBER
F	ID_FECHA	NUMBER
F	ID_CAJA	NUMBER
	NRO_FACTURA	VARCHAR2(20 BYTE)
	MONTO_TOTAL	NUMBER
	MONTO_EXENTO	NUMBER
	MONTO_GRAVADO	NUMBER
	MONTO_GRAVADO5	NUMBER
	MONTO_GRAVADO10	NUMBER
BI_DW_FACT_CABEC_PK (IDENTIFICADOR)		
BI_DW_FACT_CABEC_PK (IDENTIFICADOR)		

Figure 11. Tabla de hechos Cabecera

- **Tabla de hechos Detalles:** almacena los datos históricos del detalle de cada transacción, cada registro contiene información del producto vendido tales como: el numero de comprobante, la fecha de venta, el proveedor del producto, el cliente que realizó la compra, la cantidad vendida del producto y el precio del producto. Las métricas asociadas a la tabla de hechos son, cantidad, precio unitario, precio unitario neto, impuesto, costo y el importe total.

TESISUSER.BI_DW_FACT_DETALLES

P *	IDENTIFICADOR	NUMBER
F	ID_COMPROBANTE	NUMBER
F	ID_FECHA	NUMBER
F	ID_PRODUCTO	NUMBER
F	ID_PROVEEDOR	NUMBER
F	ID_CLIENTE	NUMBER
	CANTIDAD	NUMBER
	PRECIO_UNITARIO	NUMBER
	PRECIO_UNIT_NETO	NUMBER
	IMPUESTO_UNITARIO	NUMBER
	COSTO_UNITARIO	NUMBER
	IMPORTE_ITEM	NUMBER
BI_DW_FACT_DETALLES_PK (IDENTIFICADOR)		
BI_DW_FACT_DETALLES_PK (IDENTIFICADOR)		
BI_DW_FACT_DETALLES_PK (IDENTIFICADOR)		
BI_DW_FACT_DETALLES_PK (IDENTIFICADOR)		

Figure 12. Tabla de hechos Detalles

- **Tabla de hechos Stock:** almacena los datos históricos de cada movimiento de compra de productos y de venta de productos. Cada registro de la tabla de hechos representa un movimiento realizado que puede corresponder a una compra o una venta de un producto, el producto en movimiento, la fecha del movimiento, la cantidad y los costos. Las métricas utilizadas para la tabla de hechos son: cantidad, precio unitario y costo unitario.

TESISUSER.BI_DW_FACT_STOCK

P *	IDENTIFICADOR	NUMBER
F	ID_MOVIMIENTO	NUMBER
F	ID_PRODUCTO	NUMBER
F	ID_FECHA	NUMBER
	VR_ACCION	VARCHAR2(20 BYTE)
	CANTIDAD	NUMBER
	PRECIO_UNITARIO	NUMBER
	COSTO_UNITARIO	NUMBER
BI_DW_FACT_STOCK_PK (IDENTIFICADOR)		
BI_DW_FACT_STOCK_PK (IDENTIFICADOR)		
BI_DW_FACT_STOCK_PK (IDENTIFICADOR)		
BI_DW_FACT_STOCK_PK (IDENTIFICADOR)		

Figure 13. Tabla de hechos Stock

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 dimensión fecha se pueden establecer niveles jerárquicos en días, semanas, meses, trimestres, semestres y años.

TESISUSER.BI_DW_DIM_FECHA	
P * IDENTIFICADOR	NUMBER
FECHA	DATE
DIA	CHAR (10 BYTE)
DIASEMANA	NUMBER
DIASEM	NUMBER
DIASEMANT	NUMBER
DIASEMANTER	DATE
DIASEMANTER	DATE
SEMANADELANHO	NUMBER
MES	CHAR (10 BYTE)
MESELANHO	NUMBER
TRIMESTRELANHO	NUMBER
ANHO	NUMBER
QUINCENADELANHO	NUMBER
BI_DW_DIM_FECHA_PK (IDENTIFICADOR)	
BI_DW_DIM_FECHA_PK (IDENTIFICADOR)	

Figure 14. Tabla dimensión Fecha

- **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.

