“Automatización en la toma de decisiones para la reposición de stock a través de un modelo basado en Business Intelligence y Machine Learning”

A. Garcete, R. Benítez , D. Pinto-Roa

I. INTRODUCCIÃŞN

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.

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:

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

[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*

- 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

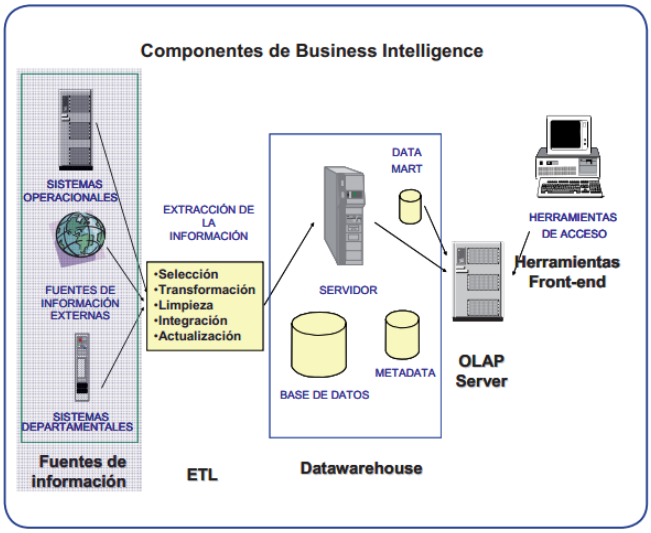


Figure 1. Fuentes de información

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

- 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 encuesta[2] 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.

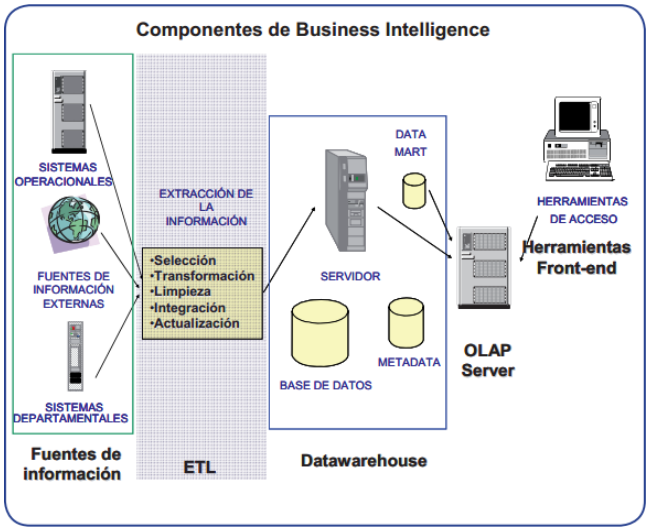


Figure 2. Proceso ETL

II-C2. ETL[3] – 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[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

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

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

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.

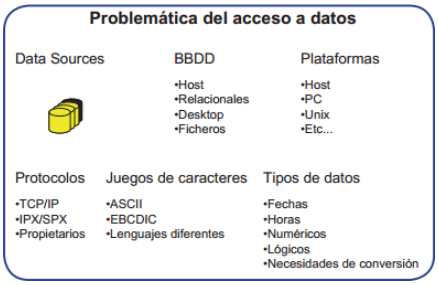


Figure 3. Problemas del acceso a datos

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 sumariza en los distintos modelos de análisis. El resultado de este proceso es la obtención de datos limpios, consistentes, sumarizados 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 lineas 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 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

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].

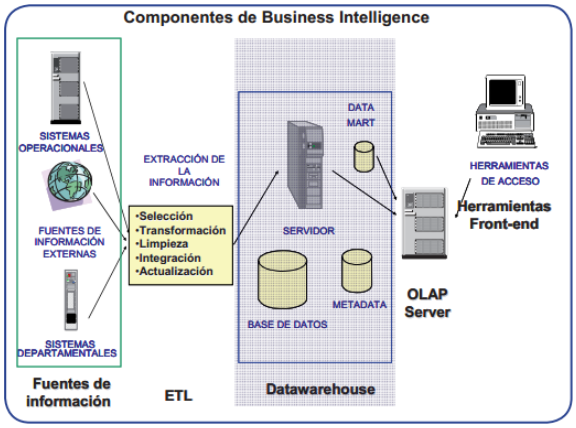


Figure 4. Datawarehouse

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.

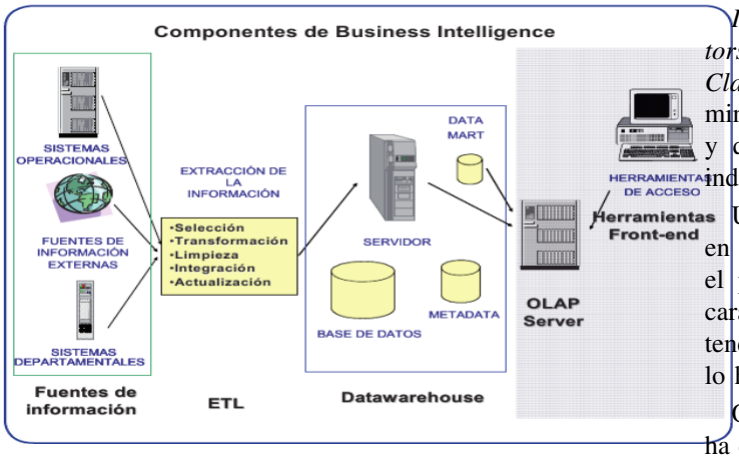


Figure 5. Herramientas de Bussines Intelligence

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 el 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 caracteristicas 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 siempres 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 buan predicción sobre datos aún no conocidos o imprevistos.

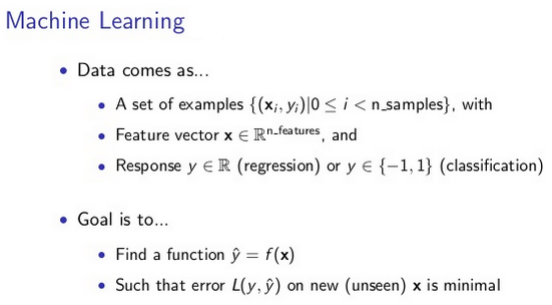


Figura 6. Definición presentada por Peter Prettenhofer y Gille Louppe.

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. Classifiation 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): :

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 ∈ {± 1} [13].:

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

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 resusstituyendo 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. Las secciones 5.3 y 5.4 revisan métodos ampliamente utilizados para hacer frente a este dilema.

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.

PARA COMPARAR varios esquemas de aprendizaje automático: Este es un trabajo para una prueba estadística basada en límites de confianza, el tipo que conocimos anteriormente al tratar de predecir el rendimiento real de una determinada tasa de error de prueba. Lo que queremos determinar es si un esquema es mejor o peor que otro en promedio, en todos los posibles conjuntos de datos de entrenamiento y prueba que se pueden extraer del dominio. Debido a que la cantidad de datos de entrenamiento afecta naturalmente el rendimiento, todos los conjuntos de datos deben ser del mismo tamaño. De hecho, el experimento podría repetirse con diferentes tamaños para obtener una curva de aprendizaje. For each learning scheme we can draw several datasets of the same size, obtain an accuracy estimate for each dataset using cross-validation, and compute the mean of the estimates. Each cross-validation experiment yields a different, independent error estimate. What we are interested in is the mean accuracy across all possible datasets of the same size, and whether this mean is greater for one scheme or the other. This is a job for a statistical device known as the T-TEST, or Student's t-test. Because the same cross-validation experiment can be used for both learning schemes to obtain a matched pair of results for each dataset, a more sensitive version of the t-test known as a paired t-test can be used.

III-H1. correct() - number of correctly classified instances.:

III-H2. pctCorrect() - percentage of correctly classified instances.:

III-H3. kappa() - KAPPA STATISTIC: CONTAR EL COSTO o COUNTING THE COST: Las evaluaciones que se

han discutido hasta ahora no tienen en cuenta el costo de tomar decisiones equivocadas, clasificaciones erróneas. Otros 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.:

III-H4. areaUnderROC(int classIndex) - CURVAS ROC o 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 × 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 × 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 .:

III-H5. recall(int classIndex) y precision(int classIndex) - RECALL-PRECISION CURVES: 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. 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, respectively. Diferentes técnicas dan diferentes compensaciones, y se pueden trazar como líneas diferentes en cualquiera de estas gráficas. TABLA 5.7:

III-H6. areaUnderPRC(int classIndex) - Area Under Precision-Recall Curve::

III-H7. Tasa de error (error rate): on the testing data. Predecir la tasa de error. Tenfold cross-validation is the standard way of measuring the error rate of a learning scheme on a particular dataset. The basic quality measure offered by the error rate is longer appropriate.:

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 retail[4], la cual trata acerca de la reposición de stock[5], 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 capitulo 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 reposision 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::

- £Cuando debemos realizar un pedido?
- £Qué cantidad debemos hacer el pedido?

