Universidad CENFOTEC

Maestría en Tecnología de Bases de Datos

MBD-802 Proyecto de Investigación Aplicada 2

Proyecto

Tema

Desarrollo de una Nueva Aplicación de Análisis de Demanda de Productos para las Compras Regionales y un Modelo de Predicción de la Variable de Flete y Seguro en las Compras para el Pre-costeo de Producto en Grupo Transmerquim al Utilizar Herramientas de Machine Learning.

Estudiante

González Moya, Julio

Fecha: Agosto del 2016

Declaratoria de derechos de autor

Yo, Julio Gonzalez Moya, número de identificación: 1-0347-0348, estudiante de la Universidad Cenfotec, de la carrera Maestría en Tecnologías de Bases de Datos, por este medio se prohíbe que el presente documento pueda ser consultado, por el tiempo establecido entre el estudiante y la Universidad Cenfotec.

Dedicatoria

A Dios todo poderoso, a mi familia y en especial a mi padre.

Resumen

Uno de los pilares en un negocio de distribución de materias primas es la determinación de los factores que inciden sobre la demanda en la cadena de suministros. Tener una certeza de los movimientos que se hacen a través del tiempo y que a la postre genera un patrón del proceso de importación de productos futuros, es por esto que se motiva la creación de un mecanismo que sea sencillo y automático en la integración de variables que formulen en el final del proceso un cálculo de posible demanda al introducir cálculos predictivos de otras variables que en conjunto suman un costo del producto listo para la venta. Con este panorama el departamento de importaciones en conjunto con el departamento de tecnología de la información se han propuesto llevar a otro nivel de análisis de información los datos que suman la base en la toma de decisión en la importación de artículos, con lo cual buscan liberar tiempo empleado para la fabricación de estos reportes y sumar ese tiempo al análisis de los mismos al llevar esto a un nuevo nivel en la automatización y en el análisis de los patrones que se generan con la información ya procesada. Adicional a esto, se busca llevar un nuevo elemento, en el pre-costeo de la mercadería con la aplicación de técnicas de Machine Learning en el cálculo de costos por flete y seguro de las cargas que se importan. Para esto se emplean algoritmos supervisados y especialmente algoritmos de regresión estadística, que es una de las técnicas más empleadas cuando se busca determinar una variable en función de una o más variables explicativas o relacionadas. Estas técnicas permiten el pre-costeo de productos en puerto con lo cual se agiliza la comercialización de los mismos.

Este trabajo se desarrolla en dos fases la primera desarrolla una herramienta multidimensional que extrae de diferentes fuentes los datos necesarios para una análisis de demanda futura de materias primas y la segunda fase un análisis experimental de variables de costeo inherentes a la importación de productos al utilizar técnicas de Machine Learning y específicamente algoritmos supervisados, como regresión lineal, regresión de redes neuronales, regresión lineal bayesiana entre otros algoritmos de regresión.

Palabras claves. Cadena de suministros, regresión estadística, análisis predictivo, Machine Learning.

Tabla de Contenido

Tabla de Figuras

Tabla de Cuadros

1 CAPÍTULO 1. INTRODUCCIÓN

1.1 GENERALIDADES

GTM es una empresa latinoamericana líder en el mercadeo de distribución de químicos y materias primas para la industria en general. Es, asimismo, proveedor de servicios logísticos y soluciones integrales para distribuidores químicos y clientes en sectores industriales como petróleo y gas, agricultura, pinturas y cubrimientos, adhesivos, tratamiento de aguas, alimentos y cuidado personal.

GTM tiene una extensa red de distribución e infraestructura logística con más de 42 instalaciones en 12 países a lo largo de Latinoamérica y cuenta con oficinas de servicios de abastecimiento en Estados Unidos, India y China. Todas las instalaciones de GTM están certificadas bajo las normas ISO 9001 e ISO 14001.

GTM es una empresa independiente que emplea alrededor de 520 personas, atraídas por su interesante cultura empresarial, donde destacan valores compartidos como Excelencia, Integridad, Espíritu Emprendedor y Trabajo en Equipo.

1.2 ANTECEDENTES DEL PROBLEMA

GTM es una empresa que busca seguir posicionándose como empresa líder en la distribución de químicos y materias primas para la industria en general, por lo que busca la manera de realizar el proceso de compra de una manera más eficiente y con herramientas de última tecnología.

Dentro de su esquema organizacional se cuenta con una oficina de compras globales en Houston, Texas. Se encarga de realizar las compras globales de las diferentes compañías distribuidas a lo largo de Latinoamérica, bajo este panorama las compras se realizan sobre la demanda que las diferentes compañías realizan. Luego se forma un solo paquete de compras para poder compartir costos de traslados del producto.

Es así que se busca un modelo predictivo que tome como base un plan de compras integral y con bases científicas, con esto se pretende ayudar en una mejora sustancial de los costos operativos de las compras grupales.

1.3 Definición y Descripción del Problema

La proyección de la demanda de productos en un lapso de tiempo es uno de los temas más críticos para la organización, este proceso se realiza actualmente dentro de una herramienta de Microsoft Access creada internamente en la compañía, pero con un proceso muy engorroso y que demanda mucho tiempo, además que se realiza sobre datos ya pasados y no sobre patrones de comportamiento, es por esto que para mejorar la operación de esta oficina de compras regionales se plantea la necesidad de contar con una herramienta más ágil que responda a las necesidades de las empresas que conforman Grupo Transmerquim, y adicional se plantea la necesidad de tener estructuras matemáticas de minería de datos que puedan mostrar patrones y predicciones de los costos secundarios en las compras al iniciar con los valores de flete y seguro de las cargas.

1.4 JUSTIFICACIÓN

El tiempo empleado en la creación de estos reportes de importaciones de productos por familia, hace que el proceso, solo se pueda realizar una vez al mes ya que el mismo emplea tiempo excesivo de los encargados de los departamentos de cada una de las afiliadas que lo hacen muy laborioso para poder realizarlo diariamente. Además, en el análisis de proyecciones se presentan varias dificultades, ya que la cantidad de variables de entrada de este proceso son considerables, casi que necesitan un análisis individual de cada una, por lo que el proceso de costeo se realiza basado en supuestos que cada compañía maneja de forma individual. Bajo este panorama las compras siempre presentan desfases de demandas internas por lo que el producto que llega a puerto en algunas ocasiones es insuficiente para suplir la demanda local o en otros casos los productos cuando llegan al país respectivo presenta una disminución significativa en su demanda. A esto se le suma los costos ocultos por atrasos en la terminal por tráfico de buques o la disponibilidad o no de terminales de descarga del producto. El pre-costeo en puerto permitirá colocar el producto listo para la venta mucho más rápido y permitirá ser más competitivos en los diferentes países donde se tiene presencia.

Una solución de inteligencia de negocios junto con modelos de predicción puede ayudar a constituir la base de análisis de información de la situación actual y hacer proyecciones, primero de flete y seguro, y luego agrupar otras variables que componen el proceso de nacionalización del producto con bases más ciertas que las actuales.

1.5 VIABILIDAD

1.5.1 Punto de Vista Técnico,

- Se cuenta con las bases de datos de compras en tránsito, inventario actual y proyecciones de ventas del grupo como base del análisis y de insumo para la creación de una nueva herramienta.
- Se cuenta con archivos de información de importaciones de algunos países, para su análisis.

1.5.2 Punto de Vista Operativo

- Se cuenta con el apoyo del área de compras del grupo para poder establecer métricas y validaciones preliminares de datos.
- Se cuenta con el apoyo del área de logística para establecer los mapas conceptuales del proceso de compras.

1.5.3 Punto de Vista Económico

- Se cuenta con el apoyo económico en esta primera y segunda fase, para su posible implementación se tendría que volver a evaluar el contenido económico.
- Se cuenta con el apoyo de la alta gerencia para la búsqueda de una solución integral y la puesta en marcha si fuera el caso del concepto y de los diferentes modelos.

1.6 OBJETIVOS

1.6.1 Objetivo General

Desarrollar una nueva aplicación de análisis de la demanda de productos para las compras regionales y un modelo de predicción de la variable de flete y seguro en las compras para el pre-costeo de productos en Grupo Transmerquim al utilizar herramientas de Machine Learning.

1.6.2 Objetivos Específicos

1. Especificar los procesos de extracción, transformación y carga en la nueva herramienta de compras, dadas las fuentes disponibles.
2. Establecer el proceso de limpieza de datos.
3. Definir las variables de entrada para el análisis predictivo
4. Definir durante el proceso el mejor algoritmo analítico de predicción de filete y seguro en las compras al modelo de GTM.
5. Construir un modelo de análisis predictivo de variables de compras mediante la programación de modelos predictivos supervisados.
6. Analizar los resultados obtenidos mediante un proceso de pruebas controlado.

1.7 ALCANCES Y LIMITACIONES

1.7.1 Alcances

- Entrega de un cubo multidimensional que sustituya la actual aplicación creada en Microsoft Access y que agilice el proceso de análisis de compras.

- Definir las variables que permitan un mejor entendimiento del modelo predictivo de compras, se inicia con los costos secundarios de flete y seguro.
- Identificar y aplicar el mejor algoritmo de predicción de compras que se adapte al modelo de compras del grupo de empresas.
- Evaluar los modelos propuestos y justificar su escogencia.

1.7.2 Limitaciones

- El proyecto no incluye la implementación del modelo matemático en todas las compañías del grupo, se limita a la oficina de compras consolidadas de Houston, Texas y a 4 compañías del grupo.
- Este proceso excluye las proyecciones de ventas y se enfoca al proceso de compras.
- El proyecto se enfoca en las compras globales del grupo y en las compras locales de algunas compañías.
- El estudio se centra en el proceso de importación y distribución de los productos a granel.
- No se incluyen todas las diferentes familias de productos, ya que, por su alto costo en la obtención de la información, se enfocaron los esfuerzos en aquellas familias de productos con un alto consumo interno en los países por analizar. Las mismas se detallan a continuación,
    - Aceites.
    - Agropecuarios y Farmacéuticos.

- Agroquímicos.
- Minerales, Sales y Otros Compuestos.
- Resinas y Aditivos Auxiliares.

1.8 MARCO DE REFERENCIA ORGANIZACIONAL Y SOCIOECONÓMICO

1.8.1 Historia

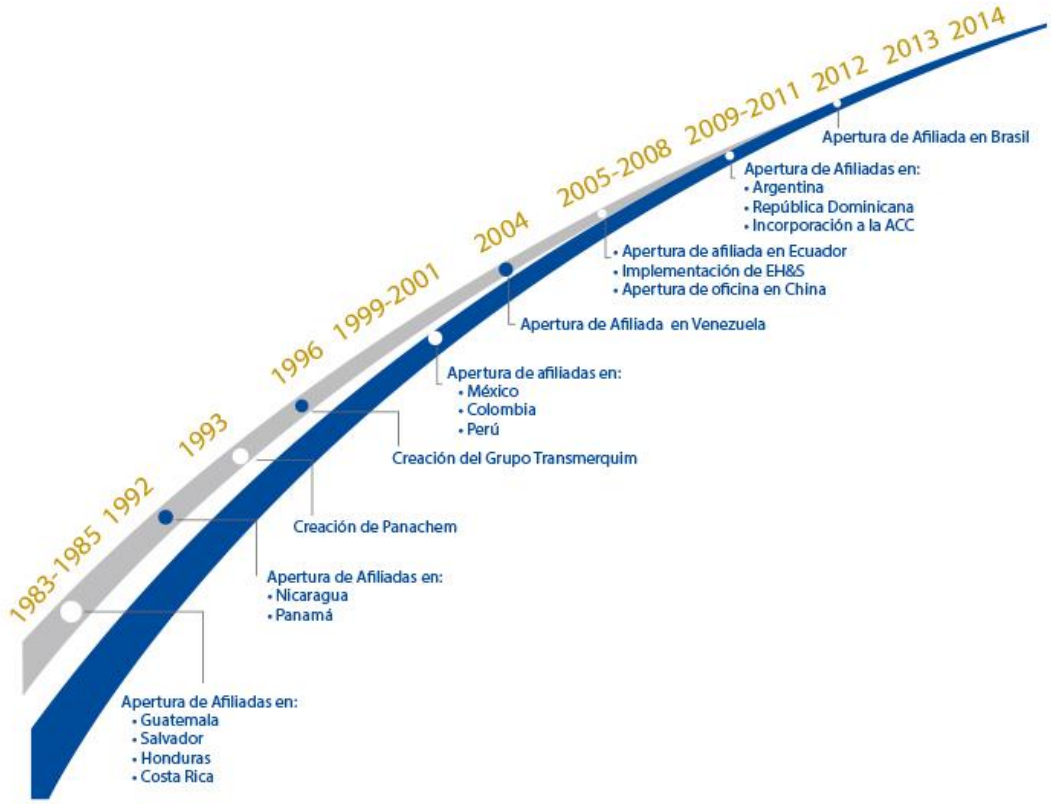


Figura 1. Evolución cronológica de Grupo Transmerquim

1.8.2 Tipo de Negocio

GTM es una empresa latinoamericana especializada en el mercadeo y distribución de químicos y materias primas para la industria en general.

1.8.3 Mercado Meta

Sectores industriales como petróleo y gas, agricultura, pinturas y cubrimientos, adhesivos, tratamiento de aguas, alimentos y cuidado personal.

1.8.4 Misión

Suministramos materias primas, soluciones y servicios innovadores, con excelencia al agregar valor a nuestros clientes y proveedores.

1.8.5 Visión

Ser un líder en los mercados de distribución de las Américas.

1.8.6 Valores Corporativos

- Excelencia
- Espíritu Emprendedor
- Integridad
- Trabajo en Equipo

1.9 ESTADO DE LA CUESTIÓN

1.9.1 Consulta de búsqueda

¿Cuáles son los algoritmos de análisis de datos más empleados en la importación de bienes y sus costos secundarios al utilizar técnicas de Machine Learning?

1.9.2 Cadena de búsqueda

(machine learning and data mining) or (machine learning and cloud computing) or (data mining and cloud computing) or (machine learning and supply chain management)

1.9.3 Fuentes Consultadas

- Springer Link 
- Coursera 
- Google Scholar 
- ELSEVIER 
- IEEE 

1.9.4 Resultados de búsqueda

El idioma de las fuentes es Inglés, al ser el más utilizado en este tipo de documentos,

- 4920 resultados iniciales
- Se ajustó y se discriminó los artículos tomando en cuenta,
  - Q1
  - Menos de 10 años de antigüedad
  - Más citados en artículos.
  - 15 artículos, en donde sobresalen algunos del Massachusetts Instititute of Technology y de Stanford University.

1.9.5 Selección de estudios

En el análisis de los estudios se definen los criterios de inclusión y de exclusión para los diferentes documentos, la siguiente tabla muestra los criterios definidos para este proceso y definición de documentación para el proyecto.

Criterio	Descripción
Criterio de Inclusión	Incluye publicaciones cuyos títulos se relacionan con la consulta de búsqueda de cadena de suministros y Machine Learning
Criterio de Inclusión	Se incluyen palabras reservadas que coincidan con las definidas en la búsqueda.
Criterio de inclusión	Incluye palabras cuyo resumen se relacione con las palabras de búsqueda.
Criterio de inclusión	Artículos más citados en otros trabajos.
Criterio de exclusión	Publicaciones de más de 10 años de antigüedad, a menos que sean muy reconocidas.
Criterios de exclusión	Publicaciones duplicadas.

Tabla 1. Criterios de inclusión y exclusión de los artículos.

Luego de aplicar los criterios de inclusión y exclusión se procede al filtrado de los mismos en las búsquedas, como resultado se obtiene la siguiente lista de artículos como base para el análisis del proyecto.

Fuente	Creadores	Nombre de la Publicación	Año Publicación
Elsevier	Gye-hang Hong, Sung ho Ha	Evaluating Spply Partner Capability for seasonal products Using Machine Learning Techniques	2007
Elsevier	Selwyn Piramunthu	Machine Learning for Dynamic Multi-Product Supply Chain Formation	2008
Springer Link	Manas Gaur, Shruti Goel, Eshaan Jain	Comparison between Nearest Neighbours and Bayesian network for Demand Forecasting in supply Chain Management	2015
IEEE	Malek Sahani, Abdellatif El Afia	Intelligent System Based Support Vector Regresion for Suppy Chain Demant Forecasting	2014
Springer Link	Hu Guosheng, Zhang Guohong	Comparison on Neural Networks and Support Vector Machines in Suppliers Selection	2006
Elsevier	Hoy-Ming Chi, Okan K. Ersoy, Herbert Moskowitz, Jim Wan	Modeling and Optimizing a Vendor Managed Replenishment System Using Machine Learning and Genetic Algorithms	2006

Springer Link	Abdallah Bashir Musa	Comparative Study on Classification Performance Between Support Vector Machine and Logistic Regression	2012
Elsevier	Real Carbonneau, Kevin Laframboise, Rustam Vahidov	Application of Machine Learning Techniques for Supply Chain Demand Forecasting	2006
Elsevier	J. Sudhir Ryan Daniel, Chandrasekharan Rajendran	A simulation- based genetic algorithm for Inventory Optimization in Serial Supply Chain.	2008
IEEE	Shoaib Bakhtyar, Lawrence Henesey	Freigth transport Prediction using Electronic Waybillis and Machine Learning	2014
Elsevier	Olga Fink, Enrico Zio, Ulrich Weidmann	Predicting Time Series of Railway Speed Restrictions with time-dependent machine Learning Techniques	2013
Google Scholar	Gragan gasevic, Carolyn Rose, George Siemens, Annika Wolff	Learning Analytics and Machine Learninng	2014
Coursera	Andrew Ng	Machine Learning, Stanford University	2015
Coursera	University Washington	Machine Learning Specialization	2016

Tabla 2. Cuadro resultante de artículos consultados para el desarrollo del proyecto.