TESISUSER.BI_DW_DIM_PRODUCTOS	
P * IDENTIFICADOR	NUMBER
DESCRIPCION	VARCHAR2 (100 BYTE)
FEC_CRE	DATE
CODIGO_BARRAS	VARCHAR2 (20 BYTE)
PRECIO_VENTA	NUMBER
COSTO_PROMEDIO	NUMBER
ID_PROVEEDOR	NUMBER
PORC_IMPUESTO	NUMBER
BI_DW_DIM_PRODUCTOS_PK (IDENTIFICADOR)	
BI_DW_DIM_PRODUCTOS_PK (IDENTIFICADOR)	

Figure 15. Tabla dimensión Productos

- **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.

TESISUSER.BI_DW_DIM_PROVEEDORES	
P * IDENTIFICADOR	NUMBER
DENOMINACION	VARCHAR2 (60 BYTE)
FEC_CRE	DATE
DIRECCION	VARCHAR2 (240 BYTE)
TELEFONO	VARCHAR2 (60 BYTE)
RUC	VARCHAR2 (15 BYTE)
BI_DW_DIM_PROVEEDORES_PK (IDENTIFICADOR)	
BI_DW_DIM_PROVEEDORES_PK (IDENTIFICADOR)	

Figure 16. Tabla dimensión Proveedores

- **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.

$$x = \frac{\sum(\text{Precio} \times \text{cantidad})}{\text{Total Tickets del período}}$$

Figure 19. Fórmula de Ticket Medio

TESISUSER.BI_DW_DIM_CLIENTES	
P * IDENTIFICADOR	NUMBER
NOMBRES	VARCHAR2 (50 BYTE)
APELLIDOS	VARCHAR2 (50 BYTE)
RAZON_SOCIAL	VARCHAR2 (100 BYTE)
DOCUMENTO	VARCHAR2 (20 BYTE)
DIRECCION	VARCHAR2 (240 BYTE)
TELEFONO	VARCHAR2 (40 BYTE)
EMAIL	VARCHAR2 (100 BYTE)
FEC_CRE	DATE
BI_DW_DIM_CLIENTES (IDENTIFICADOR)	
BI_DW_DIM_CLIENTES (IDENTIFICADOR)	

Figure 17. Tabla dimensión Clientes

- **Dimension 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.

TESISUSER.BI_DW_DIM_CAJA	
P * IDENTIFICADOR	NUMBER
NUMERO	VARCHAR2 (10 BYTE)
DESCRIPCION	VARCHAR2 (50 BYTE)
BI_DW_DIM_CAJA_PK (IDENTIFICADOR)	
BI_DW_DIM_CAJA_PK (IDENTIFICADOR)	

Figure 18. Tabla dimensión Cajas

IV-D. Definición de los KPI

Los KPI son un elemento vertebrador de la estrategia por su capacidad de comunicar resultados a todas las personas que forman parte del proyecto (directivos, gerentes, vendedores, etc). Con el uso de indicadores claves de rendimiento se trasladan a todas las personas cuáles son los elementos principales sobre los que se apoya la estrategia de organización, posibilita tener planes de accion concretos, ágiles y eficientes, esto apoya en la toma de decisiones basada en la información proporcionada por los indicadores.

En el marco de esta tesis, en esta sección se definirán los KPI que se utilizarán en el modelado de la solución de la Cantidad de Compra Óptima de Productos para la reposición de stock para el 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.

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.

Figure 28. Tabla de Valores KPI Mensual

IV-D2. 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.

KPI_CIFRA_VENTAS		
<	4.033.600	a Si el valor del KPI es menor al valor de comparación se asigna con "a"
4.033.600	6.382.100	b Si el valor del KPI está en el rango de comparación se asigna con "b"
>	6.382.100	c Si el valor del KPI es mayor al valor de comparación se asigna con "c"

KPI_COBERTURA_STOCK		
<	0,6	a Si el valor del KPI es menor al valor de comparación se asigna con "a"
0,6	1,5	b Si el valor del KPI está en el rango de comparación se asigna con "b"
>	1,5	c Si el valor del KPI es mayor al valor de comparación se asigna con "c"

KPI_ROTACION_STOCK		
<	0,4	a Si el valor del KPI es menor al valor de comparación se asigna con "a"
0,4	0,8	b Si el valor del KPI está en el rango de comparación se asigna con "b"
>	0,8	c Si el valor del KPI es mayor al valor de comparación se asigna con "c"