En nuestro medio en numerosas ocaciones 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

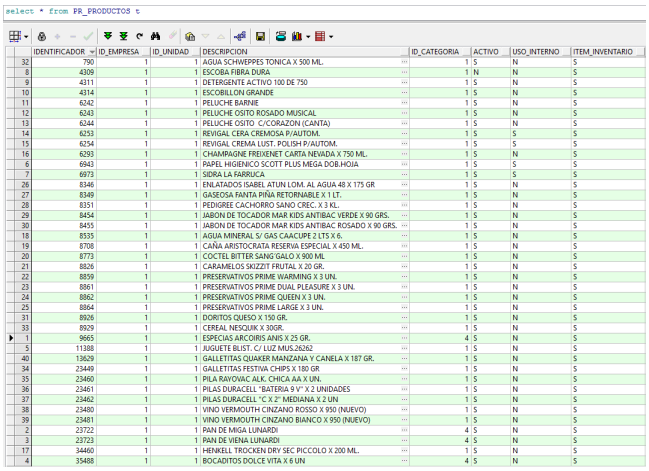
Para el presente trabajo, contamos con una base de datos auténtica con los registros de productos, proveedores, movimientos de compras, ventas y registro de stock realizados por una empresa retail. Los datos transaccionales contenidos en la base de datos corresponden a movimientos realizados entre el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

[4]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.

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

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

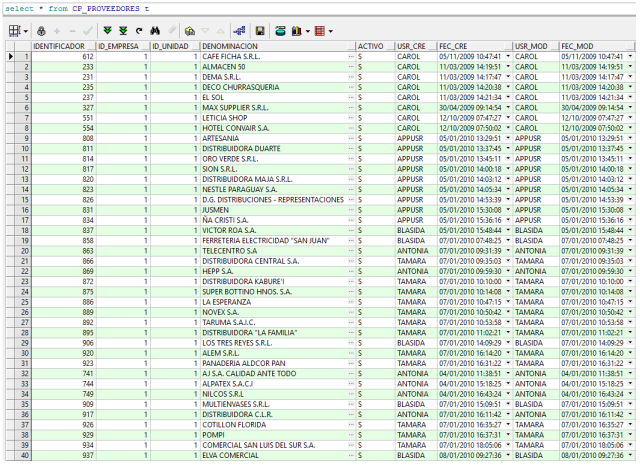


```
select * from PR_PRODUCTOS t
```

	IDENTIFICADOR	ID_EMPRESA	ID_UNIDAD	DESCRIPCION	ID_CATEGORIA	ACTIVO	USO_INTERNO	ITEM_INVENTARIO
32	790	1	1	AGUA SCHWEPPES TONICA X 500 ML.	1	S	N	S
8	4309	1	1	ESCOBA FIBRA DURA	1	N	N	S
9	4311	1	1	DETERGENTE ACTIVO 100 DE 750	1	S	N	S
10	4314	1	1	ESCOBILLON GRANDE	1	S	N	S
11	6242	1	1	PELUCHE BARNIE	1	S	N	S
12	6243	1	1	PELUCHE OSITO ROSADO MUSICAL	1	S	N	S
13	6244	1	1	PELUCHE OSITO C/CORAZON (CANTA)	1	S	N	S
14	6253	1	1	REVIGAL CERA CREMOSA P/AUTOM.	1	S	S	S
15	6254	1	1	REVIGAL CREMA LUST. POLISH P/AUTOM.	1	S	S	S
16	6293	1	1	CHAMPAGNE FREIXENET CARTA NEVADA X 750 ML.	1	S	N	S
6	6943	1	1	PAPEL HIGIENICO SCOTT PLUS MEGA DOB.HOJA	1	S	S	S
7	6973	1	1	SIDRA LA FARRUCA	1	S	S	S
26	8346	1	1	ENLATADOS ISABEL ATUN LOM. AL AGUA 48 X 175 GR	1	S	N	S
27	8349	1	1	GASEOSA FANTA PIÑA RETORNABLE X 1 LT.	1	S	N	S
28	8351	1	1	PEDIGREE CACHORRO SANO CREC. X 3 KL.	1	S	N	S
29	8454	1	1	JABON DE TOCADOR MAR KIDS ANTIBAC VERDE X 90 GRS.	1	S	N	S
30	8455	1	1	JABON DE TOCADOR MAR KIDS ANTIBAC ROSADO X 90 GRS.	1	S	N	S
18	8535	1	1	AGUA MINERAL S/ GAS CAACUPE 2 LTS X 6.	1	S	N	S
19	8708	1	1	CAÑA ARISTOCRATA RESERVA ESPECIAL X 450 ML.	1	S	N	S
20	8773	1	1	COCTEL BITTER SANG GALO X 900 ML	1	S	N	S
21	8826	1	1	CARAMELOS SKITTLES FRUTAL X 20 GR.	1	S	N	S
22	8859	1	1	PRESERVATIVOS PRIME WARMING X 3 UN.	1	S	N	S
23	8861	1	1	PRESERVATIVOS PRIME DUAL PLEASURE X 3 UN.	1	S	N	S
24	8862	1	1	PRESERVATIVOS PRIME QUEEN X 3 UN.	1	S	N	S
25	8864	1	1	PRESERVATIVOS PRIME LARGE X 3 UN.	1	S	N	S
31	8926	1	1	DORITOS QUESO X 150 GR.	1	S	N	S
33	8929	1	1	CEREAL NESQUIK X 30GR.	1	S	N	S
1	9665	1	1	ESPECIAS ARCORIS ANIS X 25 GR.	4	S	N	S
5	11388	1	1	JUGUETE BLIST. C/ LUZ MUS.26262	1	S	N	S
40	13629	1	1	GALLETITAS QUAKER MANZANA Y CANELA X 187 GR.	1	S	N	S
34	23449	1	1	GALLETITAS FESTIVA CHIPS X 180 GR	1	S	N	S
35	23460	1	1	PILA RAYOVAC ALK. CHICA AA X UN.	1	S	N	S
36	23461	1	1	PILAS DURACELL "BATERIA 9 V" X 2 UNIDADES	1	S	N	S
37	23462	1	1	PILAS DURACELL "C X 2" MEDIANA X 2 UN	1	S	N	S
38	23480	1	1	VINO VERMOUTH CINZANO ROSSO X 950 (NUEVO)	1	S	N	S
39	23481	1	1	VINO VERMOUTH CINZANO BIANCO X 950 (NUEVO)	1	S	N	S
2	23722	1	1	PAN DE MIGA LUNARDI	4	S	N	S
3	23723	1	1	PAN DE VIENA LUNARDI	4	S	N	S
17	34460	1	1	HEINKELL TROCKEN DRY SEC PICCOLO X 200 ML.	1	S	N	S
4	35488	1	1	BOCADITOS DOLCE VITA X 6 UN	4	S	N	S

Figure 7. Productos.

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



```
select * from CP_PROVEEDORES t
```