1.9.6 Extracción de resultados

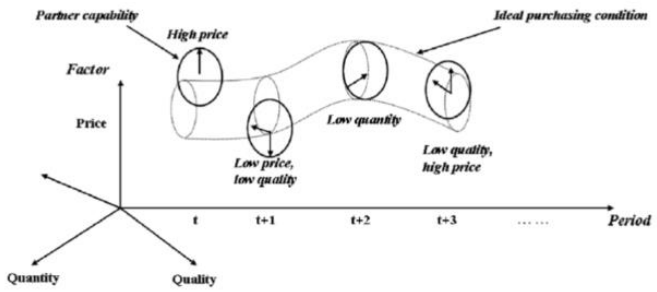
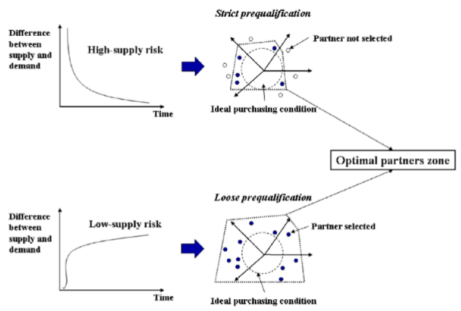
Identidicacion por Articulo
Titulo: Evaluating Supply Partner Capability for seasonal products Using Machine Learning Techniques
Año de Publicacion: 2007
Autores: Gye-hang Hong, Sung ho Ha
Referencias: 27
Descripcion del Articulo
Area: Machine Learning
Resumen: Tecnicas de Machine Learning aplicadas a un grupo de datos para la evaluacion de un grupo de proveedores.
Aspectos a destacar
*Zonas optimas de entrega de productos por parte del proveedor. *Tiempos de entrega de productos optimos.    Fig. 2. Changes in a partner's capability over a period of time.    Fig. 3. Selecting partners under different conditions of supply risk.

Tabla 3. Ficha de Artículo 1

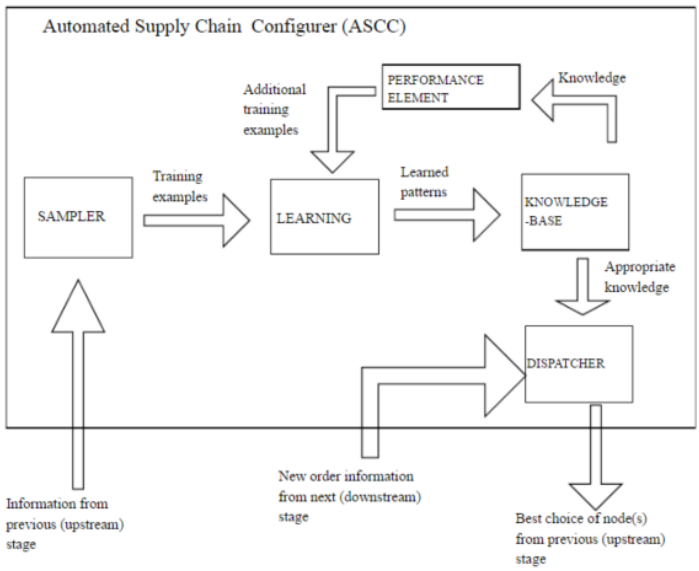
Identidicacion por Articulo
Titulo: Machine Learning for Dynamic Multi-Product Supply Chain Formation
Año de Publicacion: 2008
Autores: Selwyn Piramunthu
Referencias: 17
Descripcion del Articulo
Area: Machine Learning y cadena de suministros.
Resumen: Investigacion del proceso de produccion y su eficiente conexión con la cadena de suministros..
Aspectos a destacar
*Marco de trabajo en el trabajo de una cadena de suministros. *Incorporar tecnicas de Machine Learning. *Configuracion automatica del modelo de trabajo en la cadena de suministros. *S. Piramuthu / Expert Systems with Applications 29 (2005) 985–990*      Fig. 1. Automated Supply Chain Configurer (ASCC) Framework.

Tabla 4. Ficha Artículo 2

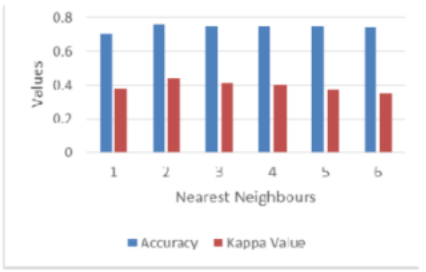
Identidicacion por Articulo
Titulo: Comparison between Nearest Neighbours and Bayesian network for Demand Forecasting in supply Chain Management
Año de Publicacion: 2015
Autores: Manas Gaur, Shruti Goel, Eshaan Jain
Referencias: 7
Descripcion del Articulo
Area: Machine Learning, algoritmos supervisados y cadena de suministros.
Resumen: Compracion entre metodos supervisados de regresion usando la matrix de confusion con metodo de mejora en las salidas de los algoritmos.
Aspectos a destacar
*Sobre alimentacion de variables. *Induccion de variables en problemas complejos. *Presicion de las variables con y sin tecnicas de Adaptive Boosting.    Fig. 2 Comparison of Accuracy and Kappa values for different neighbors

Tabla 5. Ficha Artículo 3

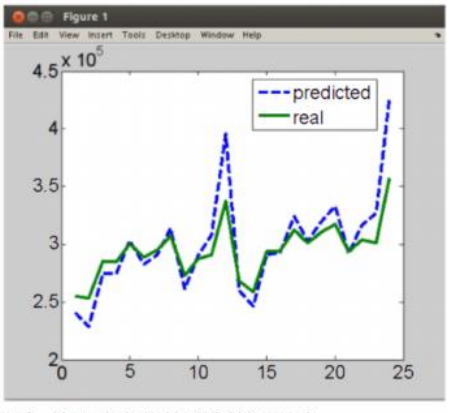
Identidicacion por Articulo
Titulo: Intelligent System Based Support Vector Regression for Supply Chain Demant Forecasting
Año de Publicacion: 2014
Autores: Malek Sahani, Abdellatif El Afia
Referencias: 10
Descripcion del Articulo
Area: Machine Learning, algoritmos no supervisados y cadena de suministros.
Resumen: Introducción a la aplicación SVR en series de tiempo en la prediccion de demanda de suministros con la incorporacion de PSO para su optimizacion.
Aspectos a destacar
*Prediccion en la demanda de artículos. *Regresion en vectores de soporte para la prediccion de la demanda. *Comparacion de SVM y SVR. *Utlizar "Particle Swarm Optimization" como mejora en la prediccion de resultados de regresión.    Fig. 2. The result obtained by SVR-PSO approach

Tabla 6. Ficha Artículo 4

Identidicacion por Articulo
Titulo: Comparison on Neural Networks and Support Vector Machines in Suppliers Selection
Año de Publicacion: 2008
Autores: Hu Guosheng, Zhang Guohong
Referencias: 7
Descripcion del Articulo
Area: Machine Learning, suport vector machine, logistica y supply chain management
Resumen: Muestra como las tecnicas de SVM son superiores a las tecnicas tradicionales de redes neuronales en la selección de proveedores.
Aspectos a destacar
*Propagación inversa. *Regresion en vectores de soporte para la prediccion de la demanda. *Comparacion de SVM y ANN. **Table 2 Comparison on SVM and ANN** S_1 S_2 S_3 S_4 S_5 ... S_{98} S_{99} S_{100} MSE Actual value 0.618 5 0.836 7 0.752 3 0.705 6 0.743 4 0.633 3 0.400 6 0.658 6 SVM 0.593 2 0.835 5 0.729 5 0.719 1 0.730 6 0.639 8 0.411 1 0.637 6 Errors/% 2.53 0.12 2.28 −1.35 1.28 ... −0.65 −1.05 2.10 5.29 BPNN 0.624 3 0.810 5 0.702 6 0.665 6 0.767 5 0.606 9 0.463 7 0.676 2 Errors/% −0.58 2.62 4.97 4.00 −2.41 2.64 −6.31 −1.76 20.88

Tabla 7. Ficha Artículo 5.

Identidicacion por Articulo
Titulo: Modeling and Optimizing a Vendor Managed Replenishment System Using Machine Learning and Genetic Algorithms
Año de Publicacion: 2006
Autores: Hoy-Ming Chi, Okan K. Ersoy, Herbert Moskowitz, Jim Wan
Referencias: 24
Descripcion del Articulo
Area: Machine Learning, algoritmos geneticos, suport vector machine y supply chain management.
Resumen: Explora el uso de algoritmos geneticos en la comprension de las redes de suministros y se aplica en los sistemas de remplazo de articulos para una salida que permite ver la solida aplicación de este tipo de algoritmos en la prediccion de reposicion de articulos.
Aspectos a destacar
*Aplicacion de tecnicas de Machine Learning en crear un modelo y la optimizacion del aanalisis en la cadena de suministros. *Suport Vector Machine en las regresiones de variables lineales. *Utilizacion del metodo random en SVM.

Table 3
Initial and final GA settings used in experiment

Parameter	Initial GA settings	Final GA settings	
		Single criterion	Bi-criteria
Parent population	20	15	50
Parent/offspring ratio	1:7	1:7	1:15
Selection type	Stochastic universal sampling	Stochastic universal sampling	Stochastic universal sampling
Crossover type	Single-point	Single-point	Single-point
Crossover rate	0.85	0.85	0.85
Mutation rate	0.125	0.125	0.125

Tabla 8. Ficha Artículo 6

Identidicacion por Articulo
Titulo: Comparative Study on Classification Performance Between Support Vector Machine and Logistic Regression
Año de Publicacion: 2012
Autores: Abdallah Bashir Musa
Referencias: 12
Descripcion del Articulo
Area: Machine Learning, suport vector machine y analisis estadistico.
Resumen: Realiza una compracion estadistica de los resultados de Suport Vector machine y Logistic Regression, incluyendo variables estadisticas que llevan a un nivel superior esta comparacion de algoritmos.
Aspectos a destacar
*Balanceo de variables. *Categorizacion de variables. *Analisis de variables con modelos estadisticos.

Table 2 The results of the performance measures for SVM

Data set	The performance measures									
	Accuracy	Sensitivity	Specificity	F-score	Precision	AUC	γ	$\rho+$	$\rho-$	DRP
Cod (sample)	0.861	0.944	0.777	0.872	0.811	0.921	0.72	4.221	0.073	58.061
Chest	0.997	0.997	0.995	0.997	0.996	1.000	0.993	217.62	0.002	108,810
Contra	0.731	0.844	0.579	0.782	0.729	0.783	0.422	2.000	0. 270	7.4074
Credit	0.875	0.888	0.865	0.866	0.845	0.946	0.753	6.569	0.129	50.9226
Liver	0.768	0.662	0.845	0.706	0.755	0.828	0.507	4.270	0.399	10.7018
Heard	0.900	0.946	0.842	0.913	0.882	0.943	0.788	5.979	0.063	94.9049
Page	0.960	0.977	0.810	0.978	0.978	0.983	0.788	5.163	0.028	184.393
Spam	0.927	0.882	0.957	0.905	0.930	0.978	0.839	20.322	0.123	165.220
Car	0.990	0.996	0.984	0.986	0.989	0.999	0.980	297.77	0.016	18,610.6
Breast	0.980	0.980	0.978	0.971	0.959	0.995	0.960	43.656	0.017	2,568.00
Diabetes	0.839	0.668	0.932	0.744	0.840	0.910	0.599	9.822	0.356	27.5899
Number	0.871	0.670	0.957	0.757	0.870	0.914	0.627	15.630	0.345	45.3044
Ionosphere	0.966	0.964	0.968	0.973	0.982	0.994	0.932	30.380	0.037	821.081

Table 3 The results of the performance measures for LR

Data set	The performance measures									
	Accuracy	Sensitivity	Specificity	F-score	Precision	AUC	γ	$\rho+$	$\rho-$	DRP
Cod (sample)	0.953	0.962	0.944	0.954	0.945	0.991	0.906	17.037	0.040	425.925
Chest	0.981	0.981	0.981	0.982	0.983	0.997	0.962	51.646	0.020	2,582.3
Contra	0.726	0.826	0.590	0.776	0.731	0.771	0.419	2.029	0.294	6.901
Credit	0.883	0.922	0.851	0.877	0.837	0.945	0.772	6.176	0.092	67.130
Liver	0.760	0.676	0.825	0.710	0.737	0.822	0.501	3.860	0.393	9.822
Heard	0.900	0.933	0.858	0.912	0.891	0.951	0.791	6.588	0.077	85.558
Page	0.962	0.992	0.696	0.979	0.966	0.975	0.689	3.269	0.011	297.182
Spam	0.932	0.894	0.957	0.912	0.931	0.979	0.851	20.772	0.111	187.135
Car	0.960	0.969	0.935	0.920	0.910	0.991	0.900	30.573	0.067	456.313
Breast	0.974	0.966	0.978	0.963	0.958	0.997	0.944	42.913	0.034	1,262.147
Diabetes	0.789	0.610	0.886	0.668	0.741	0.858	0.494	5. 335	0.442	11.312
Number	0.793	0.543	0.903	0.610	0.703	0.827	0.439	5.520	0.513	10.760
Ionosphere	0.952	0.973	0.913	0.963	0.952	0.991	0.886	11.149	0.029	384.448

Tabla 9. Ficha Artículo 7

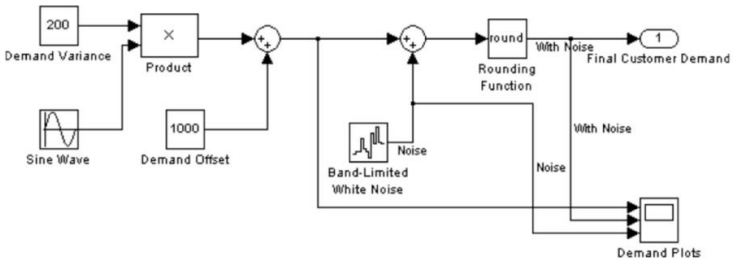
Identidicacion por Articulo
Titulo: Application of Machine Learning Techniques for Supply Chain Demand Forecasting
Año de Publicacion: 2007
Autores: Real Carbonneau, Kevin Laframboise, Rustam Vahidov
Referencias: 30
Descripcion del Articulo
Area: Machine Learning, suport vector machine y supply chain management
Resumen: Realiza una compracion de diferentes tecnicas de prediccion en la demanda de suministros, incluyendo redes neuronales o suport vector machine, con metodos tradicionales como regresion lineal multiple.
Aspectos a destacar
*Aplicacion de algoritmos de SVM en el calculo de la demanda. *Aplicacion de redes neuronales en el calculo de la demanda. *Aplicacion de redes neuronales recurrentes en el calculo de la demanda. *Ruido en lo datos.    Fig. 3. End-customer demand simulation.

Tabla 10. Ficha Artículo 8

Identidicacion por Articulo
Titulo: A simulation- based genetic algorithm for Inventory Optimization in Serial Supply Chain.
Año de Publicacion: 2008
Autores: J. Sudhir Ryan Daniel, Chandrasekharan Rajendran
Referencias: 22
Descripcion del Articulo
Area: Inventarios, algoritmos geneticos y supply chain management
Resumen: Realiza un estudio sobre la administracion del inventario y como se afecta la cadena de suministros, utilizando algoritmos geneticos. Con el proposito de mejorar los niveles de inventario con el objetivo de minimizar grandes acntidades de inventario o escases de inventario.
Aspectos a destacar
*Optimizacion de niveles de invetario utilizando algoritmos geneticos. *Comparaciones aleatorias de variables y de articulos. *Inclusión de algoritmos geneticos en el analisis del invenatario disponible.