Figure 29. Tabla de etiquetado

De la figura 29, para cada KPI el rango de montos utilizados se toma de la tabla de valores de KPI (Figura 25). Ejemplo:

- < 4.033.600, donde el monto corresponde a $\text{MIN}(\text{KPI_CIFRA_VENTAS}) + (\text{MAX}(\text{KPI_CIFRA_VENTAS}) - \text{MIN}(\text{KPI_CIFRA_VENTAS}) * 40\%)$
- Entre 4.033.600 y 6.382.100, donde el primer monto corresponde a $\text{MIN}(\text{KPI_CIFRA_VENTAS}) + (\text{MAX}(\text{KPI_CIFRA_VENTAS}) - \text{MIN}(\text{KPI_CIFRA_VENTAS}) * 40\%)$ y el segundo monto corresponde a $\text{MIN}(\text{KPI_CIFRA_VENTAS}) + (\text{MAX}(\text{KPI_CIFRA_VENTAS}) - \text{MIN}(\text{KPI_CIFRA_VENTAS}) * 65\%)$
- > 6.382.100, donde el monto corresponde a $\text{MIN}(\text{KPI_CIFRA_VENTAS}) + (\text{MAX}(\text{KPI_CIFRA_VENTAS}) - \text{MIN}(\text{KPI_CIFRA_VENTAS}) * 65\%)$

Una vez asignado los valores “a”, “b”, “c”, en la tabla de emparejamientos de la figura 30, se busca y se asigna el valor de la etiqueta correspondiente.

aaa	Nada
aab	Medio
aac	Mucho
aba	Nada
abb	Mucho
abc	Medio
aca	Nada
acb	Nada
acc	Medio
baa	Nada
bab	Medio
bac	Mucho
bba	Nada
bbb	Nada
bbc	Mucho
bca	Nada
bcb	Medio
bcc	Nada
caa	Mucho
cab	Mucho
cac	Mucho
cba	Medio
cbb	Medio
cbc	Medio
cca	Medio
ccb	Mucho
ccc	Medio

Figure 30. Tabla de emparejamientos

Una vez realizado los pasos anteriores la tabla de valores de KPI, queda de la siguiente manera:

[illegible]

Figure 31. Tabla de valores de KPI con etiquetas

Una vez finalizado el etiquetado, los 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.














Name	Date modified	Type	Size
 5.csv	24/11/2016 08:19	CSV File	8 KB
 7.csv	24/11/2016 08:19	CSV File	7 KB
 8.csv	24/11/2016 08:19	CSV File	8 KB
 11.csv	24/11/2016 08:19	CSV File	8 KB
 12.csv	24/11/2016 08:19	CSV File	8 KB
 13.csv	24/11/2016 08:19	CSV File	7 KB
 22.csv	24/11/2016 08:19	CSV File	7 KB
 28.csv	24/11/2016 08:19	CSV File	8 KB
 35.csv	24/11/2016 08:20	CSV File	9 KB
 87.csv	24/11/2016 08:20	CSV File	7 KB
 88.csv	24/11/2016 08:20	CSV File	7 KB
 89.csv	24/11/2016 08:20	CSV File	7 KB
 90.csv	24/11/2016 08:20	CSV File	7 KB

Figure 32. Archivos CSV

Estos archivos son los datos que sirven como entrada para crear el modelo de Cantidad de Compra Óptima para la reposición de stock mediante algoritmos de aprendizaje automático

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 constituida 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. Introducción al modelado del aprendizaje automático

Para el problema de estudio.

V-B. Datos proveídos por business intelligence

La salida de business intelligence se constituye de archivos CSV que podemos expresar como se muestran en la Figura 1, Figura 2 y Figura 3.