	IDENTIFICADOR	ID_EMPRESA	ID_UNIDAD	DENOMINACION	ACTIVO	USR_CRE	FEC_CRE	USR_MOD	FEC_MOD
1	612	1	1	CAFE FICHA S.R.L.	S	CAROL	05/11/2009 10:47:41	CAROL	05/11/2009 10:47:41
2	233	1	1	ALMACEN 50	S	CAROL	11/03/2009 14:19:51	CAROL	11/03/2009 14:19:51
3	231	1	1	DEMA S.R.L.	S	CAROL	11/03/2009 14:17:47	CAROL	11/03/2009 14:17:47
4	235	1	1	DECO CHURRASQUERIA	S	CAROL	11/03/2009 14:20:38	CAROL	11/03/2009 14:20:38
5	237	1	1	EL SOL	S	CAROL	11/03/2009 14:21:34	CAROL	11/03/2009 14:21:34
6	327	1	1	MAX SUPPLIER S.R.L.	S	CAROL	30/04/2009 09:14:54	CAROL	30/04/2009 09:14:54
7	551	1	1	LETICIA SHOP	S	CAROL	12/10/2009 07:47:27	CAROL	12/10/2009 07:47:27
8	554	1	1	HOTEL CONVAIR S.A.	S	CAROL	12/10/2009 07:50:02	CAROL	12/10/2009 07:50:02
9	808	1	1	ARTESANIA	S	APPUSR	05/01/2010 13:29:51	APPUSR	05/01/2010 13:29:51
10	811	1	1	DISTRIBUIDORA DUARTE	S	APPUSR	05/01/2010 13:37:45	APPUSR	05/01/2010 13:37:45
11	814	1	1	ORO VERDE S.R.L.	S	APPUSR	05/01/2010 13:45:11	APPUSR	05/01/2010 13:45:11
12	817	1	1	SION S.R.L.	S	APPUSR	05/01/2010 14:00:18	APPUSR	05/01/2010 14:00:18
13	820	1	1	DISTRIBUIDORA MAJA S.R.L.	S	APPUSR	05/01/2010 14:03:12	APPUSR	05/01/2010 14:03:12
14	823	1	1	NESTLE PARAGUAY S.A.	S	APPUSR	05/01/2010 14:05:34	APPUSR	05/01/2010 14:05:34
15	826	1	1	D.G. DISTRIBUCIONES - REPRESENTACIONES	S	APPUSR	05/01/2010 14:53:39	APPUSR	05/01/2010 14:53:39
16	831	1	1	JUSMEN	S	APPUSR	05/01/2010 15:30:08	APPUSR	05/01/2010 15:30:08
17	834	1	1	ÑA CRISTI S.A.	S	APPUSR	05/01/2010 15:36:16	APPUSR	05/01/2010 15:36:16
18	837	1	1	VICTOR ROA S.A.	S	BLASIDA	05/01/2010 15:48:44	BLASIDA	05/01/2010 15:48:44
19	858	1	1	FERRETERIA ELECTRICIDAD "SAN JUAN"	S	BLASIDA	07/01/2010 07:48:25	BLASIDA	07/01/2010 07:48:25
20	863	1	1	TELECENTRO S.A.	S	ANTONIA	07/01/2010 09:31:39	ANTONIA	07/01/2010 09:31:39
21	866	1	1	DISTRIBUIDORA CENTRAL S.A.	S	TAMARA	07/01/2010 09:35:03	TAMARA	07/01/2010 09:35:03
22	869	1	1	HEPP S.A.	S	ANTONIA	07/01/2010 09:59:30	ANTONIA	07/01/2010 09:59:30
23	872	1	1	DISTRIBUIDORA KABURE'I	S	TAMARA	07/01/2010 10:10:00	TAMARA	07/01/2010 10:10:00
24	875	1	1	SUPER BOTTINO HNOS. S.A.	S	TAMARA	07/01/2010 10:14:08	TAMARA	07/01/2010 10:14:08
25	886	1	1	LA ESPERANZA	S	TAMARA	07/01/2010 10:47:15	TAMARA	07/01/2010 10:47:15
26	889	1	1	NOVEX S.A.	S	TAMARA	07/01/2010 10:50:42	TAMARA	07/01/2010 10:50:42
27	892	1	1	TARUMA S.A.I.C.	S	TAMARA	07/01/2010 10:53:58	TAMARA	07/01/2010 10:53:58
28	895	1	1	DISTRIBUIDORA "LA FAMILIA"	S	TAMARA	07/01/2010 11:02:21	TAMARA	07/01/2010 11:02:21
29	906	1	1	LOS TRES REYES S.R.L.	S	BLASIDA	07/01/2010 14:09:29	BLASIDA	07/01/2010 14:09:29
30	920	1	1	ALEM S.R.L.	S	TAMARA	07/01/2010 16:14:20	TAMARA	07/01/2010 16:14:20
31	923	1	1	PANADERIA ALDCOR PAN	S	TAMARA	07/01/2010 16:31:22	TAMARA	07/01/2010 16:31:22
32	741	1	1	A.I.S.A. CALIDAD ANTE TODO	S	ANTONIA	04/01/2010 11:38:51	ANTONIA	04/01/2010 11:38:51
33	744	1	1	ALPATEX S.A.C.I	S	ANTONIA	04/01/2010 15:18:25	ANTONIA	04/01/2010 15:18:25
34	749	1	1	NILCOS S.R.L	S	ANTONIA	04/01/2010 16:43:24	ANTONIA	04/01/2010 16:43:24
35	909	1	1	MULTIENVASES S.R.L.	S	BLASIDA	07/01/2010 15:09:51	BLASIDA	07/01/2010 15:09:51
36	917	1	1	DISTRIBUIDORA C.L.R.	S	ANTONIA	07/01/2010 16:11:42	ANTONIA	07/01/2010 16:11:42
37	926	1	1	COTILLON FLORIDA	S	TAMARA	07/01/2010 16:35:27	TAMARA	07/01/2010 16:35:27
38	929	1	1	POMPI	S	TAMARA	07/01/2010 16:37:31	TAMARA	07/01/2010 16:37:31
39	934	1	1	COMERCIAL SAN LUIS DEL SUR S.A.	S	TAMARA	07/01/2010 18:05:06	TAMARA	07/01/2010 18:05:06
40	937	1	1	ELVA COMERCIAL	S	BLASIDA	08/01/2010 09:27:36	BLASIDA	08/01/2010 09:27:36

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 04/10/16. Contiene datos de la fecha, numero de factura, cliente, montos totales entre otros datos.

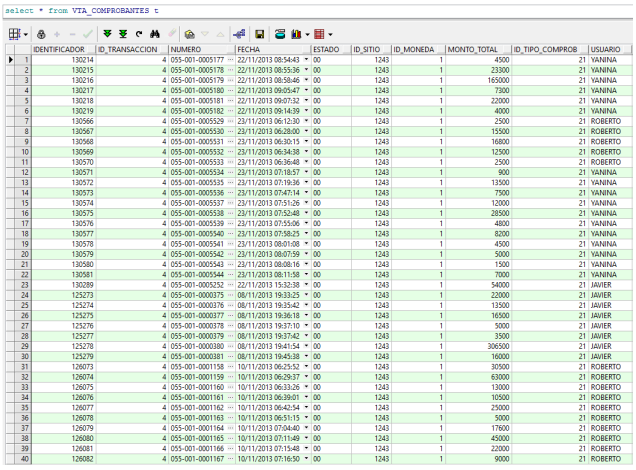


Figure 9. Tabla de Ventas Cabecera.

Tabla de Ventas Detalle: La tabla de ventas detalle contiene los registros de los productos que fueron comercializados, cada detalle esta 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.

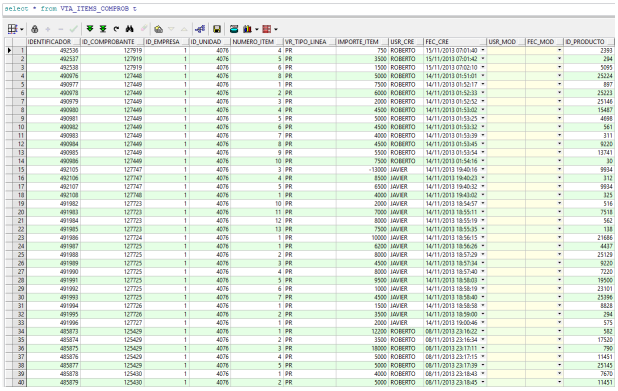


Figure 10. Tabla de Ventas Detalle.

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%.

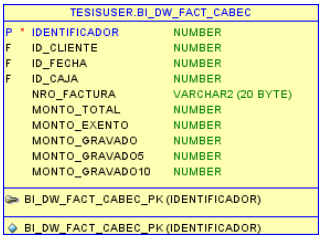


Figure 11. Tabla de hechos Cabecera

- **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.

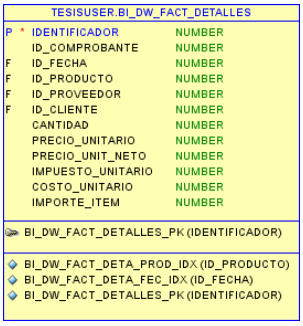


Figure 12. Tabla de hechos Detalles

- **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.

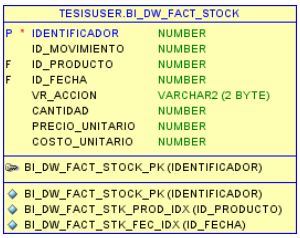


Figure 13. Tabla de hechos Stock

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 dimesión fecha se pueden establecer niveles jerárquicos en dias, semanas, meses, trimestres, semestres y años.

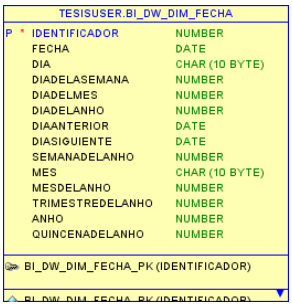


Figure 14. Tabla dimensión Fecha

- **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.

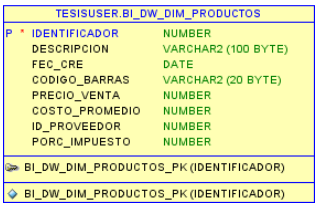


Figure 15. Tabla dimensión Productos

- **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.

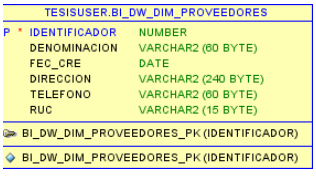


Figure 16. Tabla dimensión Proveedores

- **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.

$$x = \frac{\sum(\text{Precio} * \text{cantidad})}{\text{TotalTickes del periodo}}$$

Figure 19. Fórmula de Ticket Medio

Figure 17. Tabla dimensión Clientes

- **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.

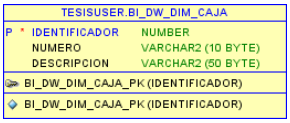


Figure 18. Tabla dimensión Cajas

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.