Table 4
Results of evaluation of different solution methodologies for different supply chain settings

Supply chain setting	Solution methodology	Base-stock levels {S, M, D, R}	Mean TSCC (over 30 repl.)	SD TSCC (over 30 repl.)	Mean THC	THC as % of TSCC	Mean TSHC	TSHC as % of TSCC
A1	COMP.ENUM.	{183, 230, 143, 52}	425,340	13,316	326,442	76.75	98,897	23.25
	GA	{183, 230, 142, 52}	425,400	13,512	321,609	75.60	103,791	24.40
	RSP-1	{130, 282, 138, 44}	578,471*	15,166	282,502	48.83	295,969	51.17
	RSP-2	{198, 239, 148, 56}	461,104*	14,436	426,766	92.55	34,338	7.45
	RSP-3	{198, 239, 148, 56}	461,104*	14,436	426,766	92.55	34,338	7.45
	RSP-4	{198, 239, 148, 56}	461,104*	14,436	426,766	92.55	34,338	7.45
	RSP-5	{182, 236, 132, 50}	453,722*	16,026	275,305	60.68	178,417	39.32
A2	COMP.ENUM.	{179, 229, 144, 54}	634,072	21,369	489,660	77.22	144,411	22.78
	GA	{179, 230, 143, 54}	634,140	21,405	487,805	76.92	146,335	23.08
	RSP-1	{130, 282, 138, 44}	871,678*	22,138	456,931	52.42	414,747	47.58
	RSP-2	{198, 239, 148, 56}	690,919*	22,675	636,007	92.05	54,912	7.95
	RSP-3	{198, 239, 148, 56}	690,919*	22,675	636,007	92.05	54,912	7.95
	RSP-4	{198, 239, 148, 56}	690,919*	22,675	636,007	92.05	54,912	7.95
	RSP-5	{198, 239, 148, 56}	690,919*	22,675	636,008	92.05	54,911	7.95
A3	COMP.ENUM.	{180, 223, 138, 50}	364,930	12,194	259,650	71.15	105,279	28.85
	GA	{182, 221, 139, 50}	365,082	12,174	259,986	71.21	105,096	28.79
	RSP-1	{179, 212, 162, 52}	487,267*	10,955	322,585	66.20	164,682	33.80
	RSP-2	{179, 212, 162, 52}	415,825*	10,802	333,339	80.16	82,486	19.84
	RSP-3	{179, 212, 162, 52}	415,825*	10,802	333,339	80.16	82,486	19.84
	RSP-4	{179, 212, 162, 52}	415,825*	10,802	333,339	80.16	82,486	19.84
	RSP-5	{182, 236, 132, 50}	381,657*	11,168	284,466	74.53	97,191	25.47
B1	COMP.ENUM.	{182, 273, 142, 95}	493,501	16,069	363,776	73.71	129,725	26.29
	GA	{182, 273, 142, 94}	493,508	16,197	356,578	72.25	136,930	27.75
	RSP-1	{152, 276, 160, 92}	566,217*	13,968	380,110	67.13	186,107	32.87
	RSP-2	{152, 276, 160, 92}	566,217*	13,968	380,110	67.13	186,107	32.87
	RSP-3	{152, 276, 160, 92}	566,217*	13,968	380,110	67.13	186,107	32.87
	RSP-4	{152, 276, 160, 92}	566,217*	13,968	380,110	67.13	186,107	32.87
	RSP-5	{200, 290, 133, 96}	530,789*	14,365	395,201	74.46	135,588	25.54
B2	COMP.ENUM.	{179, 272, 143, 98}	719,423	24,660	542,048	75.34	177,375	24.66
	GA	{178, 273, 142, 97}	719,795	25,004	529,480	73.56	190,315	26.44
	RSP-1	{152, 276, 160, 92}	819,778*	22,413	540,468	65.93	279,310	34.07
	RSP-2	{152, 276, 160, 92}	819,778*	22,413	540,468	65.93	279,310	34.07
	RSP-3	{152, 276, 160, 92}	819,778*	22,413	540,468	65.93	279,310	34.07
	RSP-4	{152, 276, 160, 92}	819,778*	22,413	540,468	65.93	279,310	34.07
	RSP-5	{200, 290, 133, 96}	807,702*	21,227	603,972	74.78	203,730	25.22

Tabla 11. Ficha Artículo 9

Identidicacion por Articulo
Titulo:Machine Learning, Stanford University
Año de Publicacion: 2015
Autores: Andrew Ng
Referencias: NA
Descripcion del Articulo
Area:Machine Learning, algoritmos supervisados y no supervisados
Resumen: Curso en donde se abacan los principios de Machine Learning, definiciones y clasificaciones dealgoritmos como supervisados y no supervisados.
Aspectos a destacar
*Aplicación de algoritmos supervisados en regresión y clasificación. *Aplicación de algoritmos no supervisados. *IInclusión de analisis estadistico. *Limpieza de datos. *Normalización de datos. *Comparacion de resultados de algoritmos de regresión.

Tabla 12. Ficha Artículo 10

2 CAPÍTULO 2. MARCO TEÓRICO

En los últimos años ha habido un creciente interés en el tratamiento y análisis de datos con el propósito de descubrir relaciones, patrones y conocimiento oculto en los mismos. Las técnicas de minería de datos o también llamadas de descubrimiento de conocimiento se han aplicado consistentemente a un gran espectro de áreas como el marketing, inversiones, detección de fraude, producción industrial, telecomunicaciones, salud, etc. Conocimiento previamente desconocido y que se encuentra en la base de datos ya sea local o en la nube al utilizar diversas áreas del conocimiento como lo es la Inteligencia artificial o la estadística.

Como consecuencia de esta búsqueda de mejora en los procesos, es que se lleva a cabo en esta investigación, con la que se espera abarcar aspectos de investigación relacionados con la importación de productos petroquímicos y su eficiente distribución.

Para resolver un problema complejo de análisis de datos como este, se puede ensayar con distintos métodos y ver con cuál de ellos se encuentra la mejor solución. En algunos casos la mejor solución no la proporciona un solo método de análisis, es por esto que el estudio se presenta con diferentes tipos de algoritmos, que proporcionan métodos que nos ayudan a entender los patrones ocultos dentro de la información proporcionada.

2.1 MACHINE LEARNING

Arthur Lee Samuel pionero de los juegos de computadoras e inteligencia artificial definió el termino de Machine Learning como: "el campo de estudio que permite a las computadoras aprender sin ser explícitamente programadas". Muy usada como definición aceptada, es ahora un poco desfasada para la actualidad.

Tom Mitchell proporciona una definición más moderna: " es un programa informático que aprende de la experiencia E con respecto a cierta clase de tareas T y mide su rendimiento P, si su rendimiento en tareas en T, según lo medido por P, mejora con la experiencia E”

Ejemplo: jugar damas chinas.

E: La experiencia de jugar damas chinas.

T: la tarea de jugar damas chinas.

P: probabilidad de que el jugador que gane, vaya a ganar el próximo juego de damas chinas.

El aprendizaje automático tiene una amplia gama de aplicaciones, incluye motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica.

2.2 MODELOS

El aprendizaje automático tiene como resultado un modelo para resolver una tarea dada. Entre los modelos se distinguen:

- Los modelos geométricos, construidos en el espacio de instancias y que pueden tener una, dos o múltiples dimensiones. Si hay un borde de decisión linear entre las clases, se dice que los datos son linealmente separables. Un límite de decisión linear se define como t = w * x, donde w es un vector perpendicular al límite de decisión, x es un punto arbitrario en el límite de decisión y t es el umbral de la decisión. (Aplicada, 2010).

- Los modelos probabilísticos, que intentan determinar la distribución de probabilidades descriptora de la función que enlaza a los valores de las características con valores determinados. Uno de los conceptos claves para desarrollar modelos probabilísticos es la estadística bayesiana. (Aplicada, 2010).

- Los modelos lógicos, que expresan las probabilidades en reglas organizadas en forma de árboles de decisión.

Los modelos pueden también clasificarse como modelos de agrupamiento y modelos de gradiente. Los primeros tratan de dividir el espacio de instancias en grupos. Los segundos, como su nombre lo indican, representan un gradiente en el que se puede diferenciar entre cada instancia. Clasificadores geométricos como las máquinas de vectores de apoyo son modelos de gradientes. (S. Kotsiantis, 2007)

Un atributo o característica es un tipo de medida realizada sobre cualquier instancia por medir. Los atributos mapean el espacio de instancias a un conjunto de valores o dominio de atributos. Los valores del dominio pueden ser números como la frecuencia de aparición de las instancias, valores binarios o un conjunto cualquiera como el de meses, estaciones o colores.

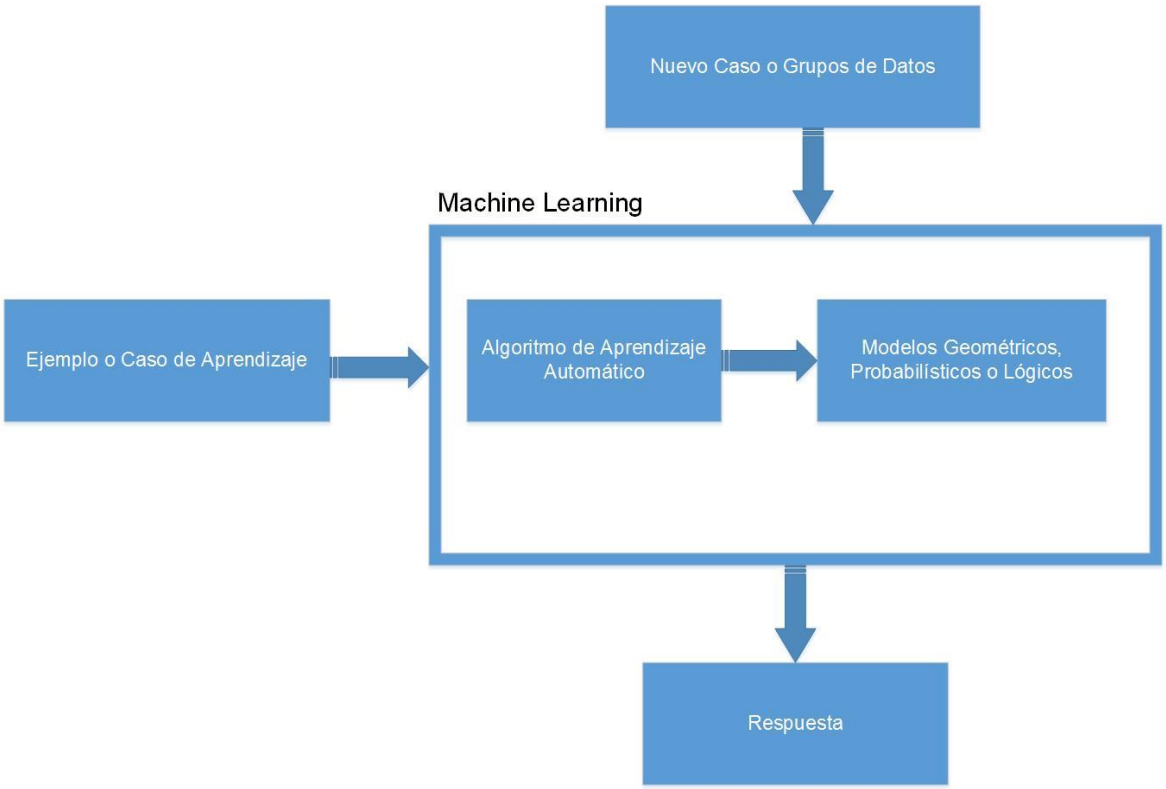


Figura 2. Modelo de aprendizaje supervisado

Como lo muestra la imagen Machine Learning se alimenta de casos de aprendizaje antes de procesar nuevos grupos y poder dar una respuesta predictiva.

2.3 APRENDIZAJE SUPERVISADO

En aprendizaje supervisado, se entrega un conjunto de datos y ya sabe lo que debería ser la salida correcta, con la idea de que existe una relación entre la entrada y la salida.

Los problemas de aprendizaje supervisado se tratan de categorizar en problemas de "regresión" y "clasificación". En un problema de regresión, se intenta predecir resultados dentro de una salida continua, lo que significa que se trata las variables de entrada y se asignan a una función continua. En un

problema de clasificación, se intenta predecir resultados en una salida discreta. En otras palabras, se está intentando asignar variables de entrada en categorías discretas. (S. Kotsiantis, 2007)

Ejemplo:

Teniendo en cuenta datos sobre el tamaño de las casas en el mercado de bienes raíces, tratar de predecir su precio. Precio en función del tamaño es una salida de continua, así que esto es un problema de regresión.

Se podría convertir este ejemplo en un problema de clasificación al hacer en cambio nuestra salida sobre si la casa "se vende por más o menos que el precio de venta." Aquí se está clasificando las casas basadas en el precio en dos categorías discretas.

2.4 APRENDIZAJE NO SUPERVISADO

El aprendizaje no supervisado, por el contrario, nos permite abordar los problemas con poca o ninguna idea de lo que nuestros resultados deben aparecer. Se puede derivar estructura de datos que no necesariamente se conoce el efecto de las variables. Se puede derivar estas estructuras en agrupamiento de datos al basarse en las relaciones entre las variables en los datos.

Con aprendizaje no supervisado no hay ninguna retroalimentación basándose en los resultados de la predicción, es decir, no hay ningún factor de corrección. Por ejemplo, la memoria asociativa. (S. Kotsiantis, 2007)

Ejemplo:

Agrupamiento. Tomar una colección de 1000 ensayos escritos sobre la economía y encontrar una manera de agrupar automáticamente estos ensayos en un pequeño número que de alguna manera son similares o relacionados por diferentes variables, como la palabra frecuencia, longitud de la frase, número de páginas y así sucesivamente.

Asociativa. El ejemplo supone que un médico con años de experiencia, hace asociaciones en su mente entre características del paciente y enfermedades que tienen. Si aparece un nuevo paciente entonces basada en las características de este paciente como síntomas, antecedentes médicos familiares, atributos físicos, mentales, etc. el médico puede asociar posibles enfermedades o enfermedades basadas en lo que el médico ha visto con pacientes similares. Esto no es lo mismo para todos o como norma base de razonamiento en sistemas expertos. En este caso se estima una función de mapeo de características en los pacientes con diferentes enfermedades. (S. Kotsiantis, 2007)

2.5 TIPOS DE ALGORITMOS

2.5.1 Árboles de decisiones

Este tipo de aprendizaje usa un árbol de decisiones como modelo predictivo. Se mapean observaciones sobre un objeto con conclusiones sobre el valor final de dicho objeto.

Los árboles son estructuras básicas en la informática. Los árboles de atributos son la base de las decisiones. Una de las dos formas principales de árboles de decisiones es la desarrollada por Quinlan de medir la impureza de la

entropía en cada rama, algo que primero desarrolló en el algoritmo ID3 y luego en el C4.5. Otra de las estrategias se basa en el índice GINI y fue desarrollada por Breiman, Friedman et alia. El algoritmo de CART es una implementación de esta estrategia. (Operaciones, 2016)

Un árbol de decisión indica las acciones por realizar en función del valor de una o varias variables. Es una representación en forma de árbol cuyas ramas se bifurcan en función de los valores tomados por las variables y que terminan en una acción concreta. Se suele utilizar cuando el número de condiciones no es muy grande, en tal caso, es mejor utilizar una tabla de decisión.

Un ejemplo al utilizar el algoritmo ID3

Ej.	Cielo	Temperatura	Humedad	Viento	Jugar Futbol
D1	Sol	Alta	Alta	Débil	-
D2	Sol	Alta	Alta	Fuerte	-
D3	Nubes	Alta	Alta	Débil	+
D4	Lluvia	Suave	Alta	Débil	+
D5	Lluvia	Baja	Normal	Débil	+
D6	Lluvia	Baja	Normal	Fuerte	-
D7	Nubes	Baja	Normal	Fuerte	+
D8	Sol	Suave	Alta	Débil	-
D9	Sol	Baja	Normal	Débil	+
D10	Lluvia	Suave	Normal	Débil	+
D11	Sol	Suave	Normal	Fuerte	+
D12	Nubes	Suave	Alta	Fuerte	+
D13	Nubes	Alta	Normal	Débil	+
D14	Lluvia	Suave	Alta	Fuerte	-

Tabla 13 Ejemplo de algoritmo ID3

En la imagen se muestra como teniendo ciertas variables y sus combinaciones se puede predecir si se podrá o no jugar futbol.

El árbol de decisión quedaría de la siguiente forma,

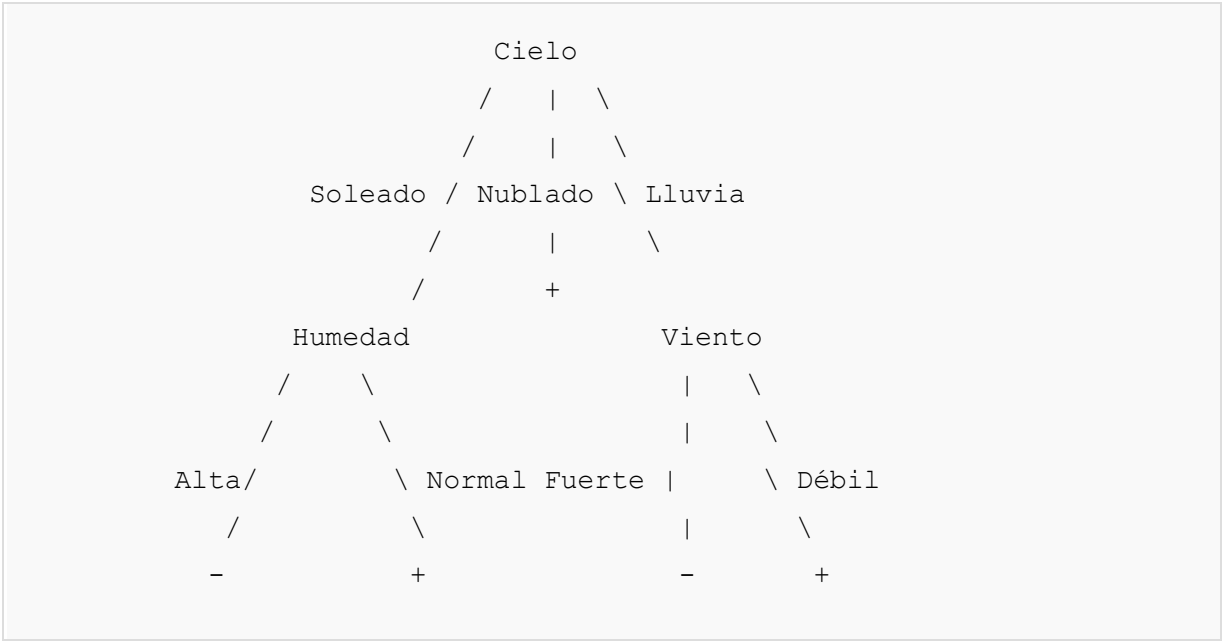


Figura 3. Árbol de decisión resultante del ejemplo.