La Figura 1 es una porción de un archivo CSV que contiene la salida de business intelligence calculada sobre las ventas mensuales de un determinado producto. Hay 812 productos diferentes analizados lo que equivale a 812 archivos CSV. En realidad la tabla de la Figura 1 tiene 34 filas sin incluir el encabezado lo que corresponde directamente a 34 instancias o ejemplos. Entonces, por cada producto analizado tenemos disponibles 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 TIKET	KPI CIFRA	KPI MARGEN	KPI ROTACION	KPI COEF	KPI COBERTURA	CANTIDAD	AÑO	MES	RESULTADO
MEDIO	VENTAS	COMERCIAL	STOCK	RENTABILIDAD	STOCK				
4222	72000	32534	1.407	45073	2.571	19	2013	12	Medio
4706	80000	34230	1.667	57017	1.364	20	2014	1	Medio
4600	72000	23095	3.6	104740	0.509	18	2014	2	Mucho
4200	84000	33944	7	237605	0.053	21	2014	3	Mucho
4222	76000	30716	3.495	106092	0.254	19	2014	4	Mucho
4000	44000	17790	2	35420	0.31	11	2014	5	Mucho
4000	36000	14517	2.25	32664	0.294	9	2014	6	Mucho
4222	76000	36162	1.152	41642	0.231	19	2014	7	Mucho
4000	76000	37616	0.927	34864	2.308	19	2014	8	Medio

Figura 33. Tabla de ejemplo que corresponde a métricas BI mensuales sobre las ventas de un producto.

La Figura 2 es una porción de un archivo CSV que contiene la salida de business intelligence calculada sobre las ventas quincenales de un determinado producto. Hay 808 productos diferentes analizados lo que equivale a 808 archivos CSV. En realidad la tabla de la Figura 2 tiene 68 filas sin incluir el encabezado lo que corresponde directamente a 68 instancias o ejemplos. Entonces, por cada producto analizado tenemos disponible 68 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 TIKET	KPI CIFRA	KPI MARGEN	KPI ROTACION	KPI COEF	KPI COBERTURA	CANTIDAD	AÑO	MES	QUINCENA	RESULTADO
MEDIO	VENTAS	COMERCIAL	STOCK	RENTABILIDAD	STOCK					
4444	40000	17195	0.789	13186	2.571	10	2013	12	23	Mucho
4000	36000	15439	0.821	9583	1.75	9	2013	12	24	Mucho
4444	40000	17195	1	1667	0.57	10	2014	1	1	Medio
5000	40000	17095	1.429	24345	0.57	10	2014	1	2	Mucho
4000	36000	14547	2	23095	0.931	9	2014	2	3	Medio
5143	36000	14547	18	26191	0	9	2014	2	4	Mucho
4364	48000	19396	1.6	38034	0.107	12	2014	3	5	Mucho
4000	36000	14547	0.947	13782	1.4	9	2014	3	6	Medio
4500	36000	14547	1.636	23805	0.5	9	2014	4	7	Mucho

Figura 34. Tabla de ejemplo que corresponde a métricas BI quincenales sobre las ventas de un producto.

La Figura 3 es una porción de un archivo CSV que contiene la salida de business intelligence calculada sobre las ventas semanales de un determinado producto. Hay 796 productos diferentes analizados lo que equivale a 796 archivos CSV. En realidad la tabla de la Figura 3 tiene 151 filas sin incluir el encabezado lo que corresponde directamente a 151 instancias o ejemplos. Entonces, por cada producto analizado tenemos disponible 151 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 TIKET	KPI CIFRA	KPI MARGEN	KPI ROTACION	KPI COEF	KPI COBERTURA	CANTIDAD	AÑO	MES	SEMANA	RESULTADO
MEDIO	VENTAS	COMERCIAL	STOCK	RENTABILIDAD	STOCK					
4657	20000	12008	0.455	5757	2.571	7	2013	12	49	Mucho
4000	4000	1715	0.061	104	3.4	1	2013	12	50	Mucho
4000	20000	8577	0.27	2218	4.364	5	2013	12	51	Nada
4000	8000	6952	0.211	1445	4.946	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
5000	12000	5146	0.353	1616	2.727	3	2014	1	2	Nada
4000	12000	5146	0.353	1616	2.1	3	2014	1	3	Nada
5600	28000	12008	1.077	12932	2.727	7	2014	1	4	Medio

Figura 35. Tabla de ejemplo que corresponde a métricas BI semanales sobre las ventas de un producto.

V-C. Esquema del procesamiento de las instancias

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

Luego, cada archivo de instancias se entrena con todos los algoritmos de clasificación WEKA posibles y la evaluación se hace tanto por el método Percentage Split así como 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 en el siguiente capítulo se analiza el significado de los resultados que contienen.