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

Figure 20. 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.

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

Figure 21. 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 22. 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 23. 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.

$$x = \text{Stock actual / Promedio cantidad venta último 3 periodos}$$

Figure 24. Fórmula Cobertura de Stock

IV-D1. 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 detallada en la figura 25. Además de los valores de los KPI, en cada registro adicionalmente se guarda la información de la cantidad, fecha, año, mes, quincena y semana.

TESISUSER.BI_DW_KPI_VALUES

P *	IDENTIFICADOR	NUMBER
	KPI_TIKET_MEDIO	NUMBER
	KPI_CIFRA_VENTAS	NUMBER
	KPI_MARGEN_COMERCIAL	NUMBER
	KPI_ROTACION_STOCK	NUMBER
	KPI_COEF_RENTABILIDAD	NUMBER
	KPI_COBERTURA_STOCK	NUMBER
	CANTIDAD	NUMBER
	FECHA	DATE
	MES	NUMBER
	ANHO	NUMBER
	SEMANA	NUMBER
	QUINCENA	NUMBER
	RESULTADO	VARCHAR2 (50 BYTE)
	ID_PRODUCTO	NUMBER
	PERIODO	VARCHAR2 (1 BYTE)

BI_DW_KPI_VALUES_PK (IDENTIFICADOR)

KPI_TIKET_MEDIO	KPI_CIFRA_VENTAS	KPI_MARGEN_COMERCIAL	KPI_ROTACION_STOCK	KPI_COEF_RENTABILIDAD	KPI_COBERTURA_STOCK	CANTIDAD	FECHA	ANHO	MES	QUINCENA	SEMANA	ID_PRODUCTO	PERIODO
6933,33333333333	104000	25213,0808080808	16	403409,292929293	0	16	01/12/2014	2014	12	23		10644	Q
6964,28571428571	97500	23637,202020202	0,810810810810811	19165,2989352989	0,193548387096774	15	16/12/2014	2014	12	24		10644	Q
6500	117000	28364,5454545455	0,692307692307692	19636,993006993	2,83783783783784	18	01/01/2015	2015	1	1		10644	Q
6861,11111111111	123500	29940,3939393939	0,844444444444444	25282,9993265993	1,04081632653061	19	16/01/2015	2015	1	2		10644	Q
6500	110500	26788,9090909091	0,871794871794872	23354,4335664336	1,61538461538462	17	01/02/2015	2015	2	3		10644	Q
6500	71500	17334	0,536585365853658	9301,17073170732	0,611111111111111	11	16/02/2015	2015	2	4		10644	Q
12000	156000	37819,5151515152	0,888888888888889	33617,3468013468	1,91489361702128	24	01/03/2015	2015	3	5		10644	Q
6500	84500	20485,5353535354	0,4	8194,21414141414	1,38461538461538	13	16/03/2015	2015	3	6		10644	Q
6500	78000	18909,8181818182	0,342857142857143	6483,36623376623	2,5625	12	01/04/2015	2015	4	7		10644	Q
6500	58500	14182,3636363636	0,36734693877551	5209,84786641929	1,77551020408163	9	16/04/2015	2015	4	8		10644	Q
6500	65000	15758,2121212121	0,476190476190476	7503,91053391053	1,76470588235294	10	01/05/2015	2015	5	9		10644	Q
7222,22222222222	65000	15758,3333333333	0,588235294117647	9269,60784313726	2,12903225806452	10	16/05/2015	2015	5	10		10644	Q

Figure 25. Tabla Valores de los KPI

Cada registro de los valores obtenidos de los KPI corresponden a un determinado producto y un periodo de tiempo determinado. Ejemplo el primer registro del figura 25 corresponde a la información obtenida para el producto con ID igual a 10644 y corresponde al periodo de la primera quincena del mes de diciembre del 2014, esta quincena corresponde a la número 23 del año.

PERIODOS DE TIEMPO: Agrupamos el cálculo de los valores de los KPI de cada producto en 3 periodos de tiempo: semanal, quincenal y mensual.

SEMANAL: En la figura 26 se puede observar una parte de los valores obtenidos para un producto por el rango de tiempo semanal.

El campo SEMANA se completa de acuerdo al periodo calculado, en el campo PERIODO, se guarda el valor S para semanal.

KPI_TIKET_MEDIO	KPI_CIFRA_VENTAS	KPI_MARGEN_COMERCIAL	KPI_ROTACION_STOCK	KPI_COEF_RENTABILIDAD	KPI_COBERTURA_STOCK	CANTIDAD	FECHA	ANHO	MES	QUINCENA	SEMANA	ID_PRODUCTO	PERIODO
5000	25000	7380,95238095238	0,769230769230769	5677,65567765568	1,2	5	01/12/2013	2013	12		49	133	S
5000	20000	5904,7619047619	0,571428571428571	3374,14965986395	2,07692307692308	4	08/12/2013	2013	12		50	133	S
5000	15000	4428,57142857143	0,272727272727273	1207,79220779221	0,9375	3	15/12/2013	2013	12		51	133	S
5000	10000	2952,38095238095	0,125	369,047619047619	4,25	2	22/12/2013	2013	12		52	133	S
5000	10000	2952,38095238095	0,142857142857143	421,768707482993	5	2	29/12/2013	2013	12		53	133	S
5000	5000	1476,19047619048	0,08	118,095238095238	5,57142857142857	1	01/01/2014	2014	1		1	133	S
5000	15000	4428,57142857143	0,285714285714286	1265,30612244898	7,2	3	05/01/2014	2014	1		2	133	S
5000	20000	5904,7619047619	0,571428571428571	3374,14965986395	4,5	4	12/01/2014	2014	1		3	133	S
5000	20000	5904,7619047619	1,33333333333333	7873,01587301587	1,875	4	19/01/2014	2014	1		4	133	S
5000	10000	2952,38095238095	0,2	590,47619047619	0,272727272727273	2	26/01/2014	2014	1		5	133	S

Figure 26. Tabla de Valores KPI Semanal

QUINCENAL: En la figura 27 se puede observar una parte de los valores obtenidos para un producto por el rango de tiempo quincenal.

El campo QUINCENA se completa de acuerdo al periodo calculado, en el campo PERIODO, se guarda el valor Q para quincenal.

	KPI_TIKET_MEDIO	KPI_CIFRA_VENTAS	KPI_MARGEN_COMERCIAL	KPI_ROTACION_STOCK	KPI_COEF_RENTABILIDAD	KPI_COBERTURA_STOCK	CANTIDAD	FECHA	ANHO	MES	QUINCENA	SEMANA	ID_PRODUCTO	PERIODO
1	5000	45000	13285,7142857143	2	26571,4285714286	1,2	9	01/12/2013	2013	12	23		133	Q
2	5000	35000	10333,3333333333	0,777777777777778	8037,03703703704	0,789473684210526	7	16/12/2013	2013	12	24		133	Q
3	5000	35000	10333,3333333333	0,736842105263158	7614,0350877193	1,625	7	01/01/2014	2014	1	1		133	Q
4	5000	35000	10333,3333333333	0,56	5786,66666666667	0,782608695652174	7	16/01/2014	2014	1	2		133	Q
5	5000	35000	10333,3333333333	0,451612903225806	4666,66666666667	2,71428571428571	7	01/02/2014	2014	2	3		133	Q
6	5000	15000	4428,57142857143	0,285714285714286	1265,30612244898	1,71428571428571	3	16/02/2014	2014	2	4		133	Q
7	5000	5000	1476,19047619048	0,117647058823529	173,669467787115	1,58823529411765	1	01/03/2014	2014	3	5		133	Q
8	0	0	0	0	0	2,18181818181818	0	16/03/2014	2014	3	6		133	Q
9	0	0	0	0	0	6	0	01/04/2014	2014	4	7		133	Q
10	2500	5000	1476,19047619048	0,133333333333333	196,825396825397	24	1	16/04/2014	2014	4	8		133	Q
11	5000	10000	2952,38095238095	0,333333333333333	984,126984126984	21	2	01/05/2014	2014	5	9		133	Q
12	5000	10000	2952,38095238095	0,5	1476,19047619048	5	2	16/05/2014	2014	5	10		133	Q
13	5000	10000	2952,38095238095	1	2952,38095238095	1,8	2	01/06/2014	2014	6	11		133	Q

Figure 27. Tabla de Valores KPI Quincenal

MENSUAL: En la figura 28 se puede observar una parte de los valores obtenidos para un producto por el rango de tiempo mensual.

El campo MES se completa para todos los casos del periodo calculado, en el campo PERIODO, se guarda el valor M para mensual.