2.5.2 Reglas de asociación

Los algoritmos de reglas de asociación procuran descubrir relaciones interesantes entre variables. Entre los métodos más conocidos se hallan el algoritmo a priori, el algoritmo Eclat y el algoritmo de Patrón Frecuente. (IBM, 2013)

Se han investigado ampliamente diversos métodos para aprendizaje de reglas de asociación que han resultado ser muy interesantes para descubrir relaciones entre variables en grandes conjuntos de datos.

Un ejemplo de reglas de asociación en ventas de un supermercado,

{cebollas, vegetales} $\Longrightarrow$ {carne}

En los datos de ventas de un supermercado, esta información indicaría que un consumidor que compra cebollas y verdura a la vez, es probable que compre también carne. Esta información se puede utilizar como base para tomar

decisiones sobre marketing como precios promocionales para ciertos productos o dónde ubicar éstos dentro del supermercado. Además del ejemplo anterior, las reglas de asociación también son de aplicación en otras muchas áreas como la minería de datos en la Web, la detección de intrusos o la bioinformática. (Reglas de Asociación, 2015)

Al ampliar el conjunto de datos de ventas en un supermercado,

C = {Leche, Pan, Mantequilla, Cerveza}

En la siguiente tabla se muestra la distribución de compras, en donde un 1 significa la existencia del producto en el escenario, lo contrario sería un 0 que indica la no existencia en el mismo.

ID	Leche	Pan	Mantequilla	Cerveza
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Tabla 14. Ejemplo de reglas de asociación

Basado en la tabla anterior se podría decir la siguiente regla de asociación,

{Leche, Pan} $\Longrightarrow$ {Mantequilla}

Significaría que, si el cliente compró leche y pan, también compró mantequilla.

Se nota que el ejemplo anterior es muy pequeño, en la práctica, una regla necesita un soporte de varios cientos de registros o transacciones antes de que ésta pueda considerarse significativa desde un punto de vista estadístico. A menudo las bases de datos contienen miles o incluso millones de registros. Para

seleccionar reglas de asociación del conjunto de todas las reglas posibles que se pueden derivar de un conjunto de datos se pueden utilizar restricciones sobre diversas medidas. Las restricciones más conocidas son los umbrales mínimos de soporte y confianza.

El soporte de un conjunto de ítemes en una base de datos se define como la proporción de transacciones en la base de datos que contiene dicho conjunto de ítemes.

$$sop(x) = \frac{x}{D}$$

En el ejemplo de leche y pan el soporte seria de,

$$sop(x) = \frac{2}{5} = 0.4$$

Es decir, el soporte es del 40% o sea que 2 de cada 5 transacciones.

La confianza de una regla se define como:

$$conf(X \rightarrow Y) = \frac{sop(X \cup Y}{sop(X)} = \frac{|X \cup Y|}{|X|}$$

Para el ejemplo anterior la confianza seria de,

$$\text{conf}(\{\text{Leche}, \text{Pan}\} \Rightarrow \{\text{Mantequilla}\}) = \frac{\text{sop}(\{\text{Leche}, \text{Pan}\} \cup \{\text{Mantequilla}\})}{\text{sop}(\{\text{Leche}, \text{Pan}\})} = \frac{0.2}{0.4} = 0.5$$

Este cálculo significa que el 50% de las reglas de la base de datos que contienen leche y pan también tienen mantequilla.

Las reglas de asociación deben satisfacer las especificaciones del usuario en cuanto a umbrales mínimos de soporte y confianza. Para conseguir esto, el proceso de generación de reglas de asociación se realiza en dos pasos. Primero se aplica el soporte mínimo para encontrar los conjuntos de ítemes más

frecuentes en la base de datos. En segundo lugar, se forman las reglas partiendo de estos conjuntos frecuentes de ítemes y de la restricción de confianza mínima.

2.5.3 Algoritmos genéticos

Los algoritmos genéticos son procesos de búsqueda del conocimiento que simulan la selección natural. Usan métodos tales como la mutación y el cruzamiento para generar nuevas clases que puedan ofrecer una buena solución a un problema dado.

Estos algoritmos hacen evolucionar una población de individuos sometiéndola a acciones aleatorias semejantes a las que actúan en la evolución biológica, mutaciones y recombinaciones genéticas, así como también a una selección de acuerdo con algún criterio, en función del cual se decide cuáles son los individuos más adaptados que sobreviven, y cuáles los menos aptos, que son descartados. Los algoritmos genéticos se enmarcan dentro de los algoritmos evolutivos, que incluyen también las estrategias evolutivas, la programación evolutiva y la programación genética. (Guervos, 2014)

Los algoritmos genéticos funcionan entre el conjunto de soluciones de un problema llamado fenotipo, y el conjunto de individuos de una población natural al codificar la información de cada solución en una cadena, generalmente binaria, llamada cromosoma. Los símbolos que forman la cadena son llamados los genes. Cuando la representación de los cromosomas se hace con cadenas de dígitos binarios se le conoce como genotipo. Los cromosomas evolucionan a través de iteraciones, llamadas generaciones. En cada generación, los cromosomas son evaluados al usar alguna medida de aptitud. Las siguientes

generaciones, son generadas al aplicar los operadores genéticos repetidamente, estos son los operadores de selección, cruzamiento, mutación y reemplazo. (Marrero, 2013)

Los algoritmos genéticos son de probada eficacia en caso de querer calcular funciones no derivables o de derivación muy compleja, aunque su uso es posible con cualquier función.

Un algoritmo genético puede presentar diversas variaciones, dependiendo de cómo se aplican los operadores genéticos, de cómo se realiza la selección y de cómo se decide el reemplazo de los individuos para formar la nueva población. En general, el pseudocódigo consiste de los siguientes pasos:

- Inicialización. Se genera aleatoriamente la población inicial, que está constituida por un conjunto de cromosomas los cuales representan las posibles soluciones del problema. En caso de no hacerlo aleatoriamente, es importante garantizar que, dentro de la población inicial, se tenga la diversidad estructural de estas soluciones para tener una representación de la mayor parte de la población posible o al menos evitar la convergencia prematura.
- Evaluación. A cada uno de los cromosomas de esta población se aplicará la función de aptitud para saber qué tan "buena" es la solución que se está codificando.
- Condición de término. Se deberá detener cuando se alcance la solución óptima, pero ésta generalmente se desconoce, por lo que se deben utilizar otros criterios de detención. Normalmente se usan dos criterios:

- Correr el algoritmo genético un número máximo de iteraciones o generaciones.
- Detenerlo cuando no haya cambios en la población.

Mientras no se cumpla la condición de término se hace lo siguiente:

- Selección. Después de saber la aptitud de cada cromosoma se procede a elegir los cromosomas que serán cruzados en la siguiente generación. Los cromosomas con mejor aptitud tienen mayor probabilidad de ser seleccionados.
- Recombinación o Cruzamiento. La recombinación es el principal operador genético, representa la reproducción sexual, opera sobre dos cromosomas a la vez para generar dos descendientes donde se combinan las características de ambos cromosomas padres.
- Mutación. Modifica al azar parte del cromosoma de los individuos y permite alcanzar zonas del espacio de búsqueda que no estaban cubiertas por los individuos de la población actual.
- Reemplazo. Una vez aplicados los operadores genéticos, se seleccionan los mejores individuos para conformar la población de la generación siguiente.

Algunas aplicaciones funcionales del algoritmo genético:

- Diseño automatizado, incluye investigación en diseño de materiales y diseño multiobjetivo de componentes automovilísticos, en el mejor comportamiento ante choques, ahorros de peso, mejora de aerodinámica.
- Diseño automatizado de equipamiento industrial.

- Diseño automatizado de sistemas de comercio en el sector financiero.
- Optimización de carga de contenedores.
- Diseño de sistemas de distribución de aguas.
- Diseño de topologías de circuitos impresos.
- Diseño de topologías de redes computacionales.
- Análisis de expresión de genes.
- Aprendizaje de comportamiento de robots.

2.5.4 Redes neuronales artificiales

Las redes de neuronas artificiales son un paradigma de aprendizaje automático inspirado en las neuronas de los sistemas nerviosos de los animales. Se trata de un sistema de enlaces de neuronas que colaboran entre sí para producir un estímulo de salida. Las conexiones tienen pesos numéricos que se adaptan según la experiencia. De esta manera, las redes neurales se adaptan a un impulso y son capaces de aprender. (Malaga, 2013)

Una red neuronal se compone de unidades llamadas neuronas. Cada neurona recibe una serie de entradas a través de interconexiones y emite una salida. Esta salida viene dada por tres funciones:

- Una función de propagación o función de excitación, que por lo general consiste en el sumatorio de cada entrada multiplicada por el peso de su interconexión o su valor neto. Si el peso es positivo, la conexión se denomina excitatoria; si es negativo, se denomina inhibitoria.

- Una función de activación, que modifica a la anterior. Puede no existir, es en este caso la salida la misma función de propagación.
- Una función de transferencia, que se aplica al valor devuelto por la función de activación. Se utiliza para acotar la salida de la neurona y generalmente viene dada por la interpretación que se quiera darles a dichas salidas.

Con un paradigma convencional de programación en ingeniería del software, el objetivo del programador es modelar matemáticamente el problema en cuestión y posteriormente formular una solución o programa mediante un algoritmo codificado que tenga una serie de propiedades que permitan resolver dicho problema. En contraposición, la aproximación basada en las RNA parte de un conjunto de datos de entrada suficientemente significativo y el objetivo es conseguir que la red aprenda automáticamente las propiedades deseadas. En este sentido, el diseño de la red tiene menos que ver con cuestiones como los flujos de datos y la detección de condiciones, y más que ver con cuestiones tales como la selección del modelo de red, la de las variables por incorporar y el pre-procesamiento de la información que formará el conjunto de entrenamiento. (Malaga, 2013)

2.5.5 Algoritmos de agrupamiento

El análisis por agrupamiento (clustering en Inglés) es la clasificación de observaciones en subgrupos para que las observaciones en cada grupo se asemejen entre sí según ciertos criterios.

Las técnicas de agrupamiento hacen inferencias diferentes sobre la estructura de los datos; se guían usualmente por una medida de similitud

específica y por un nivel de compactamiento interno y la separación entre los diferentes grupos al utilizar diferentes criterios.

Esos criterios son por lo general distancia o similitud. La cercanía se define en términos de una determinada función de distancia, como la euclidiana, aunque existen otras más robustas o que permiten extenderla a variables discretas. La medida más utilizada para medir la similitud entre los casos es la matriz de correlación entre los NxN casos. Sin embargo, también existen muchos algoritmos que se basan en la maximización de una propiedad estadística llamada verosimilitud. (Rousseeuw & Kaufman, 1995)

Generalmente, los vectores de un mismo grupo o clúster comparten propiedades comunes. El conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. De ahí su uso en minería de datos. Esta descripción sintética se consigue sustituyendo la descripción de todos los elementos de un grupo por la de un representante característico del mismo. En algunos contextos, como el de la minería de datos, se lo considera una técnica de aprendizaje no supervisado puesto que busca encontrar relaciones entre variables descriptivas, pero no la que guardan con respecto a una variable objetivo.

Existen dos grandes técnicas para el agrupamiento de casos:

Agrupamiento jerárquico, que puede ser aglomerativo o divisivo.

Agrupamiento no jerárquico, en los que el número de grupos se determina de antemano y las observaciones se van asignando a los grupos en función de su cercanía. Existen los métodos de k-mean y k-medioid. (Flach, 2012)

2.5.6 Redes bayesianas

Una red bayesiana es un modelo probabilístico que representa una serie de variables de azar y sus independencias condicionales a través de un grafo acíclico dirigido. Una red bayesiana puede representar, las relaciones probabilísticas entre enfermedades y síntomas. Dados ciertos síntomas, la red puede usarse para calcular las probabilidades de que ciertas enfermedades estén presentes en un organismo. Hay algoritmos eficientes que infieren y aprenden al usar este tipo de representación.

Existen algoritmos eficientes que llevan a cabo la inferencia y el aprendizaje en redes bayesianas. Las redes bayesianas que modelan secuencias de variables, por ejemplo, señales del habla o secuencias de proteínas, son llamadas redes bayesianas dinámicas. Las generalizaciones de las redes bayesianas que pueden representar y resolver problemas de decisión bajo incertidumbre son llamados diagramas de influencia. (Faltin F. & Kenett R., 2008)

2.5.6.1 Inferencia y aprendizaje

Hay tres tareas principales de inferencia para las redes bayesianas. Las cuales se detallan a continuación,

Deducción de variables no observadas

Debido a que una red bayesiana es un modelo completo de las variables y sus relaciones, se puede utilizar para responder a las consultas de probabilidad acerca de ellos. Por ejemplo, la red se puede utilizar para averiguar el conocimiento actualizado del estado de un subconjunto de variables cuando otras variables se observan. Este proceso de cálculo de la distribución posterior

de las variables dada la evidencia que se llama inferencia probabilística. La posterior da un suficiente estadístico universal para aplicaciones de detección, cuando se quiere elegir los valores para la variable de un subconjunto que minimizan alguna función de pérdida esperada, por ejemplo, la probabilidad de error de decisión. Una red bayesiana de esta manera, puede considerarse como un mecanismo para aplicar automáticamente el teorema de Bayes a problemas complejos.

Aprendizaje de Parámetros

Para especificar completamente la red bayesiana y por lo tanto representar plenamente a la distribución de probabilidad conjunta, es necesario especificar para cada nodo X la distribución de probabilidad de X condicional dado a sus padres. La distribución de X condicional dado a sus padres puede tener cualquier forma. Es común trabajar con distribuciones discretas o gaussianas ya que simplifica los cálculos. A veces sólo restricciones sobre una distribución son conocidas; uno puede entonces utilizar el principio de máxima entropía para determinar una distribución única. A menudo, estas distribuciones condicionales incluyen parámetros que son desconocidos y deben estimarse a partir de los datos, a veces al utilizar el enfoque de máxima probabilidad. La maximización directa de la probabilidad o de la probabilidad posterior, es a menudo compleja cuando hay variables no observadas. Un método clásico de este problema es el algoritmo de expectación-maximización el cual alterna los valores esperados computados de las variables condicionales no observadas a datos observados, con la maximización de la probabilidad total, suponiendo que previamente calculados los valores esperados son correctos. Bajo condiciones de regularidad leves este proceso converge en valores de probabilidad máxima

o máximo posterior, para los parámetros. Un enfoque más bayesiano es tratar a los parámetros como variables no observadas adicionales y para calcular la distribución posterior completa sobre todos los nodos condicionales de los datos observados, después, integrar los parámetros. Este enfoque puede ser costoso y llevar a modelos de grandes dimensiones, por lo que en la práctica los enfoques de ajuste de parámetros clásicos son más comunes.

Aprendizaje de Estructuras

En el caso más simple, una red bayesiana se especifica por un experto y se utiliza entonces para realizar inferencia. En otras aplicaciones, la tarea de definir la red es demasiado compleja para los seres humanos. En este caso la estructura de la red y los parámetros de las distribuciones locales debe ser aprendido de datos.

El aprendizaje automático de la estructura gráfica de una red bayesiana es un reto dentro del aprendizaje de máquina. La idea básica se remonta a un algoritmo de recuperación desarrollado por Rebane y Pearl y se basa en la distinción entre los tres tipos posibles de triplos adyacentes permitidos en un gráfico acíclico dirigido:

1. $X \rightarrow Y \rightarrow Z$
2. $X \leftarrow Y \rightarrow Z$
3. $X \rightarrow Y \leftarrow Z$

Tipo 1 y tipo 2 representan las mismas dependencias X y Z son independientes dada Y y son, por tanto, indistinguibles. Tipo 3, sin embargo, puede ser identificado de forma única, ya que X y Z son marginalmente independientes y todos los otros pares son dependientes. Así, mientras que los

esqueletos, los grafos despojados de flechas, de estos tres triplos son idénticos, la direccionalidad de las flechas es parcialmente identificable. La misma distinción se aplica cuando X y Z tienen padres comunes, excepto que uno debe condicionar primero en esos padres. Se han desarrollado algoritmos para determinar sistemáticamente el esqueleto del grafo subyacente y, a continuación, orientar todas las flechas cuya direccionalidad está dictada por las independencias condicionales observadas.