V-D. Entrenamiento y evaluación de las instancias

En la Figura 4 se muestra 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.

...

VII-A4. Resultados de valores mínimos alcanzados: ...

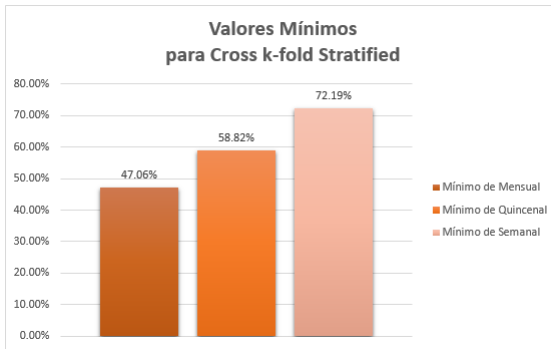


Figura 47. Min para Cross k-fold Stratified.

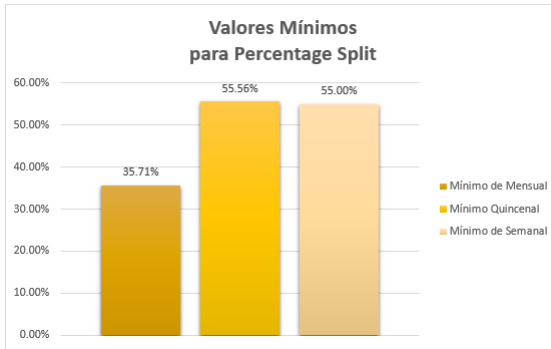


Figura 48. Min para Percentage Split.

...

VII-A5. Conteo de clasificadores para Cross k-fold Stratified: ...

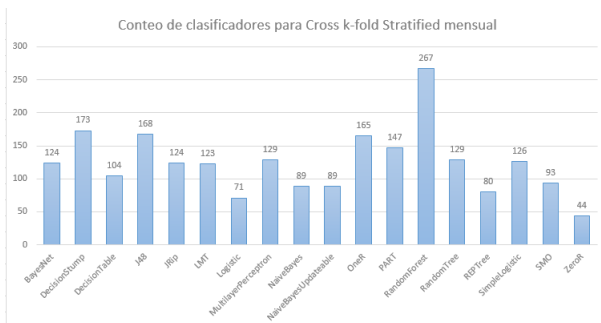


Figura 49. Cross Stratified mensual.

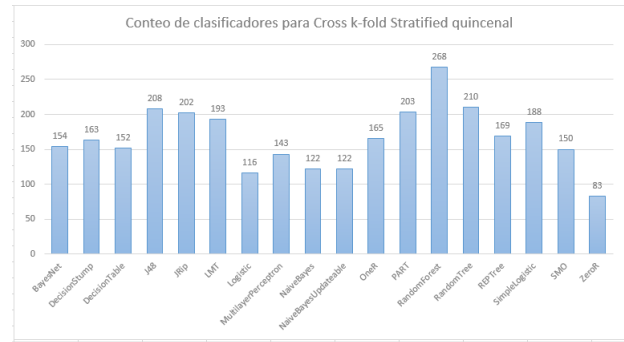


Figura 50. Cross Stratified quincenal.

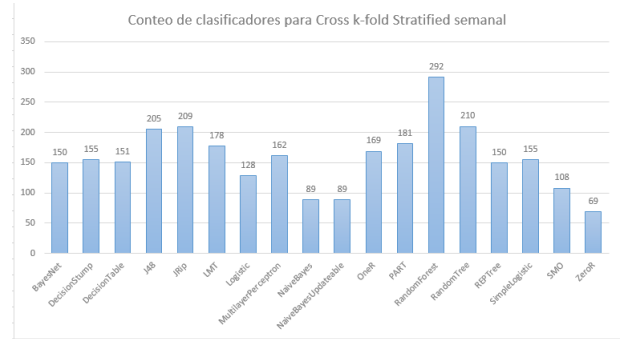


Figura 51. Cross Stratified semanal.

...

VII-A6. Conteo de clasificadores para Percentage Split:

...