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 MIN(KPI_CIFRA_VENTAS) + (MAX(KPI_CIFRA_VENTAS) - MIN(KPI_CIFRA_VENTAS)* 40%)
- Entre 4.033.600 y 6.382.100, donde el primer monto corresponde a MIN(KPI_CIFRA_VENTAS) + (MAX(KPI_CIFRA_VENTAS) - MIN(KPI_CIFRA_VENTAS)* 40%) y el segundo monto corresponde MIN(KPI_CIFRA_VENTAS) + (MAX(KPI_CIFRA_VENTAS) - MIN(KPI_CIFRA_VENTAS)* 65%)
- > 6.382.100, donde el monto corresponde a MIN(KPI_CIFRA_VENTAS) + (MAX(KPI_CIFRA_VENTAS) - MIN(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:

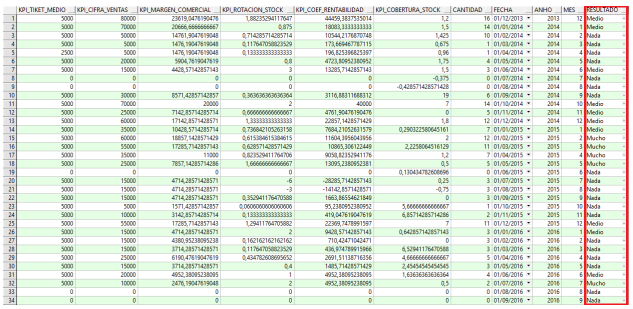


Figure 31. Tabla de valores de KPI con etiquetas

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.

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 reposicion 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 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. 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 MEDIO	KPI CIFRA VENTAS	KPI MARGEN COMERCIAL	KPI ROTACION STOCK	KPI COEF RENTABILIDAD	KPI COBERTURA STOCK	CANTIDAD	AÑO	MES	RESULTADO
4222	76000	32594	1.407	45873	2.571	19	2013	12	**Medio**
4706	80000	34210	1.667	57017	1.364	20	2014	1	**Medio**
4500	72000	29095	3.6	104740	0.509	18	2014	2	**Mucho**
4200	84000	33944	7	237605	0.053	21	2014	3	**Mucho**
4222	76000	30711	3.455	106092	0.254	19	2014	4	**Mucho**
4000	44000	17710	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 MEDIO	KPI CIFRA VENTAS	KPI MARGEN COMERCIAL	KPI ROTACION STOCK	KPI COEF RENTABILIDAD	KPI COBERTURA STOCK	CANTIDAD	AÑO	MES	QUINCENA	RESULTADO
4444	40000	17155	0.769	13196	2.571	10	2013	12	23	**Mucho**
4000	36000	15439	0.621	9583	1.75	9	2013	12	24	**Mucho**
4444	40000	17155	1	17155	1.667	10	2014	1	1	**Medio**
5000	40000	17155	1.429	24365	0.517	10	2014	1	2	**Mucho**
4000	36000	14547	2	29095	0.931	9	2014	2	3	**Medio**
5143	36000	14547	18	261851	0	9	2014	2	4	**Mucho**
4364	48000	19396	1.6	31034	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 MEDIO	KPI CIFRA VENTAS	KPI MARGEN COMERCIAL	KPI ROTACION STOCK	KPI COEF RENTABILIDAD	KPI COBERTURA STOCK	CANTIDAD	AÑO	MES	SEMANA	RESULTADO
4667	28000	12008	0.483	5797	2.571	7	2013	12	49	**Mucho**
4000	4000	1715	0.061	104	3.4	1	2013	12	50	**Nada**
4000	20000	8577	0.27	2316	4.364	5	2013	12	51	**Nada**
4000	16000	6862	0.211	1445	4.846	4	2013	12	52	**Nada**
4000	8000	3431	0.125	429	5.1	2	2013	12	53	**Nada**
4000	20000	8577	0.4	3431	4.091	5	2014	1	1	**Medio**
6000	12000	5146	0.353	1816	2.727	3	2014	1	2	**Nada**
4000	12000	5146	0.353	1816	2.1	3	2014	1	3	**Nada**
5600	28000	12008	1.077	12932	2.727	7	2014	1	4	**Medio**

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.

```
//Conjunto de clasificadores.
Classifier[] modelos = {
    //BAYES.
    new BayesNet(),
    new NaiveBayes(),
    new NaiveBayesUpdateable(),
    //FUNCTIONS.
    new Logistic(),
    new MultilayerPerceptron(),
    new SimpleLogistic(),
    new SMO(),
    //RULES.
    new OneR(),
    new DecisionTable(),
    new JRip(),
    new PART(),
    new ZeroR(),
    //TREES.
    new DecisionStump(),
    new J48(),
    new LMT(),
    new RandomForest(),
    new RandomTree(),
    new REPTree(),
};
```

Figura 36. Clasificadores WEKA utilizados en la implementeación.

En la Figura 5 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 37. Esquema de entrenamiento y evaluación en WEKA.