Un método alternativo de aprendizaje estructural utiliza la optimización basada en búsquedas. Se requiere una función de puntuación y una estrategia de búsqueda. Una función de puntuación común es la probabilidad posterior de la estructura dado los datos de entrenamiento. El requisito de tiempo de una búsqueda exhaustiva retornando una estructura que maximice la puntuación es súper-exponencial en el número de variables. Una estrategia de búsqueda local hace cambios incrementales destinados a mejorar la puntuación de la estructura. Un algoritmo de búsqueda global como la cadena de Markov Monte Carlo puede evitar quedar atrapado en mínimos locales. Friedman et al, habla acerca del uso de la información mutua entre las variables y encontrar una estructura que maximiza esto. Lo hacen mediante la restricción del conjunto de padres candidatos a k nodos y exhaustivamente buscando en el mismo.

Un ejemplo de este tipo de algoritmo lo se puede detallar a continuación, con la influencia de la lluvia si el rociador está activo e influencia de la lluvia y el rociador si la hierba se encuentra húmeda.

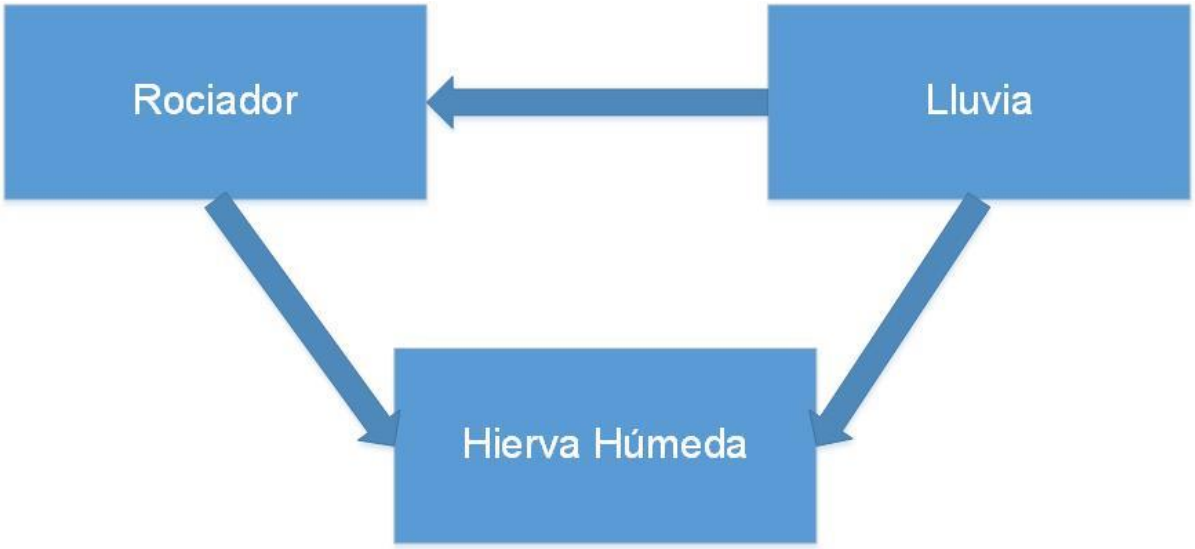


Figura 4. Ejemplo del rociador y la lluvia en las redes bayesianas

En el ejemplo se suponen que hay dos eventos, los cuales pueden causar que la hierba esté húmeda:

- Que el rociador esté activado
- Que esté lloviendo.

También se suponen que la lluvia tiene un efecto directo sobre el uso del rociador, usualmente cuando llueve el rociador se encuentra apagado. Entonces la situación puede ser modelada con una red bayesiana. Las tres variables tienen dos posibles valores, V para verdadero y F para falso. La función de probabilidad conjunta es:

$$P(H,R,L) = P(H|R,L)P(R|L)P(L)$$

Donde los nombres de las variables han sido abreviados a H = Hierba húmeda, R = Rociador activado, y L = Lloviendo.

El modelo puede responder preguntas como ¿Cuál es la probabilidad de que esté lloviendo dado que la hierba está húmeda? O ¿Cuál es la probabilidad de que llueva dado que la hierba está húmeda?

Como lo muestra la siguiente imagen se detalla ese tipo de posibilidad.

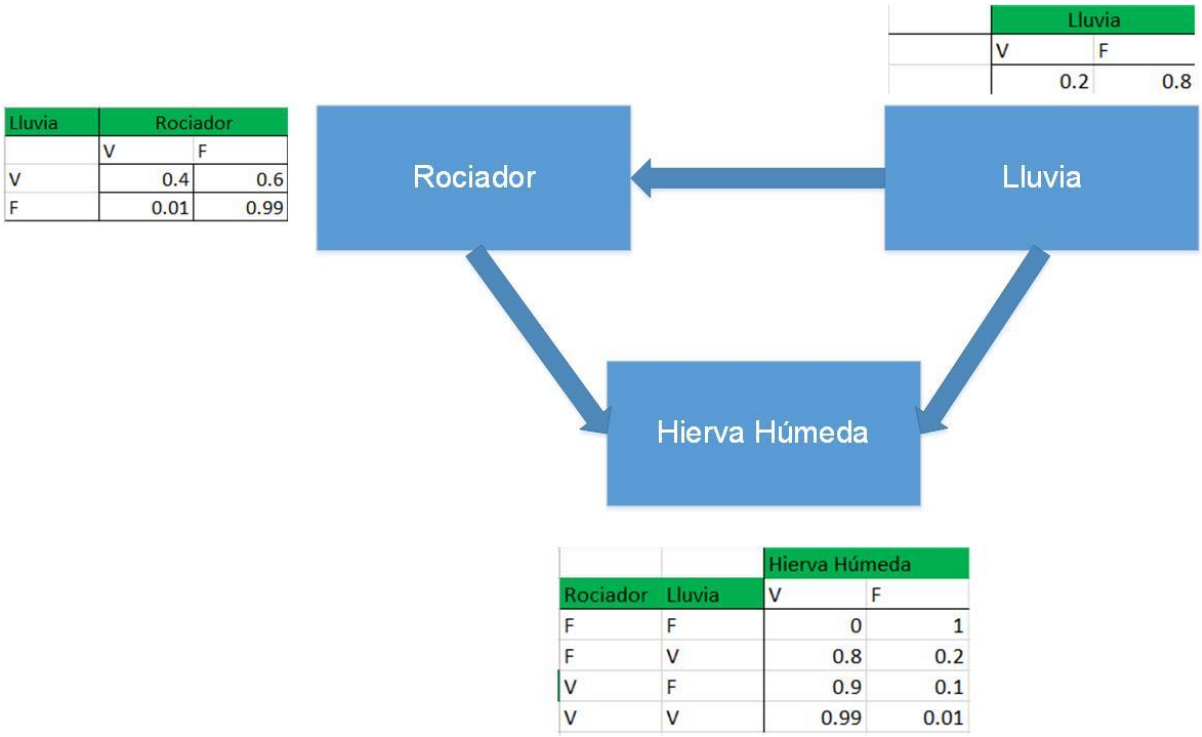


Figura 5. Muestra el resultado luego de la compilación del algoritmo.

Las redes Bayesianas se han utilizado para aplicaciones en diversas áreas, tales como aprendizaje automático, minería de texto, procesamiento del lenguaje natural, reconocimiento de voz, procesamiento de señales, bioinformática, códigos de control de errores, diagnóstico médico, pronóstico del tiempo y redes celulares.

3 CAPÍTULO 3. MARCO METODOLÓGICO

3.1 TIPO DE INVESTIGACIÓN

El tipo de investigación que se adecua al propósito del proyecto es investigación aplicada ya que su principal objetivo se basa en resolver problemas prácticos, con un margen de generalización limitado y delimitado a Grupo Transmerquim.

3.2 ALCANCE INVESTIGATIVO

Descriptiva. Describir características del proceso de importaciones y especificar características, propiedades, rasgos del fenómeno analizado.

Correlacionales. Determinar junto con el departamento de compras regionales las relaciones del descubrimiento de patrones que se mueven con base en el ingreso de ciertas variables, como es el flete y el seguro.

3.3 ENFOQUE

- Abordaje Alternativo. Así como detalla Sampieri en su libro "Metodología de la Investigación", los abordajes que incluyen los dos mundos cuantitativos y cualitativos son mucho más robustos, ya que por sí solos son una parte de un proceso, pero juntos son el todo en una investigación. En otras palabras, no se pueden desligar uno del otro ya que se encuentran intrincadamente unidos desde cualquier punto de vista metodológico.
- Epistemología. Observación. La dimensión epistemológica, nos da detalles de eficacia, eficiencia y facilidad de implementación en condiciones similares, para poder contrastar y documentar los resultados de los experimentos.
- Comparación de algoritmos y cotejar resultados.
- Eficiencia y facilidad en condiciones similares.
- Contrastar y documentar.
- Ontológica:

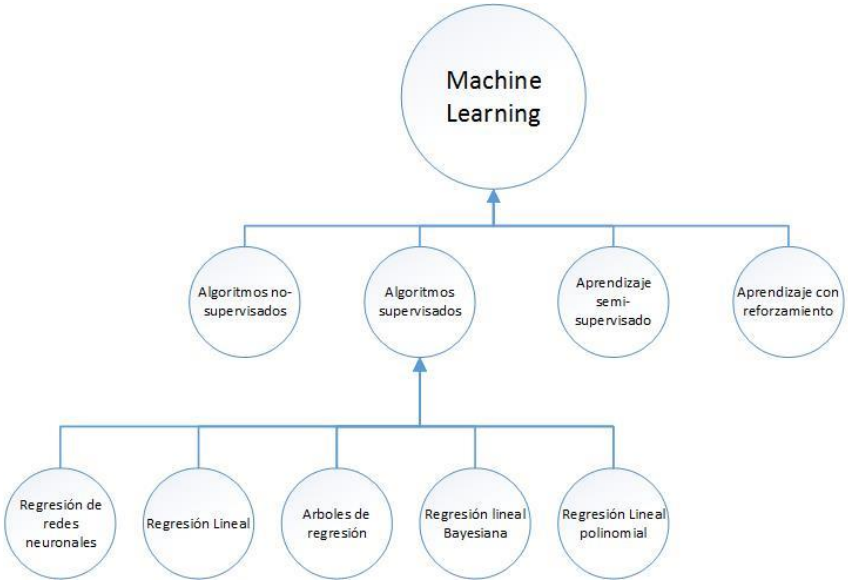


Figura 6. Ontología de algoritmos de Machine Learning basado en el modelo de implementación

Bajo este esquema propuesto se validarían los siguientes conceptos,

- Regresión de las variables de importación con esto se persigue la formación de un modelo de regresión y la predicción de un valor numérico.
- Modelo Multiclase. El objetivo es predecir los valores que pertenecen a un conjunto de valores admisibles limitado, previame nte definido.
- Número de ciclos válidos. Determinar el número de ciclos que Machine Learning debe de utilizar para que el modelo sea funcional.
- Regularización. La regularización es una técnica que se puede utilizar para obtener modelos de mayor calidad de aprendizaje.

- Axiológica.

- Exactitud y complejidad de la implementación

Metrica	Valor	Opcion uno	Opcion Dos	Opcion Tres
Coeficiente de determinación	25%	< = 0.5 10%	> 0.5 & < 0.9 20%	>= 0.9 25%
Error cuadrático relativo	25%	< 1.0 & > 0.5 10%	<= 0.5 & > 0.2 20%	< = 0.2 25%
Cantidad de líneas de código	15%	Mas de 200 líneas: 5%	De 200 a 100 líneas: 10%	Menos de 100 líneas: 15%
Cantidad de variables de entrada	35%	Mas de 10 variables: 10%	De 5 a 10 variables: 25%	Menos de 4 variables: 35%

Tabla 15. Parámetros de calificación para aplicar a los diferentes algoritmos.

3.4 Diseño Mixto

Se decanta por la creación de un diseño mixto en el proyecto, ya que con esto podemos combinar varios métodos, aprovechar sus fortalezas y minimizar sus debilidades. Con la combinación de los procesos cuantitativos y cualitativos nos aseguramos de los valiosos aportes para esta investigación ya que ambos presentan visiones distintas de abordaje del problema por un lado la visión de los procesos y su forma de cómo se interviene en las tareas diarias para mejorar la puesta en marcha del proyecto y por otro el aseguramiento del resultado con métricas justas, que observan como el resultado de la combinación de ambas perspectivas confluyen para obtener datos valiosos.

3.5 Población y Muestreo

El estudio de la información se centra en los siguientes países

- El Salvador
- Ecuador
- Brasil

3.6 INSTRUMENTOS DE RECOLECCIÓN DE DATOS

El proceso de recolección incluirá las siguientes herramientas,

- Cuestionarios, los cuestionarios nos permitirán identificar los distintos procesos de cómo y en qué momento se llevan a cabo. Se aplican los cuestionarios a las siguientes oficinas y gerencias del grupo de empresas.
    - Oficina de compras.
    - Gerentes Financieros.
    - Gerente de compras y oficina de compras.
- Entrevista, se podrá realizar entrevistas al Director de cadena de suministros para corroborar el alcance del proyecto, con esto nos aseguramos que el proyecto tendrá realmente el enfoque que la alta gerencia requiere del mismo y además surge como el punto de partida para iniciar a delimitar el proyecto y sus entregables finales.
- Focus groups, a través de herramientas como Skype, que nos ayude a identificar las falencias del proceso actual y como la nueva aplicación ayude a solventar estas carencias.
- Técnicas observación, esto se llevará a cabo con los datos ya recolectados y su composición, a través de herramientas como Excel de Microsoft, para comprender las relaciones y la importancia de cada una de las variables que el proceso consume o da como resultado.

- Archivos de texto de importaciones, estos archivos permiten la inclusión de nuevas variables de análisis de información, con lo que se persigue dar una mayor profundidad del análisis de importaciones.

3.7 Técnicas de Análisis de la Información

Dentro de las técnicas de análisis se incluirá una herramienta altamente comprobada como lo es la Espina de Ishikawa o de causa y efecto como se muestra a continuación.

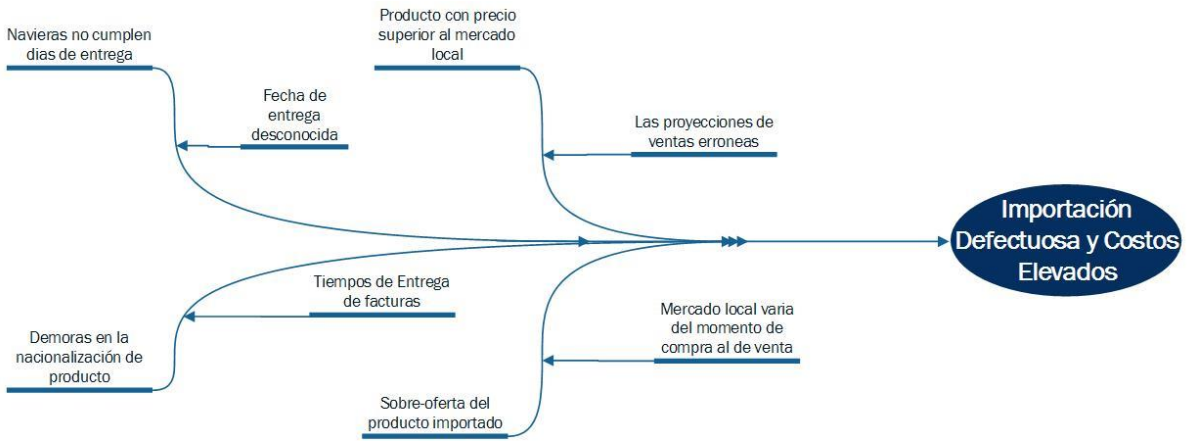


Figura 7. Espina de Ishikawa o de Causa y Efecto en el proceso de importación.

3.8 Estrategia de Desarrollo de la Propuesta

Se plantea la utilización de las siguientes estrategias de desarrollo, junto con sus diferentes usos en el proceso de desarrollo.

- Desarrollo Ágil, entregas constantes a los interesados de manera tal que se vaya validando los diferentes resultados y aprobar su entrega de manera tal que los avances se realicen de forma rápida y continua sin perder la perspectiva de la calidad de los entregables.

- Entregas incrementales o de modo Discovery, de la misma manera que la anterior se aplicaría para la primera fase del proyecto específicamente en el modelo multidimensional.
- CRISP-DM, esquema que se utiliza principalmente en este tipo de proyectos y que para nuestro caso se aplicara en la segunda fase del proyecto.