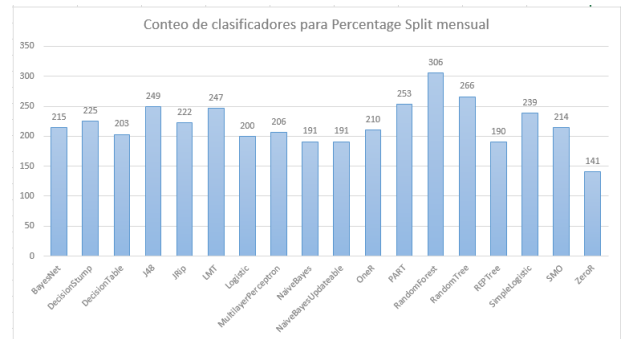


Figura 52. Percentage Split mensual.

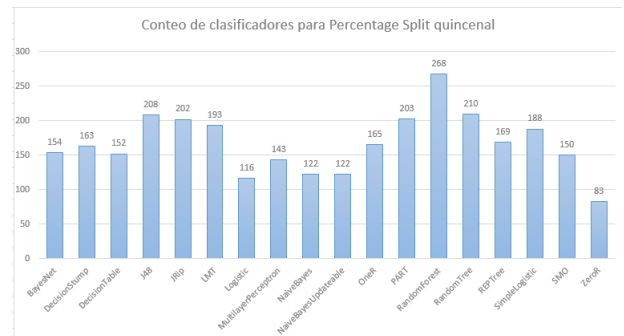


Figura 53. Percentage Split quincenal.

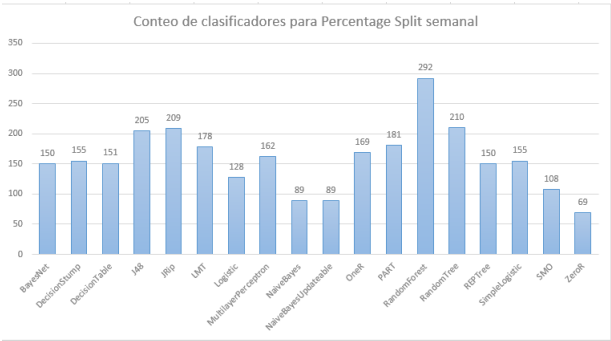


Figura 54. Percentage Split semanal.

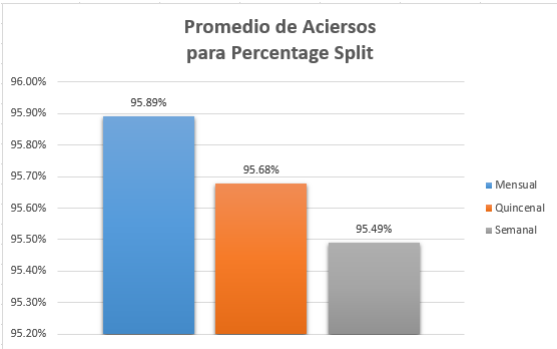


Figura 58. PACI para Percentage Split.

...

VII-B. Resultados con un limite inferior de Porcentaje de Aciertos del 90 %

VII-B1. Conteo de cantidad de productos analizados: ...

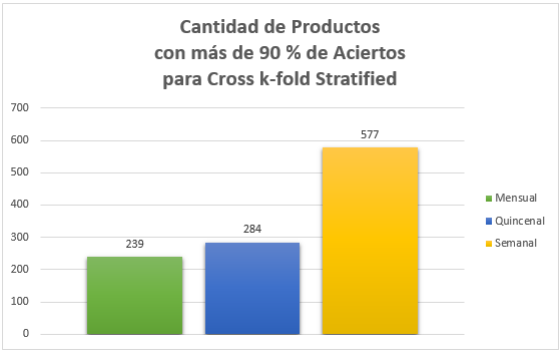


Figura 55. Conteo de la cantidad de productos analizados.

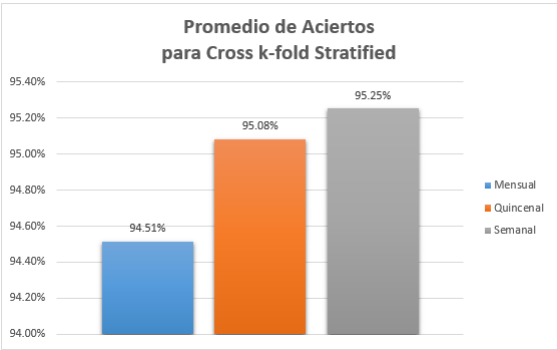


Figura 57. PACI para Cross k-fold Stratified.