V-E. *Métricas de los resultados de la evaluación*

En la Figura 6 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, Recall, Precision, F-Measure y Area Under Precision-Recall Curve (AUPRC).

```
//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 38. Esquema de obtención de las métricas de evaluación en WEKA.

V-F. *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 1.3, hay dos tablas que se generan al concluir el proceso de aprendizaje automático.

En la Figura 7 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.

PROD	PERIODO	TIPO CLASIF	CLASIFICADOR	TIPO EVAL	PORC ACIER	KAPPA	Nada - PREC	Nada - RECALL	Nada - FMEA	Nada - ROCA	Nada - PRCA	Medio - PREC	Medio - RECALL	Medio - FMEA	Medio - ROCA	Medio - PRCA	Mucho - PREC	Mucho - RECALL	Mucho - FMEA	Mucho - ROCA	Mucho - PRCA
100	MENSUAL	bayes	BayesNet	split	57.14	0.22	0.5	0.75	0.6	0.73	0.45	0	0	0	0.25	0.14	0.63	0.63	0.63	0.56	0.6
100	MENSUAL	bayes	NaiveBayes	split	42.86	-0.1	0.33	0.25	0.29	0.78	0.53	0	0	0	0.54	0.33	0.5	0.63	0.56	0.59	0.7
100	MENSUAL	bayes	NaiveBayesUpdateable	split	42.86	-0.1	0.33	0.25	0.29	0.78	0.53	0	0	0	0.54	0.33	0.5	0.63	0.56	0.59	0.7
100	MENSUAL	functions	Logistic	split	57.14	0.25	0.5	0.5	0.5	0.73	0.47	0.5	0.5	0.5	0.63	0.67	0.63	0.63	0.63	0.67	0.61
100	MENSUAL	functions	MultilayerPerceptron	split	50	0.02	0	0	0	0.65	0.28	0.5	0.5	0.5	0.92	0.75	0.55	0.75	0.63	0.56	0.73
100	MENSUAL	functions	SimpleLogistic	split	71.43	0.5	0.75	0.75	0.75	0.85	0.73	0.5	0.5	0.5	0.96	0.83	0.75	0.75	0.75	0.85	0.82
100	MENSUAL	functions	SMO	split	57.14	0	0	0	0	0.65	0.3	0	0	0	0.92	0.5	0.67	1	0.73	0.5	0.67
100	MENSUAL	rules	OneR	split	57.14	0.22	0.5	0.75	0.6	0.73	0.45	0	0	0	0.5	0.14	0.63	0.63	0.63	0.56	0.6
100	MENSUAL	rules	DecisionTable	split	57.14	0.22	0.5	0.75	0.6	0.73	0.45	0	0	0	0.25	0.14	0.63	0.63	0.63	0.56	0.6
100	MENSUAL	rules	JRip	split	57.14	0.07	0	0	0	0.38	0.24	0	0	0	0.46	0.14	0.62	1	0.76	0.59	0.62
100	MENSUAL	rules	PART	split	42.86	0.02	0.33	0.25	0.29	0.53	0.3	0	0	0	0.38	0.14	0.63	0.63	0.63	0.59	0.62
100	MENSUAL	rules	ZeroR	split	57.14	0	0	0	0	0.5	0.29	0	0	0	0.5	0.14	0.57	1	0.72	0.5	0.57
100	MENSUAL	trees	DecisionStump	split	57.14	0.22	0.5	0.75	0.6	0.73	0.45	0	0	0	0.25	0.14	0.63	0.63	0.63	0.56	0.6
100	MENSUAL	trees	J48	split	42.86	0.02	0.33	0.25	0.29	0.53	0.3	0	0	0	0.38	0.14	0.63	0.63	0.63	0.59	0.62
100	MENSUAL	trees	LMT	split	71.43	0.5	0.75	0.75	0.75	0.85	0.73	0.5	0.5	0.5	0.96	0.83	0.75	0.75	0.75	0.85	0.92
100	MENSUAL	trees	RandomForest	split	64.29	0.33	0.6	0.75	0.67	0.74	0.68	0	0	0	0.58	0.24	0.67	0.75	0.71	0.76	0.83
100	MENSUAL	trees	RandomTree	split	50	0.01	0.5	0.25	0.33	0.59	0.34	0	0	0	0.38	0.14	0.67	0.75	0.71	0.63	0.64
100	MENSUAL	trees	REPTree	split	42.86	0.02	0.33	0.25	0.29	0.53	0.3	0	0	0	0.38	0.14	0.63	0.63	0.63	0.59	0.62
100	QUINCENAL	bayes	BayesNet	split	55.56	0.34	0.5	0.5	0.5	0.67	0.42	0.71	0.45	0.56	0.71	0.67	0.5	0.75	0.6	0.72	0.46
100	QUINCENAL	bayes	NaiveBayes	split	66.67	0.49	0.5	0.5	0.5	0.74	0.66	0.82	0.82	0.82	0.91	0.84	0.63	0.63	0.63	0.69	0.79
100	QUINCENAL	bayes	NaiveBayesUpdateable	split	66.67	0.49	0.5	0.5	0.5	0.74	0.66	0.82	0.82	0.82	0.91	0.84	0.63	0.63	0.63	0.69	0.79
100	QUINCENAL	functions	Logistic	split	55.56	0.34	0.5	0.25	0.33	0.57	0.4	0.86	0.55	0.67	0.85	0.77	0.44	0.88	0.58	0.72	0.44
100	QUINCENAL	functions	MultilayerPerceptron	split	55.56	0.35	0.6	0.38	0.46	0.73	0.58	1	0.55	0.71	0.88	0.8	0.38	0.75	0.5	0.7	0.45
100	QUINCENAL	functions	SimpleLogistic	split	59.26	0.39	0.5	0.25	0.33	0.72	0.63	0.79	0.64	0.7	0.92	0.95	0.5	0.92	0.64	0.84	0.72
100	QUINCENAL	functions	SMO	split	66.67	0.5	1	0.25	0.4	0.99	0.47	1	0.73	0.84	0.94	0.89	0.47	1	0.64	0.76	0.47
100	QUINCENAL	rules	OneR	split	59.26	0.39	0.5	0.38	0.43	0.61	0.37	0.89	0.73	0.8	0.83	0.76	0.42	0.63	0.5	0.63	0.37
100	QUINCENAL	rules	DecisionTable	split	48.15	0.24	0.44	0.5	0.47	0.61	0.38	0.6	0.27	0.37	0.61	0.47	0.46	0.75	0.57	0.7	0.42
100	QUINCENAL	rules	JRip	split	74.07	0.61	1	0.25	0.4	0.69	0.53	0.91	0.91	0.91	0.92	0.95	0.57	1	0.73	0.84	0.57
100	QUINCENAL	rules	PART	split	55.56	0.33	0.5	0.25	0.33	0.6	0.37	0.7	0.64	0.67	0.72	0.59	0.46	0.75	0.57	0.69	0.42
100	QUINCENAL	rules	ZeroR	split	28.63	0	0	0	0	0.5	0.3	0	0	0	0.5	0.41	0.3	1	0.46	0.5	0.3
100	SEMANAL	bayes	BayesNet	split	71.67	0.53	0.6	0.53	0.56	0.81	0.63	0.84	0.84	0.84	0.91	0.9	0.54	0.64	0.58	0.85	0.51
100	SEMANAL	bayes	NaiveBayes	split	53.33	0.2	0.29	0.12	0.17	0.73	0.43	0.71	0.84	0.77	0.88	0.92	0.2	0.27	0.23	0.63	0.25
100	SEMANAL	bayes	NaiveBayesUpdateable	split	53.33	0.2	0.29	0.12	0.17	0.73	0.43	0.71	0.84	0.77	0.88	0.92	0.2	0.27	0.23	0.63	0.25
100	SEMANAL	functions	Logistic	split	63.33	0.34	0.63	0.29	0.4	0.83	0.62	0.7	0.94	0.8	0.94	0.96	0.33	0.27	0.3	0.79	0.42
100	SEMANAL	functions	MultilayerPerceptron	split	55	0.22	0.33	0.24	0.28	0.68	0.41	0.71	0.84	0.77	0.87	0.92	0.2	0.18	0.19	0.67	0.27
100	SEMANAL	functions	SimpleLogistic	split	56.67	0.16	0.38	0.18	0.24	0.79	0.49	0.63	0.97	0.77	0.91	0.94	0	0	0	0.73	0.32
100	SEMANAL	functions	SMO	split	60	0.2	0.6	0.18	0.27	0.55	0.35	0.59	0.97	0.73	0.59	0.56	0	0.18	0.26	0.79	0.43

Figura 39. Resultados de la evaluación del aprendizaje automático con el método Percentage Split.