4 CAPÍTULO 4. ANÁLISIS DEL DIAGNÓSTICO

4.1 ANÁLISIS DEL PROCESO ACTUAL

La evolución del mercado del transporte es consecuencia de su complejidad en los deseos de los clientes y en el cual grupo Transmerquim también se ve envuelto. Siendo un mercado global muy anterior al actual proceso globalizador existente en otros sectores, el número de agentes intervinientes y las características específicas de legislación, formas de contratación, etc. condicionan su evolución y por ende su proceso de distribución.

Bajo este contexto económico-empresarial, los analistas de importaciones de Grupo Transmerquim vieron la necesidad de optimizar los recursos que se emplean actualmente en la cadena de suministros de las diferentes compañías en distintas zonas de América, pudiendo así responder al fenómeno de la globalización de la economía. Este fenómeno se ha visto notablemente influenciado por las tecnologías de la información que son capaces de transformar grandes cantidades de datos en información útil para la toma de

decisiones, contribuyendo a disminuir las barreras físicas entre mercados distantes como lo son el nuestro y el europeo o asiático. También esto ha traído que las empresas deban de trabajar en un entorno más hostil con un elevado índice de competencia. Esta situación es imprescindible para adoptar decisiones estratégicas que a Grupo Transmerquim les proporcione ventajas con respecto de sus competidores.

Estas decisiones estratégicas requieren un esfuerzo excepcional de parte de los encargados del área de cadena de suministros, así como gran cantidad de información normalmente histórica, procedente de diferentes fuentes y que permite al analista descubrir fenómenos y tendencias “escondidas” en los datos.

En la actualidad se tiene un sistema desarrollado en Access de Microsoft, lo que representa una limitante grande para compartir información, carga de datos, visualización de información y accesos concurrentes a la información. Otra del problema es la poca escalabilidad del sistema actual y su obsolescencia del producto que no representa la corriente actual y que representa una desventaja competitiva.

Otro de los puntos en contra es el tiempo por invertir en la preparación de la información inicial como suministro para el proceso y que representa muchas horas de diferentes colaboradores, información que se utiliza para la toma de decisiones de inicio de mes y que luego entra en desuso inmediatamente.

Otra de las limitantes actuales es la integración de análisis paralelos de la competencia que en este momento no es posible por el proceso de carga de información que debería de hacer y además que la herramienta no soportaría esta carga y el análisis analítico se vuelve imposible.

A continuación, se muestra a grandes rasgos el proceso en la herramienta actual,

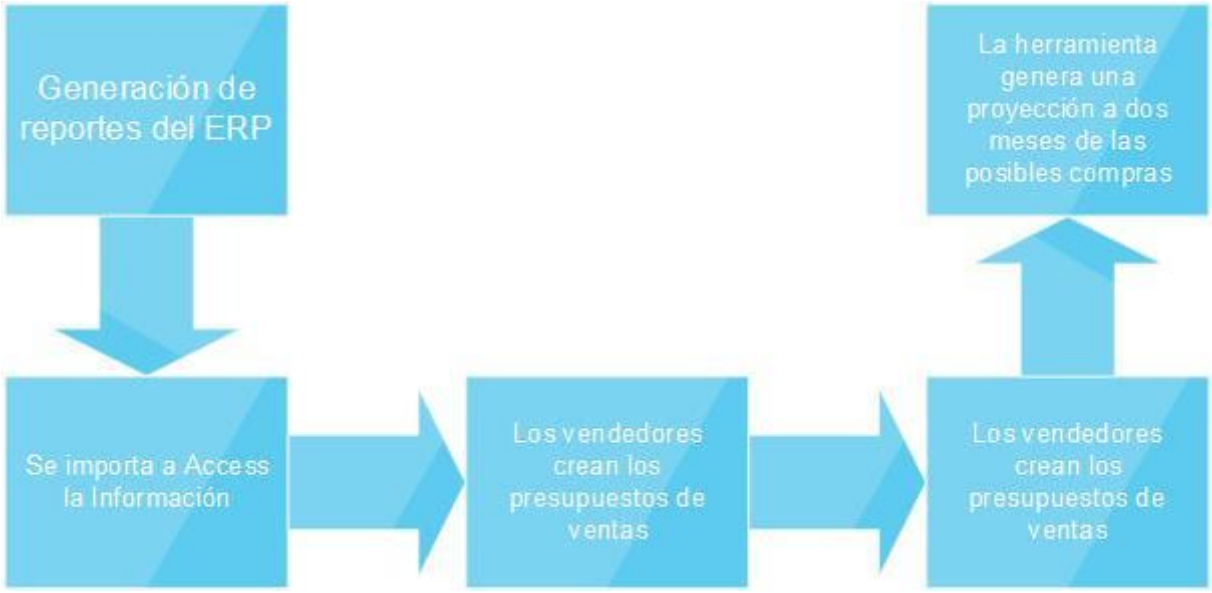
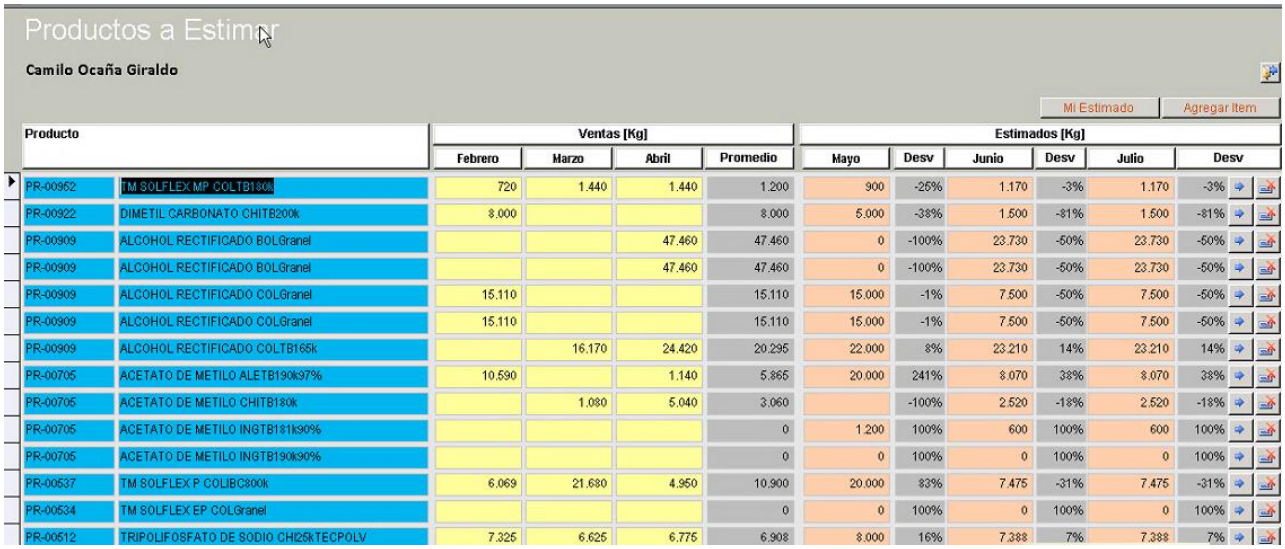


Figura 8 Proceso actual de análisis de información de importaciones.

Como se observa en la anterior imagen, los datos se extraen del ERP como fuente inicial del proceso, estos datos son datos postmortem de las transacciones realizadas y que trata de interpretar lo que podría suceder en un futuro, pero sin una tendencia clara.

Luego de esto, los vendedores hacen sus proyecciones basado en sus ventas y una desviación estándar configurada previamente en la herramienta. Como lo muestra la imagen siguiente,



Productos a Estimar

Camilo Ocaña Giraldo

Mi Estimado | Agregar Item

Producto		Ventas [Kg]				Estimados [Kg]					
		Febrero	Marzo	Abril	Promedio	Mayo	Desv	Junio	Desv	Julio	Desv
PR-00952	TM SOLFLEX MP COLTB180k	720	1.440	1.440	1.200	900	-25%	1.170	-3%	1.170	-3%
PR-00922	DIMETIL CARBONATO CHITB200k	8.000			8.000	5.000	-38%	1.500	-81%	1.500	-81%
PR-00909	ALCOHOL RECTIFICADO BOLGranel			47.460	47.460	0	-100%	23.730	-50%	23.730	-50%
PR-00909	ALCOHOL RECTIFICADO BOLGranel			47.460	47.460	0	-100%	23.730	-50%	23.730	-50%
PR-00909	ALCOHOL RECTIFICADO COLGranel	15.110			15.110	15.000	-1%	7.500	-50%	7.500	-50%
PR-00909	ALCOHOL RECTIFICADO COLGranel	15.110			15.110	15.000	-1%	7.500	-50%	7.500	-50%
PR-00909	ALCOHOL RECTIFICADO COLTB165k		16.170	24.420	20.295	22.000	8%	23.210	14%	23.210	14%
PR-00705	ACETATO DE METILO ALETB190k97%	10.590		1.140	5.865	20.000	241%	8.070	38%	8.070	38%
PR-00705	ACETATO DE METILO CHITB180k		1.080	5.040	3.060		-100%	2.520	-18%	2.520	-18%
PR-00705	ACETATO DE METILO INGTB181k90%				0	1.200	100%	600	100%	600	100%
PR-00705	ACETATO DE METILO INGTB190k90%				0	0	100%	0	100%	0	100%
PR-00537	TM SOLFLEX P COLIBC800k	6.069	21.680	4.950	10.900	20.000	83%	7.475	-31%	7.475	-31%
PR-00534	TM SOLFLEX EP COLGranel				0	0	100%	0	100%	0	100%
PR-00512	TRIPOLIFOSFATO DE SODIO CHI25kTECPOLV	7.325	6.625	6.775	6.908	8.000	16%	7.388	7%	7.388	7%

Figura 9. Muestra los resultados parciales de los productos por estimar

Como resultado, los encargados obtienen datos de proyecciones a dos meses del mes actual, basado en las variables antes mencionadas, con lo que se provee una base para negociar las futuras compras y un presupuesto de los siguientes meses, aunque como se menciona esto no toma ciertas variables que por su complejidad lo hace muy costoso para su inclusión dentro de este análisis. El resultado del proceso se muestra en la siguiente imagen.

Resumen de Compras Pendientes y Transito

Transmerquim de Colombia S.A.

Producto		Mayo			Junio			Julio		
		TTo	Compra	Total	TTo	Compra	Total	TTo	Compra	Total
PR-00016	ACEITE MINERAL	16.000	0	16.000	0	32.000	32.000	0	0	0
PR-00046	ACETATO DE BUTILO	0	0	0	0	16.000	16.000	0	0	0
PR-00047	ACETATO DE ETILO	304.000	0	304.000	0	96.000	96.000	0	0	0
PR-00047	ACETATO DE ETILO	201.600	0	201.600	0	0	0	0	0	0
	ACIDO ACETICO	0	0	0	0	4.000	4.000	0	0	0

Figura 10. Productos en tránsito y compras pendientes, herramienta actual.

Otro de los obstáculos que deben de sortear los usuarios es la poca integración con los sistemas transaccionales que no permiten sincronizar los artículos nuevos o las configuraciones de estos, que en ocasiones provoca que

los productos nuevos queden fuera de este análisis, trayendo consigo desabastos en productos emergentes.

4.2 ANÁLISIS DESEADO DE LA INFORMACIÓN

El aprovechamiento de la información actual hace que los competidores de un grupo de empresas en un mismo sector de negocios, sean más rentables y mucho más atractivos de inversión, esto hace que Grupo Transmerquim vea la necesidad de echar a la mano las nuevas herramientas de análisis de información, primero al obtener información de una forma más sencilla y sin tantas horas hombre en la elaboración de los registros y luego que permita aprender de la información y sus errores para luego tener un proceso más depurado en la toma de decisiones.

La empresa desea mantener un análisis similar al actual que base sus proyecciones en variables que provengan del sistema transaccional y que además integre un nuevo recurso, que se instaló en la actividad normal de la empresa, el CRM, herramienta que permite llevar las relaciones con los clientes, los presupuestos de ventas y las ventas finales en forma automática. Dentro de este análisis se debe de automatizar también la extracción, la transformación, la integración y la presentación de la información para que se lleve de forma casi en tiempo real los análisis y las proyecciones de las compras.

Otra de las premisas sobre la nueva solución es la liberación de los recursos y horas que se emplean en la creación de los reportes para que sean más bien utilizados en el ajuste y corrección de los datos, que se emplean creando una depuración con el tiempo.

Con esto también se persigue enfocar esfuerzos en aquellos grupos de productos o categorías que pudieran tener esa tendencia de crecimiento que no es posible con la actual herramienta, lo que permita cumplir con una de las mayores necesidades del grupo de empresas, hacer sus compras más proactivas que reactivas al hacer el modelo obsoleto y sin margen para el error, o en su defecto no tener propuestas salidas de contexto o desfasadas para el mercado.

El escenario idóneo, en una primera fase, debe de automatizar,

- La extracción de datos de sistemas transaccionales como lo es Microsoft Dynamics AX, Dynamics CRM y tablas de Excel.
- Limpieza de datos con la ayuda de herramientas que permitan crear reglas
- Integrar los datos en un almacén desde donde se puedan hacer consultas multidimensionales
- Visualizaciones en reportes estáticos.

Como segunda fase del mismo debe de integrar otras variables de datos como lo son,

- Archivos de importación de los países propuestos
- Aplicar algoritmos de aprendizaje supervisado de Machine Learning.
- Integrar la información junto con el análisis multidimensional de la primera fase.

Con esto se busca obtener una toma de decisión más robusta y que abarque diferentes aristas de un mismo problema. En donde se pueda obtener,

- Patrones de compras.
- Proyecciones de compras en el tiempo.
- Mejoras en los presupuestos de compras.
- Crecimiento en productos.
- Mejora en la cadena de suministros.
- Mejora en los inventarios y su obsolescencia.
- Mejora en el margen de negociación en los transportes de carga.
- Aseguramiento de disponibilidad de producto.
- Aumento en la satisfacción del cliente.
- Nuevos negocios con productos afines.

Con el análisis de esta información, se esperaría que el riesgo de nuevos negocios se vea disminuido y se saque mejor ganancia a los datos que actualmente se tiene e incorporar nuevas fuentes de datos que haga más robusta la información obtenida de las diferentes herramientas. Además, se puedan hacer simulaciones de proyecciones que indiquen la veracidad de la información y así obtener un porcentaje más alto de éxito. Este análisis nos debería arrojar pistas que nos haga girar de rumbo en otras actividades en donde se pueda ir bien pero que con el paso del tiempo se note leves declinaciones en sus proyecciones, así como las causas del porqué sucede esto y sus posibles soluciones.

5 CAPÍTULO 5. PROPUESTA DE LA SOLUCIÓN

Teniendo como base las limitantes mencionadas anteriormente se busca que la solución recomendada abarque las diferentes aristas del problema y además de un avance en el proceso, tanto a nivel de disponibilidad, como de tiempo y manejo de nuevas variables, para que así, se fomente la innovación y la mejora continua en los procesos de importación de productos y distribución.

Para lo cual se establece que el procedo debe ir en tres etapas, de las cuales el proyecto abarcara dos etapas de las mismas y dejara como puntos de trabajos a futuro la tercera etapa.

En la primera fase del proceso se desarrollaron reuniones con los colaboradores del área de compras y logística de Grupo Transmerquim a nivel regional, así como sus homólogos a nivel local en los países donde se desarrolla el proyecto actual. Esto para comprender la estructura y el ciclo completo de compras y distribución de los productos. Con la ayuda de técnicas agiles se busca la entrega del producto lo antes posible.

Se determinó que la primera fase del proceso debe de sustituir la herramienta actual, y permitir la automatización de los procesos de extracción de información de las diferentes fuentes de datos, procesamiento de la información, transformación en estructuras multidimensionales y visualizaciones amigables al usuario final, para que en el final pueda ser utilizada como parte de

un conjunto de variables en la toma de decisiones que afecten los procesos futuros de compras.

Para la segunda etapa del proceso se llevaron a cabo reuniones con los funcionarios corporativos de Grupo Transmerquim para la inclusión de una nueva variable en el análisis de información, como lo son los archivos de información de importaciones de los países en estudio, que contienen información de productos similares que los competidores importan y que el mercado de los países en estudio consume. En esta etapa se buscará predecir la información del valor CIF de un determinando grupo de artículos o familia de artículos.

Para la tercera etapa se busca analizar los datos de otras variables como son fletes globales, procesos de transportes y variaciones externas que afecten de una u otra forma el proceso de consolidación de la carga y su distribución a los diferentes países.

Es por esto que se llega a determinar que los procesos normales de extracción, transformación de información y análisis no son tan flexible y tan natural como se desean, cuando se está tratando de ejecutar minería de datos con las herramientas actuales, principalmente porque las fuentes de datos se encuentran en la nube de información, el tamaño de dichas bases en algunos casos se vuelve inmanejable y el procesamiento de la información comprendería tener servidores muy robustos y que procesarían dicha información por un par de días y luego entrarían en procesos improductivos que encarecería el proyecto y que además como etapa de pruebas se debería de ver los resultados en muy corto plazo. Además, que las estructuras de minería de datos normales

nos presentan ciertas limitantes con algoritmos de regresión, neuronales y de clasificación que son los que principalmente que planea ejecutar para su análisis. Es por esto que se decidió llevar esta fase del proyecto a Azure Machine Learning y consumir los diferentes algoritmos a través de Web Services de la herramienta.