...

VII-B3. Resultados de valores mínimos alcanzados: ...

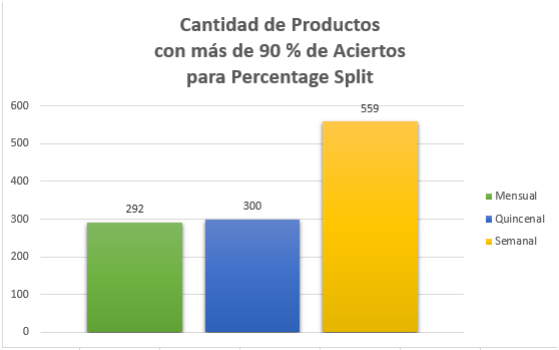


Figura 56. Conteo de la cantidad de productos analizados.

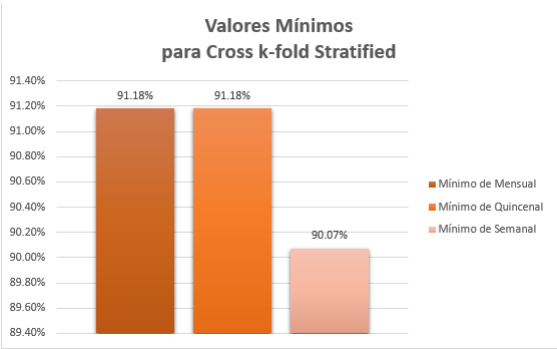


Figura 59. Min para Cross k-fold Stratified.

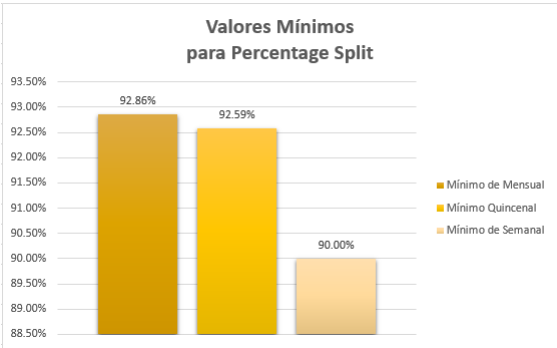


Figura 60. Min para Percentage Split.

...

VII-B2. Resultados del porcentaje de aciertos: ...

...

VIII. CONCLUSIONES PARA BI

Es necesario conclusiones para BI?

IX. CONCLUSIONES PARA MACHINE LEARNING

REFERENCIAS

- [1] Interface classifier.
- [2] Marcos Alvarez. *Cuadro de Mando Retail*. Profit, 2013.
- [3] Josep Lluís Cano. *Business Intelligence: Competir con informaci3n*. 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] Gartner. Glosario de gartner, www.gartner.com, enero 2006. gartner es una consultora internacional especializada en tecnolog3as de informaci3n y comunicaci3n, 01 2006.
- [6] Machine Learning Group. Weka 3: Data mining software in java.
- [7] W.H. Inmon. *Building the datawarehouse*. QED Press, 1992.
- [8] W.H. Inmon. *Building the datawarehouse*. Willey, 1996.
- [9] Jordi Conesa Josep Curto. *Introducci3n al Business Intelligence*. Editorial UOC, 2010.
- [10] Ralph Kimball. *The datawarehouse Toolkit*. John Wiley & Sons, Inc, 1992.
- [11] Jean Francois Puget. What is machine learning?, May 2016.
- [12] Arthur Samuel. Some studies in machine learning using the game of checker. *IBM Journal* 3, 211-229, 1959.
- [13] Alex Smola and S.V.N. Vishwanathan. *Introduction to Machine Learning*. The Press Syndicate of The University of Cambridge, 2008.
- [14] Hugh James Watson. Recent developments in datawarehousing: A tutorial. 2006.
- [15] Colin White Wayne Eckerson. Evaluating etl and data integration platforms. Technical report, TDWI Report Series, 2003.
- [16] Ian H. Witten, Frank Eibe, and Mark A. Hall. *Data Mining - Practical Machine Learning Tools and Techniques*. Copyright 2011 Elsevier Inc. All rights reserved, 2011.