En la Figura 8 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.

PROD.	PERIODO	TIPO CLASIF	CLASIFICADOR	TIPO EVAL	PORC ACIER.	KAPPA	Nada - PRESC	Nada - RCALL	Nada - FMEA	Nada - ROCA	Medio - PRICA	Medio - PRESC	Medio - RCALL	Medio - FMEA	Medio - ROCA	Medio - PRICA	Mucho - PRESC	Mucho - RCALL	Mucho - FMEA	Mucho - ROCA	Mucho - PRICA
100	MENSUAL	bagas	BagaRan	crossEstratificado	59.92	0.2	0.75	0.6	0.67	0.73	0.52	0.2	0.31	0.43	0.41	0.24	0.67	0.9	0.72	0.6	0.67
100	MENSUAL	bagas	NaiveBagas	crossEstratificado	62.94	0.25	0.29	0.4	0.33	0.59	0.3	0.38	0.56	0.45	0.67	0.49	0.79	0.95	0.65	0.79	0.78
100	MENSUAL	bagas	NaiveBagasUpdateable	crossEstratificado	62.94	0.25	0.29	0.4	0.33	0.59	0.3	0.38	0.56	0.45	0.67	0.49	0.79	0.95	0.65	0.79	0.78
100	MENSUAL	functions	Logistic	crossEstratificado	64.71	0.27	0.4	0.4	0.4	0.69	0.44	0.56	0.56	0.56	0.62	0.37	0.75	0.75	0.75	0.62	0.67
100	MENSUAL	functions	MultilayerPerceptron	crossEstratificado	55.88	0.07	0.5	0.2	0.29	0.56	0.3	0.4	0.44	0.62	0.62	0.41	0.64	0.7	0.67	0.67	0.78
100	MENSUAL	functions	SimpleLogistic	crossEstratificado	74.47	0.57	0.6	0.6	0.6	0.78	0.62	0.71	0.58	0.63	0.75	0.65	0.82	0.9	0.86	0.83	0.86
100	MENSUAL	functions	SMO	crossEstratificado	55.88	-0.05	0	0	0	0.57	0.36	0	0	0	0.47	0.26	0.58	0.95	0.72	0.48	0.56
100	MENSUAL	rules	ZeroR	crossEstratificado	59.92	0.42	0	0	0	0.6	0.35	0.64	0.78	0.7	0.81	0.55	0.74	0.95	0.79	0.71	0.72
100	MENSUAL	rules	DecisionTable	crossEstratificado	62.94	0.07	1	0.4	0.57	0.7	0.55	0.14	0.11	0.13	0.27	0.24	0.6	0.75	0.67	0.43	0.54
100	MENSUAL	rules	JRip	crossEstratificado	70.59	0.44	1	0.6	0.75	0.81	0.69	0.5	0.44	0.47	0.62	0.39	0.74	0.95	0.79	0.74	0.74
100	MENSUAL	rules	PART	crossEstratificado	64.71	0.36	0.75	0.6	0.67	0.73	0.44	0.44	0.44	0.44	0.67	0.36	0.71	0.75	0.73	0.64	0.67
100	MENSUAL	rules	ZeroR	crossEstratificado	59.92	0	0	0	0	0.6	0.35	0	0	0	0.4	0.24	0.59	0.9	0.74	0.41	0.56
100	MENSUAL	trees	DecisionStump	crossEstratificado	55.88	0.11	1	0.4	0.57	0.73	0.57	0.17	0.11	0.13	0.41	0.26	0.62	0.8	0.7	0.46	0.56
100	MENSUAL	trees	J48	crossEstratificado	61.76	0.27	1	0.6	0.75	0.88	0.76	0.38	0.33	0.35	0.54	0.33	0.85	0.75	0.7	0.64	0.68
100	MENSUAL	trees	LMT	crossEstratificado	76.47	0.57	0.6	0.6	0.6	0.78	0.62	0.71	0.56	0.63	0.75	0.59	0.82	0.9	0.86	0.83	0.86
100	MENSUAL	trees	RandomForest	crossEstratificado	73.53	0.49	0.75	0.6	0.67	0.8	0.67	0.67	0.44	0.53	0.76	0.68	0.79	0.9	0.82	0.86	0.91
100	MENSUAL	trees	RandomTree	crossEstratificado	64.71	0.34	0.6	0.6	0.6	0.77	0.42	0.5	0.33	0.4	0.61	0.34	0.7	0.8	0.74	0.65	0.67
100	MENSUAL	trees	REPTree	crossEstratificado	61.76	0.22	1	0.2	0.33	0.63	0.35	0.43	0.33	0.38	0.61	0.27	0.85	0.95	0.74	0.63	0.7
100	QUINCENAL	bagas	BagaRan	crossEstratificado	69.82	0.52	0.74	0.71	0.72	0.79	0.66	0.53	0.59	0.56	0.69	0.41	0.77	0.74	0.75	0.8	0.65
100	QUINCENAL	bagas	NaiveBagas	crossEstratificado	70.59	0.56	0.74	0.83	0.78	0.8	0.65	0.55	0.65	0.59	0.76	0.59	0.81	0.63	0.71	0.89	0.62
100	QUINCENAL	bagas	NaiveBagasUpdateable	crossEstratificado	70.59	0.56	0.74	0.83	0.78	0.8	0.65	0.55	0.65	0.59	0.76	0.59	0.81	0.63	0.71	0.89	0.62
100	QUINCENAL	functions	Logistic	crossEstratificado	66.18	0.48	0.74	0.71	0.72	0.82	0.68	0.44	0.41	0.42	0.72	0.45	0.72	0.79	0.75	0.81	0.69
100	QUINCENAL	functions	MultilayerPerceptron	crossEstratificado	66.18	0.49	0.75	0.75	0.75	0.86	0.61	0.53	0.53	0.53	0.73	0.5	0.67	0.67	0.67	0.82	0.74
100	QUINCENAL	functions	SimpleLogistic	crossEstratificado	69.12	0.53	0.77	0.71	0.74	0.85	0.9	0.53	0.53	0.53	0.82	0.75	0.72	0.79	0.75	0.93	0.82
100	QUINCENAL	functions	SMO	crossEstratificado	63.24	0.41	0.66	0.79	0.72	0.81	0.61	0.5	0.06	0.11	0.57	0.29	0.62	0.95	0.72	0.76	0.59
100	QUINCENAL	rules	ZeroR	crossEstratificado	55.88	0.24	0.75	0.67	0.71	0.78	0.62	0.41	0.53	0.43	0.64	0.33	0.82	0.95	0.8	0.59	0.66
100	QUINCENAL	rules	DecisionTable	crossEstratificado	70.59	0.55	0.74	0.71	0.72	0.84	0.67	0.5	0.47	0.48	0.57	0.33	0.79	0.85	0.82	0.83	0.69
100	QUINCENAL	rules	JRip	crossEstratificado	70.59	0.54	0.86	0.79	0.83	0.89	0.82	0.55	0.35	0.43	0.7	0.5	0.86	0.95	0.74	0.79	0.82
100	QUINCENAL	rules	PART	crossEstratificado	70.59	0.55	0.75	0.75	0.75	0.85	0.77	0.54	0.41	0.47	0.66	0.45	0.74	0.85	0.79	0.83	0.66
100	QUINCENAL	rules	ZeroR	crossEstratificado	39.71	0	0	0	0	0.42	0.32	0	0	0	0.4	0.22	0.4	1	0.57	0.45	0.39
100	SEMANAL	bagas	BagaRan	crossEstratificado	72.39	0.52	0.95	0.77	0.85	0.88	0.94	0.43	0.55	0.48	0.8	0.45	0.56	0.76	0.64	0.89	0.47
100	SEMANAL	bagas	NaiveBagas	crossEstratificado	69.23	0.4	0.91	0.83	0.82	0.83	0.89	0.45	0.42	0.43	0.74	0.45	0.46	0.44	0.45	0.79	0.43
100	SEMANAL	bagas	NaiveBagasUpdateable	crossEstratificado	68.21	0.4	0.91	0.83	0.82	0.83	0.89	0.45	0.42	0.43	0.74	0.45	0.46	0.44	0.45	0.79	0.43
100	SEMANAL	functions	Logistic	crossEstratificado	72.39	0.45	0.85	0.82	0.88	0.92	0.95	0.38	0.32	0.35	0.78	0.43	0.82	0.44	0.46	0.85	0.54
100	SEMANAL	functions	MultilayerPerceptron	crossEstratificado	67.65	0.37	0.82	0.87	0.85	0.89	0.94	0.26	0.23	0.32	0.73	0.28	0.42	0.4	0.45	0.78	0.34
100	SEMANAL	functions	SimpleLogistic	crossEstratificado	72.39	0.44	0.79	0.91	0.94	0.91	0.95	0.46	0.39	0.43	0.82	0.52	0.65	0.44	0.52	0.94	0.57
100	SEMANAL	functions	SMO	crossEstratificado	66.21	0.29	0.72	0.94	0.81	0.86	0.71	0.5	0.42	0.46	0.64	0.32	1	0.94	0.66	0.57	0.32