Luego de estos procesos el comité tomará sus decisiones de compras en el futuro y así poder negociar los contratos de compras, teniendo como meta bajar los costos que implica la importación del producto y mejorar la distribución a los diferentes países de la región.

5.1 Primera Fase: Modelo Multidimensional

En la construcción de la primera parte del proceso de recambio de la herramienta actual se procede a tener claro los actores de la construcción de la información y la forma en cómo se extrae la información, la fuente de datos y su procesamiento dentro de Access y luego de esto se pretende replicar lo mismo en estructuras multidimensionales.

Para esta parte del proyecto se procede a aplicar la metodología Ágil o entregas incrementales, ya que esta parte del proyecto busca el cambio inmediato de una herramienta que por su complejidad se ha vuelto insostenible en el tiempo.

Luego de revisar a detalle la información, se determinó que se tiene tres fuentes distintas de datos,

- El ERP, Microsoft Dynamics AX en su versión 2009 RU8. Desde donde se extrae
    - Ventas actuales, son las ventas que se han realizado en un rango de fecha de un determinado vendedor. Esta información del usuario la extrae del módulo de clientes.
    - Inventario a hoy, son los artículos disponibles a la fecha del reporte, desglosados en sus diferentes dimensiones del inventario. Esta información el usuario la extrae del módulo de inventarios.
    - Compras pendientes, son las compras que aún no se han facturado pero que se están en proceso de recibir o de nacionalizar. Esta información el usuario la extrae del módulo de proveedores.
- EL CRM, Microsoft Dynamics CRM en su versión 2011
    - Prospectos de ventas, son las posibles ventas proyectadas en un rango de fechas determinado.
- Excel, con el presupuesto de compras en un rango determinado.

Con esta información se inició con el análisis de las estructuras de datos, sus dimensiones y métricas que se implementaran. Para lo cual se proponen las siguientes dimensiones y métricas.

5.1.1 Dimensiones

1. Artículos. Esta dimensión corresponde al catálogo de artículos de GTM, campos disponibles dentro de la dimensión:

- Artículo
- Clasificación ABC
- Código Familia Artículo
- Código Sub Familia Artículo
- Familia Artículo
- Código de articulo
- Sub Familia Artículo

2. Clientes. Esta dimensión corresponde a los clientes de GTM, campos disponibles:
   - Cédula Jurídica
   - Código Cliente
   - Nombre Cliente
   - Nombre Consolidado
   - VIP
3. Compañías. Esta dimensión hace referencia a todas las compañías bajo GTM, campos disponibles:
   - Código Compañía
   - Descripción Compañía
4. Facturas. Esta dimensión corresponde a las facturas, campos disponibles:
   - Factura Compras

- Factura Venta

5. Grupos de Clasificación de los Clientes. Esta dimensión indica el grupo económico al cual pertenece el cliente, campos disponibles:
   - Grupo Clasificación Cliente
6. Grupos de Clientes. Esta dimensión indica qué tipo de cliente los campos disponibles:
   - Código Grupo Cliente
   - Grupo Cliente
7. Inventario. Esta dimensión hace referencia a dimensiones del inventario, campos disponibles:
   - Almacén
   - Configuración
   - Origen, Presentación y Grado
   - Situación Fiscal
   - Ubicación
   - Zona Costo
8. Pedidos de Ventas. Esta dimensión corresponde a los pedidos de las ventas, campos disponibles:
   - Pedido Venta
9. Segmentos: esta dimensión corresponde a los segmentos de negocio de los clientes, campos disponibles:
   - Código Segmento

- Descripción Segmento

10. Sub Segmentos. Esta dimensión corresponde a los sub segmentos de negocio de los clientes, campos disponibles:
    - Descripción Sub Segmento
    - Sub Segmento
11. Tiempo. Esta dimensión corresponde al tiempo descompuesto en:
    - Día
    - Día del Año
    - Semestre
    - Mes
    - Trimestre
    - Cuatrimestre
    - Semestre
    - Años
12. Tipo de Cliente. Esta dimensión hace referencia a los tipos de cliente, campos disponibles:
    - Tipo Cliente
13. Vendedores. Esta dimensión hace referencia a los vendedores en GTM, campos disponibles:
    - Grupo Vendedor

- Nombre Vendedor

14. Zonas de Venta. Esta dimensión hace referencia a las zonas de ventas, campos disponibles:
    - Zona Venta
15. Pedidos de Compras. Esta dimensión corresponde a los pedidos de las compras en tránsito, campos disponibles:
    - Pedido compra
16. Proveedores. Esta dimensión corresponde a los proveedores de GTM, campos disponibles:
    - Cédula Jurídica
    - Código Proveedor
    - Nombre Proveedor
    - Nombre Consolidado del Proveedor

5.1.2 Métricas

1. Margen de Contribución en Dólares y Moneda Local. Esta métrica corresponde a la utilidad (venta menos costos asociados a la venta) disponible en dos monedas, dólares estadounidenses y moneda local.
2. Monto de Venta en Dólares y Moneda Local. Esta métrica corresponde al monto de la venta disponible en dos monedas, dólares estadounidenses y moneda local.
3. Costo Gastos Varios Asignación y Reversión en Dólares y Moneda Local: esta métrica corresponde al monto asignado al detalle de la factura

asociado a gastos sobre la venta y la compra, disponible en dos monedas, dólares estadounidenses y moneda local.

4. Costo Producto en Dólares y Moneda Local. Esta métrica corresponde al Costo Promedio del Producto a la Fecha disponible en dos monedas, dólares estadounidenses y moneda local.
5. Toneladas Vendidas. Esta métrica corresponde a la cantidad en toneladas de la venta.
6. Toneladas Compradas. Esta métrica corresponde a la cantidad en toneladas de la compra.

5.1.3 Proceso de Extracción

Con base en estas dimensiones y métricas se crea el proyecto de Integration Services que ejecuta la extracción diaria de los datos, la misma se detalla a continuación,

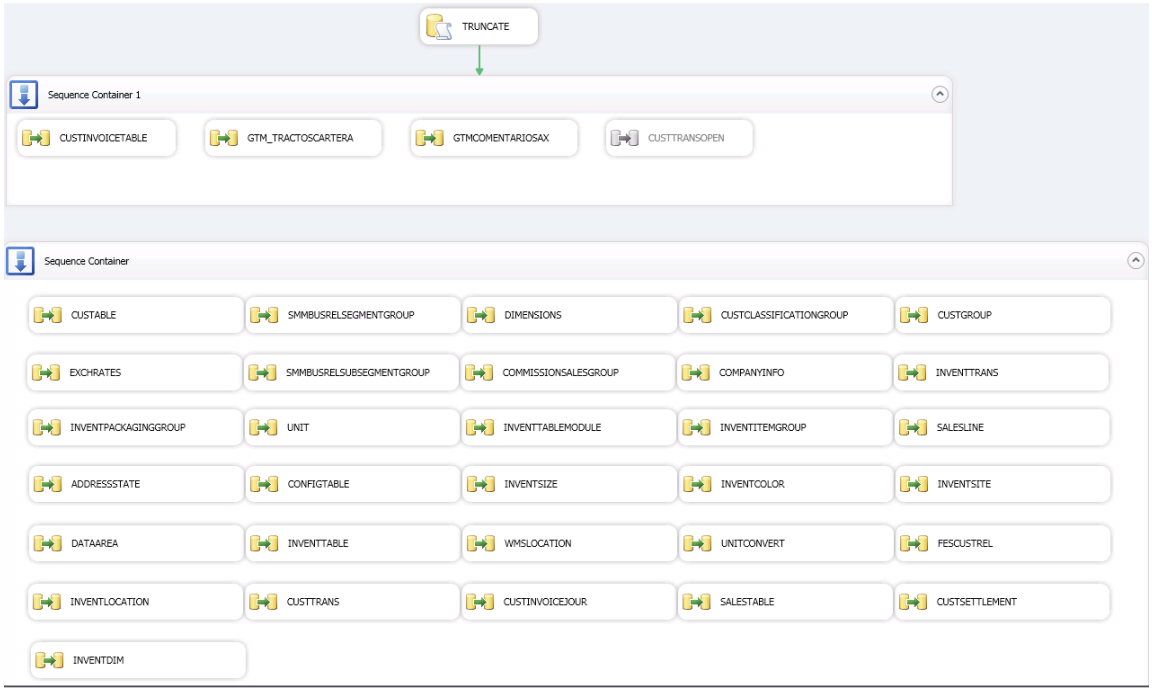


Figura 11. Proceso de extracción de información.

La misma se encarga de automatizar la extracción de la información de tres distintas fuentes, a saber:

- Microsoft Dynamics AX 2009, su base de datos está en SQL Server 2008 R2.
- Microsoft Dynamics CRM 2011, su base de datos está en SQL Server 2012.
- Excel Budget.

Los catálogos que se extraen abarcan tanto las compras en tránsito, el inventario disponible a la fecha de la consulta y las ventas finales de los clientes. En el Apéndice 1, ahí se muestra el diseño de bases de datos de dos de las fuentes de datos. Con esta información en nuestro “Staging Area”, se procede a la transformación de la información y a cargarla en el “Data Warehouse”.

5.1.4 Proceso de Transformación

El proceso de trasformación se incluye la creación y carga de las diferentes dimensiones que el modelo implica, durante el mismo se implementa el tema de tipo de cambio promedio para poder consolidar entre aquellos países que la moneda local sea diferente a dólares estadounidenses, para poder implementar esto se llevara un tipo de cambio promedio del mes que implicara tomar el tipo de cambio inicial del mes más el tipo de cambio del último día del mes y dividirlo entre dos. Este proceso se lleva en el “Staging Area” como un proceso separado por conveniencia de programación y se invocara en los procesos de Integration Services. Dentro de estos procesos se incluye información de las tres fuentes de datos como se mencionó anteriormente.

Los procesos de transformación se muestran en la siguiente imagen.

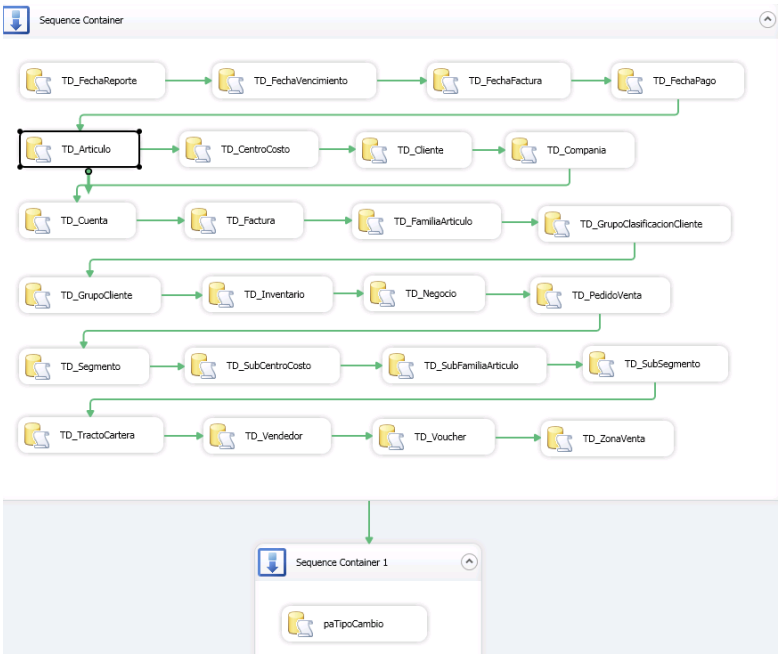


Figura 12. Proceso de transformación de importaciones.

Se crea un proceso separado que implica ejecutar un procedimiento dentro del “Staging Area” que se encarga de llevar cuáles facturas fueron

anuladas de forma total o parcial para que en el final del proceso se lleve la información lo más detallada y fidedigna esta fue una de las mejoras al sistema anterior ya que el mismo no disponía de una lógica de asociación de facturas y notas de crédito. Y por consiguiente hacía la información no tan exacta para su análisis. En la siguiente imagen se muestra el paquete que ejecuta dicho proceso.

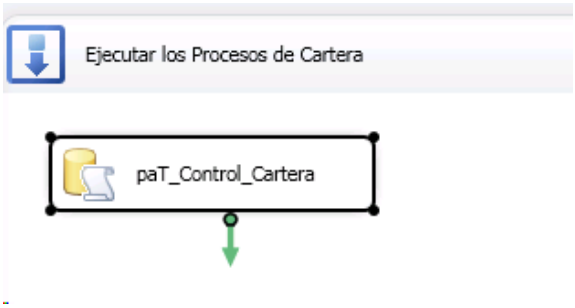


Figura 13. Proceso de asociación de facturas y notas de crédito.

5.1.5 Proceso de Carga

En el siguiente paso se cargan la estructura de la tabla de hechos, para esto se ejecutan varios procedimientos almacenados, algunos se detallan a continuación,

- THCostos: agrupa los costos de las facturas y notas de crédito.
- THCyR: agrupan los cierres y recálculos de inventarios.
- THGV: agrupan los costos de gastos varios en la venta.
- UnirCostos: une los costos asociados a una factura
- VentasProductosNC: agrupa los montos de ventas.
- VentasUnirResultados: resultados de márgenes por factura.

En la siguiente imagen se detallan los procesos que se ejecutan en este paso,

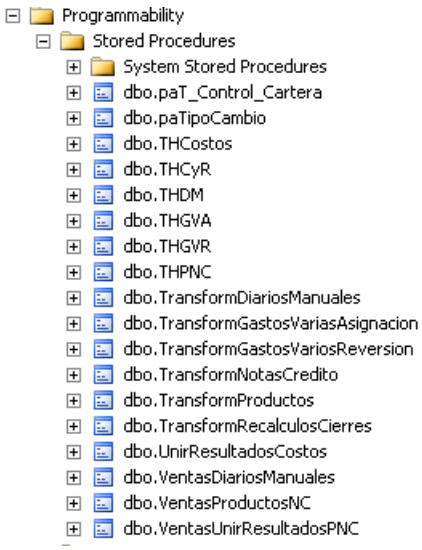


Figura 14. Procedimientos almacenados que se ejecutan en la carga de información.

Una vez ejecutado el proceso la información se carga en el Data Warehouse dispuesto para tal proceso. En la siguiente imagen se ilustra la base de datos utilizada,

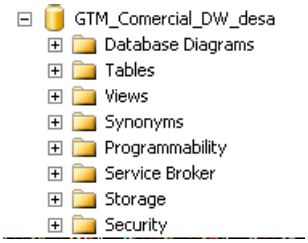


Figura 15. Base de datos resultante de la carga de información.

5.1.6 Estructuras Multidimensionales

Luego de esto se procede a la creación de las estructuras multidimensionales que servirán de fuente de consulta y de remplazo para la herramienta actual. En la misma se podrá hacer las consultas que actualmente se llevan de forma muy manual y con esto se cumple el primer paso del proceso que es la extracción automática de ventas en un periodo dado, el inventario a la fecha del análisis y las compras en tránsito. Dentro del mismo proceso de

transformación se ejecutan secciones de limpieza y aseguramiento de información, como lo es la integración de facturas y sus notas de créditos parciales o totales, y ayuda al usuario a tener consultas diferentes vistas desde dimensiones del cubo que anteriormente no las tenía.

En la siguiente imagen se muestra el resultado de una consulta de facturas por código de cliente en un rango de fechas determinado para una compañía X.