Figura 40. Resultados de la evaluación del aprendizaje automático con el método Stratified K-fold Cross Validation.

VI. Resultados de Bussines Intelligence

VI-A. Valores obtenidos

Para el conjunto total de datos disponibles en el dataware para cada producto se realizaron 3 ejecuciones del algoritmo de cálculo de los valores de los KPI para rangos de tiempo diferentes: semanal, quincenal y mensual.

VI-A1. Rango de tiempo.:

Semanal: Para todos los productos existentes en la tabla de dimensiones Productos se realizó el cálculo de los valores de los KPI por periodos de tiempo semanal, como restricción se aplicó que si un producto no fuese vendido en un periodo de 75 semanas consecutivas se descartan los cálculos para dicho producto, se dicha restricción debido a que al no ser vendido por un periodo largo de tiempo los valores KPI generados dan como resultado 0(Cero) lo cual no tiene relevancia en el modelado.

Para cada producto considerado se genera 151 registros que serán utilizados como valores de entrada para el modelado de la solución mediante el aprendizaje automático.

Del universo total de productos, sólo 992 productos cumplieron con las restricciones, es decir, esto productos no cesaron las ventas por un periodo de tiempo mayor o igual a 75 semanas.

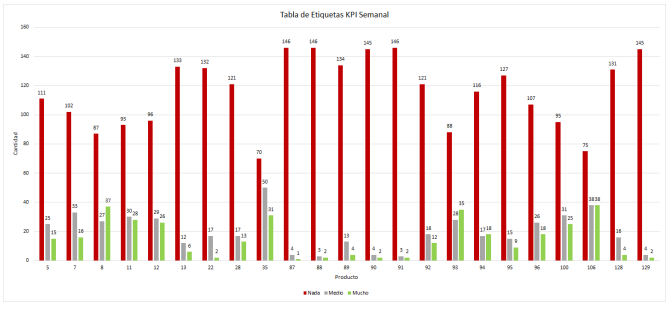


Figure 41. Tabla de etiquetas KPI Semanal

Quincenal.: Para todos los productos existentes en la tabla de dimensiones Productos se realizó el cálculo de los valores de los KPI por periodos de tiempo quincenal, como restricción se aplicó que si un producto no fuese vendido en un periodo de 36 quincenas consecutivas se descartan los cálculos para dicho producto, se dicha restricción debido a que al no ser vendido por un periodo largo de tiempo los valores KPI generados dan como resultado 0(Cero) lo cual no tiene relevancia en el modelado.

Para cada producto considerado se genera 68 registros que serán utilizados como valores de entrada para el modelado de la solución mediante el aprendizaje automático.

Del universo total de productos, sólo 1050 productos cumplieron con las restricciones, es decir, esto productos no cesaron las ventas por un periodo de tiempo mayor o igual a 36 quincenas.

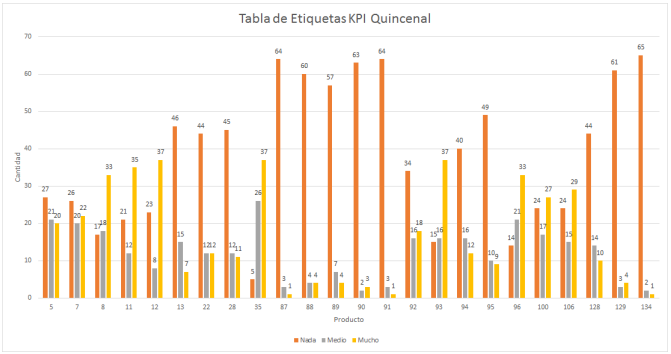


Figure 42. Tabla de etiquetas KPI Quincenal

Mensual.: Para todos los productos existentes en la tabla de dimensiones Productos se realizó el cálculo de los valores de los KPI por periodos de tiempo mensual, como restricción se aplicó que si un producto no fuese vendido en un periodo de 18 meses consecutivas se descartan los cálculos para dicho producto, se dicha restricción debido a que al no ser vendido por un periodo largo de tiempo los valores KPI generados dan como resultado 0(Cero) lo cual no tiene relevancia en el modelado.

Para cada producto considerado se genera 34 registros que serán utilizados como valores de entrada para el modelado de la solución mediante el aprendizaje automático.

Del universo total de productos, sólo 1052 productos cumplieron con las restricciones, es decir, esto productos no cesaron las ventas por un periodo de tiempo mayor o igual a 18 meses.

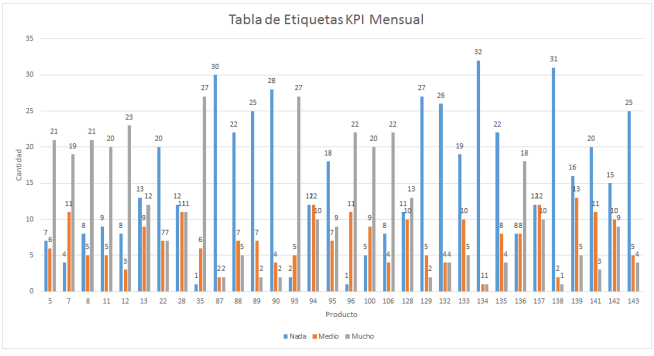


Figure 43. Tabla de etiquetas KPI Mensual

VII. Resultados del Aprendizaje Automático

Resultados preliminares del Aprendizaje Automático.

VII-A. Resultados sin un limite inferior permitido

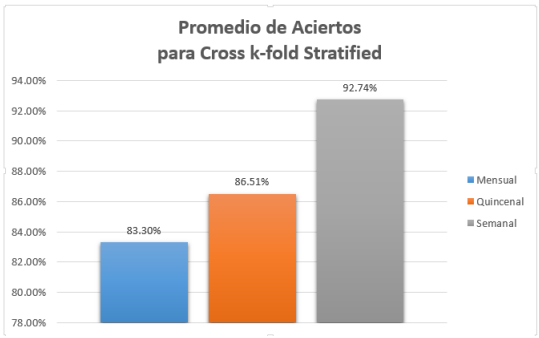


Figura 44. PACI para Cross k-fold Stratified.

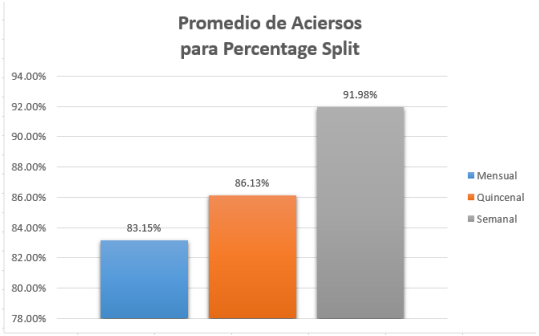


Figura 45. PACI para Percentage Split.

VII-A1. Resultados del porcentaje de aciertos:

Figura 46. Max1

VII-A2. Resultados de valores máximos:

Figura 47. Max2

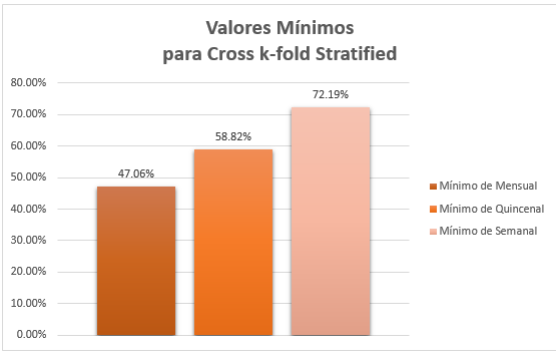


Figura 48. Min1

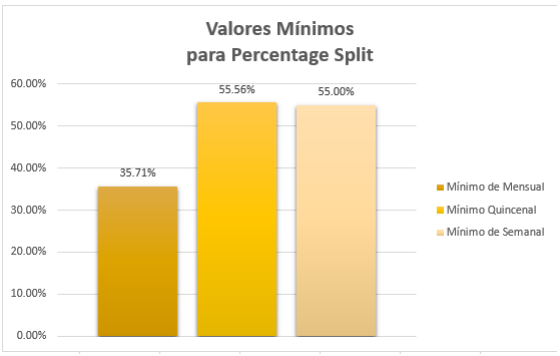


Figura 49. Min2

VII-A3. Resultados de valores mínimos:

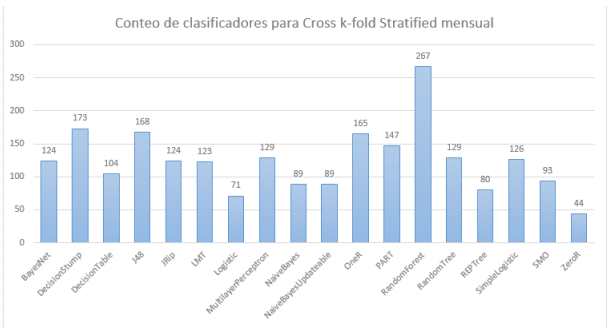


Figura 50. Cross Stratified mensual.

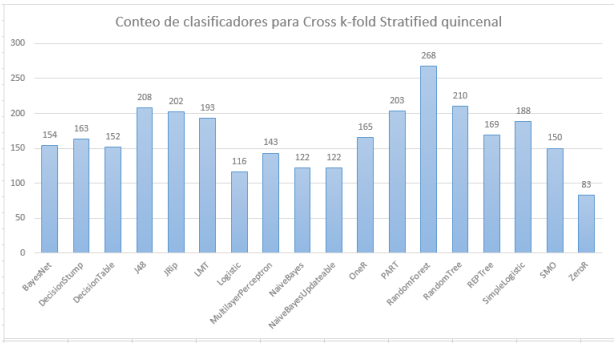


Figura 51. Cross Stratified quincenal.

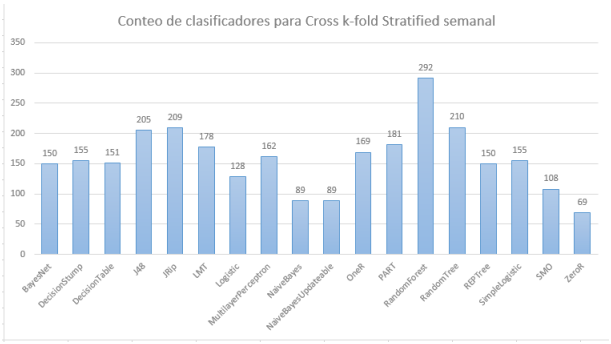


Figura 52. Cross Stratified semanal.

VII-A4. Conteo de clasificadores para Cross k-fold Stratified:

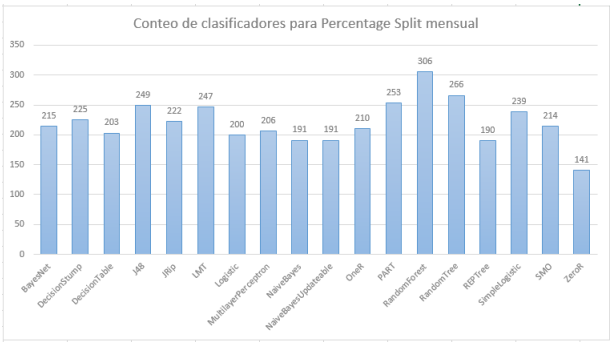


Figura 53. Percentage Split mensual.

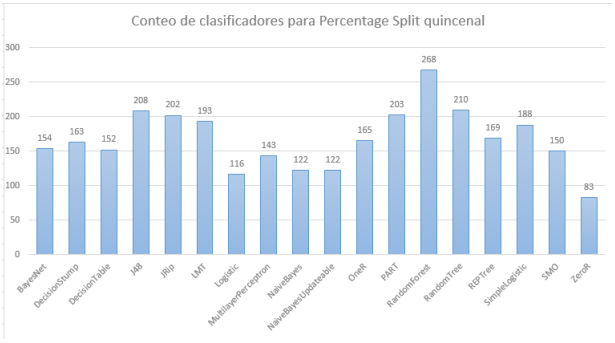


Figura 54. Percentage Split quincenal.

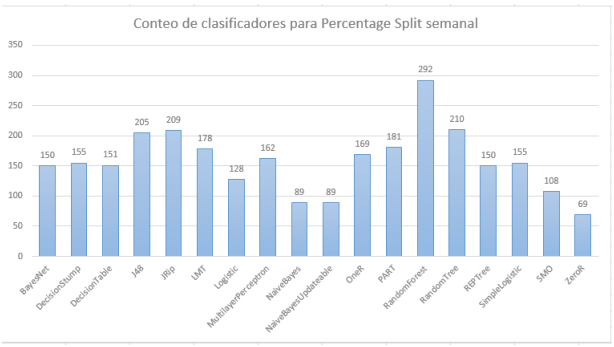


Figura 55. Percentage Split semanal.

VII-A5. Conteo de clasificadores para Percentage Split:

VII-B. Resultados con un límite inferior ZeroR

VIII. Referencias

[1] Interface classifier.

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

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

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

[5] Gartner. Glosario de gartner, www.gartner.com, enero 2006. gartner es una consultora internacional especializada en tecnologÃ■as de informaciÃşn y comunicaciÃşn, 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. *IntroducciÃşn al Busines 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 Âl' 2011 Elsevier Inc. All rights reserved, 2011.