Id Compania	(Multiple Items)			
Año	2011			
Numero Mes	(Multiple Items)			
Monto Original Local	**Column Labels**			
Row Labels	**AGOSTO**	**OCTUBRE**	**SETIEMBRE**	**Grand Total**
07-000002	₡ 24,511,388.14	₡ 22,644,510.11	₡ 43,371,753.23	₡ 90,527,651.48
07-000004	₡ 3,213,076.58	₡ 1,314,467.19	₡ 2,844,021.86	₡ 7,371,565.63
07-000005	₡ 2,709,497.67	₡ 4,401,273.14	₡ 1,912,531.05	₡ 9,023,301.86
07-000008			₡ 196,535.17	₡ 196,535.17
07-000009	₡ 7,663,615.77	₡ 3,034,617.03	₡ 8,248,312.53	₡ 18,946,545.33
07-000010	₡ 490,564.11	₡ 249,652.83	₡ 995,021.50	₡ 1,735,238.44
07-000011	₡ 4,051,855.99	₡ 2,533,330.78	₡ 2,904,604.22	₡ 9,489,790.99
07-000012	₡ 631,828.03	₡ 1,055,820.93		₡ 1,687,648.96
07-000013	₡ 325,968.04	₡ 165,671.16	₡ 495,560.44	₡ 987,199.64
07-000015	₡ 3,065,475.92	₡ 3,421,610.27	₡ 3,642,186.79	₡ 10,129,272.98
07-000016		₡ 45,621.49		₡ 45,621.49
07-000017	₡ 41,209,541.63	₡ 24,132,242.10	₡ 386,460.00	₡ 65,728,243.73
07-000018	₡ 2,390,944.40	₡ 1,937,017.12		₡ 4,327,961.52
07-000019	₡ 20,471,747.10	₡ 13,958,154.73	₡ 4,458,263.76	₡ 38,888,165.59
07-000020	₡ 4,259,584.64	₡ 3,334,207.45	₡ 2,639,309.25	₡ 10,233,101.34
07-000022	₡ 1,129,815.81			₡ 1,129,815.81
07-000023	₡ 467,516.93	₡ 646,802.17	₡ 466,186.70	₡ 1,580,505.80
07-000024	₡ 8,371,630.48	₡ 8,254,317.21	₡ 8,082,893.96	₡ 24,708,841.65
07-000026		₡ 1,882,969.28	₡ 1,187,244.90	₡ 3,070,214.18
07-000027	₡ 1,821,900.59	₡ 516,439.72	₡ 657,872.21	₡ 2,996,212.52
07-000028	₡ 8,217,718.03	₡ 6,779,896.74	₡ 5,741,806.46	₡ 20,739,421.23
07-000029	₡ 700,332.30	₡ 997,441.39	₡ 321,650.95	₡ 2,019,424.64
07-000030			₡ 66,013.47	₡ 66,013.47
07-000031	₡ 17,090,751.47	₡ 16,578,492.13	₡ 22,877,072.58	₡ 56,546,316.19
07-000033	₡ 2,822,728.72	₡ 8,516,410.27	₡ 5,145,399.78	₡ 16,484,538.78
07-000034	₡ 5,997,354.40	₡ 6,128,060.90	₡ 17,296,740.56	₡ 29,422,155.86
07-000035	₡ 198,619.51	₡ 269,209.20	₡ 266,498.72	₡ 734,327.43
07-000038	₡ 423,754.92			₡ 423,754.92
07-000042	₡ 3,815,694.56	₡ 1,311,274.60	₡ 661,863.60	₡ 5,788,832.76
07-000043	₡ 9,137,763.20	₡ 7,362,405.70	₡ 3,943,266.06	₡ 20,443,434.96
07-000045	₡ 158,510.92	₡ 162,134.83	₡ 187,242.70	₡ 507,888.45
07-000046	₡ 268,235.73	₡ 490,059.64	₡ 873,597.21	₡ 1,631,892.58
07-000047		₡ 190,000.00	₡ 380,000.00	₡ 570,000.00

Figura 16. Consulta de facturas resultante.

Código Compañía	(Multiple Items)	
Año	2015	
Numero Mes	01	
Row Labels	**TON - UNI Vendidas**	**Costo Producto Moneda Local**
Aceites	13.44	-13,742,632.53
Acidos y Alcalis	531.86	-189,125,322.43
Agroquimicos	323.78	-88,884,560.90
Carbohidratos	9.70	-3,531,719.72
Complementos para Ferreteria	1.00	-3,320.00
Condimentos, Saborizantes y Grasas	2.51	-5,951,398.77
Insumos Farmaceuticos y Agropecuarios	2.40	-1,221,067.50
Lubricantes y Grasas	18.56	-19,553,088.02
Material de Empaque	0.02	-253,475.11
Minerales, Sales y Otros Compuestos	1,488.82	-177,656,799.24
No Definido	0.00	897,933.65
Otros	16.70	-7,373,113.24
Otros Servicios	0.00	0.00
Productos de Limpieza y Perforacion de Pozos	3,176.75	-724,860.95
Productos de Limpieza, Desinfeccion y Aromatizacion	22.50	-7,884,441.26
Resinas y Aditivos Auxiliares	8.00	-7,958,933.45
Solventes	646.53	-537,364,504.76
Thinner y Mezclas	311.58	-252,678,051.15
Vaselina	7.01	-7,470,511.93
Grand Total	**6,581.16**	**-1,320,479,867.31**

Figura 17. Consulta de inventario disponible por familia.

En la imagen anterior se muestra una consulta del inventario disponible dado en toneladas de un grupo de empresas para un mes dado.

En el Apéndice 2 se adjunta el diagrama de la vista multidimensional.

5.2 Segunda Fase: Técnicas de Machine Learning

La siguiente fase se involucra el análisis de nuevas variables dentro del proceso de importaciones, dentro de estas están los archivos de importaciones de diferentes países en donde se tiene presencia.

Para esta fase del proyecto se implementará la metodología CRISP-DM, la cual consiste en la siguiente metodología de acción,

- Comprensión del negocio. Objetivos y requerimientos desde una perspectiva no técnica. Como primer paso del análisis, será reconocer las

verdaderas necesidades del negocio y realizar las preguntas correctas, que permitan conocer las necesidades de conocimiento sobre los datos por analizar.

- Comprensión de los datos. Conocer los datos teniendo presente los objetivos del negocio.

- Preparación de los datos. Al obtener una muestra de datos se selecciona los datos por ser analizados, se procede con una herramienta de limpieza de datos. Se construye de un proceso escalable según la necesidad y se integra el conocimiento de los diferentes actores del proceso de importaciones. Para que en el final de esta etapa los datos salgan con un formato establecido, para ser consumido por las técnicas de Machine Learning.

- Modelado. En esta etapa se selecciona las técnicas o algoritmos por aplicar a los datos “limpios”, se crea un diseño acorde con la necesidad del negocio y se evalúa el modelo propuesto para así ser ajustado, si es requerido. Haciendo comparaciones con diferentes algoritmos y su salida según la necesidad del negocio.

- Evaluación. Se evalúan los modelos de las fases anteriores para determinar si son útiles a las necesidades del negocio. Se coteja la información existente con los resultados de los existentes. Este proceso lo realiza los usuarios de importaciones.

- Despliegue. En el final se procede a explotar la utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización. Para esto se ha determinado que los Web Services de Microsoft Azure

Machine Learning es la forma más adecuada de servir esta herramienta a los usuarios finales. Se prevé un seguimiento de ajuste y evaluación de la herramienta.

5.2.1 Desarrollo de la Propuesta

Uno de los principales problemas es la información con la que se cuenta ya que la proyección de diferentes productos se basa en información pasada y los cálculos se hacen con esta información, previendo un mismo comportamiento en el tiempo. Además, que los márgenes en la búsqueda de nuevos nichos de mercado asociados al negocio se hacen basados en informes de probabilidades poco fiables del mercado y no tan científicas, no toma en cuenta la diversificación según temporalidad y su rendimiento, así como la oportunidad de participación en nuevos mercados. Para esto se plantea un proceso que involucra diferentes pasos que se detallaron antes y se basan en la metodología CRISP-DM.

5.2.1.1 Comprensión de las necesidades del negocio

Para iniciar con este proceso se plantearon entrevistas con los usuarios involucrados y encuestas, sobre temas de importaciones y su necesidad más puntual de análisis de información. En el Apéndice 5 se podrá observar la entrevista y encuesta hecha a los líderes de importaciones de los diferentes países donde se planea introducir la herramienta. Ante las consultas necesarias se lleva a evaluar las carestías de una herramienta que sea dinámica, pueda hacer comparaciones automáticas, extracciones continuas y sobre todo como requisito que se consolide en una herramienta indispensable en el análisis de información, Office Excel de Microsoft. Se evalúa la situación actual y se establecen los objetivos de la fase en el nivel del negocio.

5.2.1.2 Comprensión de los datos por evaluar

Luego de las consultas de negocio vienen las consultas un poco más técnicas entre las cuales se desea entender las relaciones que existen en los datos y comprender el problema de manejo de los grupos de datos. Se hace un muestreo de los datos y se hace una descripción inicial. Se valida la calidad de los datos y se hacen las recomendaciones y el plan de acción. Los archivos con los que se cuentan son archivos separados por comas y de texto plano.

Los archivos se evalúan y se determinan que su estructura varia en campos según el país desde donde se recolecta. Para lo cual se hace un análisis de los diferentes archivos y luego un consolidado de campos, esto se determina con la ayuda del encargado del proceso la importancia del campo en el análisis de información.

Como primer obstáculo se determinó que no todos los archivos tienen información similar o la expresión en una de sus columnas se homologa por otro campo, en total se cuenta con 139 campos diferentes en los distintos archivos de importaciones. En el Apéndice 3 se puede observar una muestra de la consolidación de estos campos.

Luego se creó una fuente de homologación en la que según criterios técnicos de importaciones se pueden homologar campos de distintos archivos, el archivo final sirve de base de análisis para los archivos de los países en donde se está haciendo el estudio y que además sirve como insumo para la siguiente parte del proyecto. Este archivo contiene el país, el campo original, el campo homologado y su importancia para el análisis. En el Apéndice 4 se puede observar una muestra del archivo de homologación final.

Como resultado se tienen que los campos por evaluar y que se homologan entre todos son los siguientes,

- Fecha de registro.
- Detalle de la mercadería
- Importador del producto.
- País de origen del producto.
- Peso en kilogramos del producto.
- Proveedor del producto
- Partida arancelaria, que es el código que agrupa una categoría de mercadería determinada.
- Valor CIF del producto, es el valor del producto más seguro y flete.
- Valor FOB del producto, es el valor del producto sin flete y sin seguro.
- Valor del Flete y el Seguro
- Valor del costo por kilogramo importado

5.2.1.3 Preparación de los datos

1. Proveedores de archivos de importaciones. Las empresas Sicex, Penta Transactions envían los archivos de importaciones de aquellos países en donde la información no sea pública o no se pueda descargar del ministerio correspondiente, países como Honduras, Nicaragua y Costa Rica se extraen los archivos de importaciones de los sitios web de las instituciones encargadas.

2. Esta información se puede decir que estará “cruda” se descarga a una carpeta compartida en el servidor de archivos destinado para tal fin. Los archivos deben de ser cargados con una determinada nomenclatura que el analista de importaciones ya debe de tener clara.

3. Extracción a una base de datos Temporal 1. la información de los archivos en formato Excel y de texto plano se cargan a una base de datos llamada Temporal 1, por medio de un proyecto de Integration Services, en este proyecto se lleva a cabo la extracción de los datos, la asignación de una fecha de carga de los datos y por último la compresión de los archivos, movimiento a una carpeta de procesados y la eliminación de los archivos originales. En la siguiente imagen se ilustra el proceso descrito.

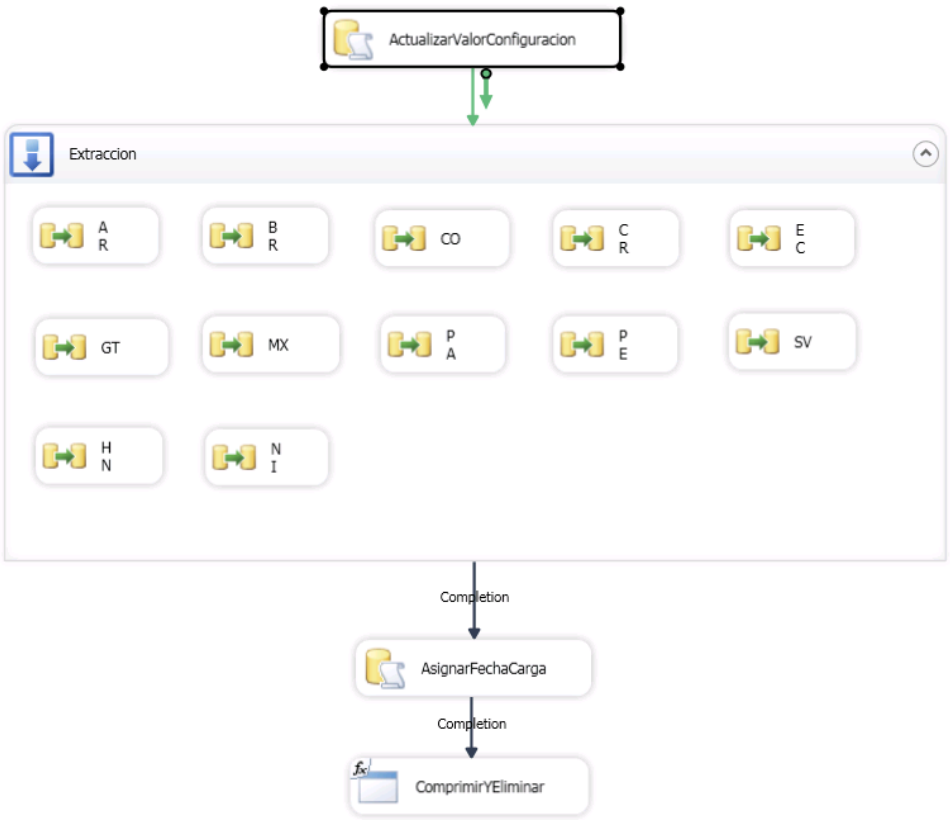


Figura 18. Extracción a una base de datos Temporal 1.

4. Limpieza y Calidad de Datos. Luego de la extracción, el proceso continua con la intervención del analista de importaciones y su aplicación de conocimiento del negocio, apoyado por una herramienta de calidad de datos para el proyecto se propone el uso de Data Quality Services de Microsoft SQL Server 2012, la misma se configura con las reglas de negocio o de homologación de datos y va ampliando su base de conocimiento conforme se procesen más y más fuentes de datos al dar los primeros indicios de un proceso automático de calidad y filtrado de información inicial. Este proceso también se programó un paquete de Integration Services para su ejecución automática. En la siguiente imagen se ilustra lo mencionado.

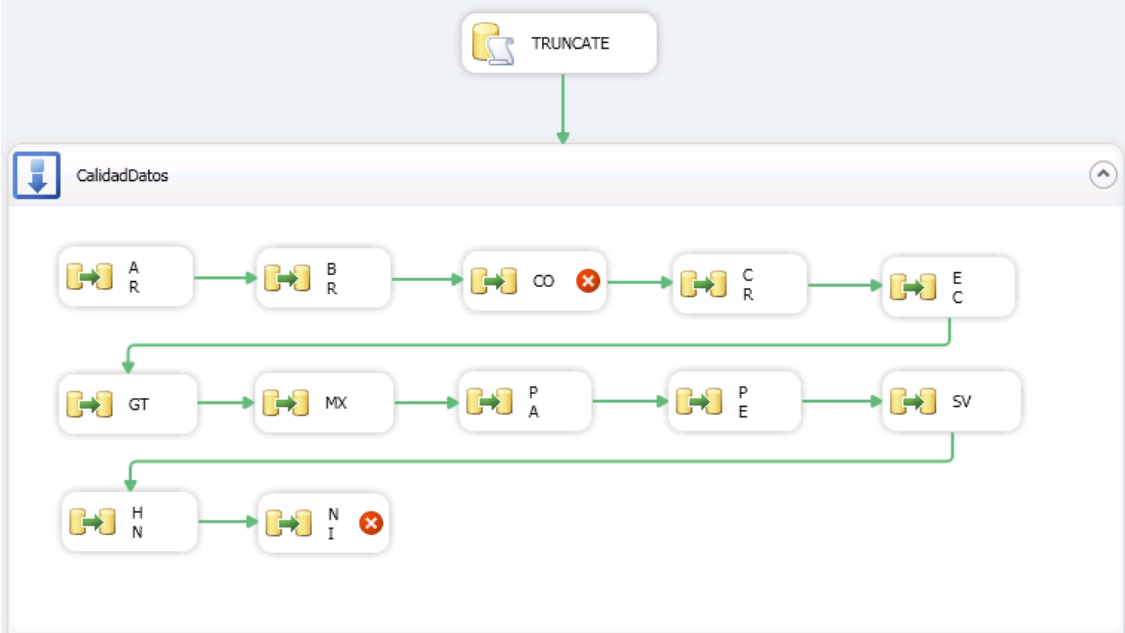


Figura 19. Procesos de limpieza y calidad de datos.

5. Extracción a una base de datos Temporal 2. Cada fuente de datos pasa por un proceso de limpieza y se crean sub conjuntos de datos temporales que se cargan en una segunda base de datos llamada Temporal 2, en

este paso el analista de importaciones toma decisiones con respecto de si aprueba o rechaza alguna corrección que sugiera la herramienta de calidad de datos.

6. Coincidencias. Estos sub conjuntos de datos separan la información en datos nuevos, correcciones posibles que sugiere la herramienta de calidad de datos basado en las reglas programadas, correcciones hechas por la herramienta, registros correctos e información inválida. En este punto el analista ajusta la herramienta al evaluar los sub conjuntos de datos resultantes y decide que pasa y que no corresponde a lo implementado en las reglas de negocio programadas en la herramienta de calidad de datos. Los datos que deben pasar nuevamente por las reglas del negocio se ejecutan nuevamente.

7. Extracción a una base de datos Temporal 3. Luego del proceso anterior se tiene como resultado una fuente de datos más ajustada y limpia según las necesidades del negocio, a esta base de datos, llamada Temporal 3, que consistiría en los datos aprobados, correctos o corregidos en el proceso. De esta información se desprende información puntual para el análisis en Machine Learning como lo es:

    a. País de origen del producto.

    b. País de destino.

    c. Código del artículo.

    d. Descripción comercial del artículo.

    e. Proveedor del producto.

f. Partida Arancelaria.

g. Importador del producto.

h. Fecha de carga.

i. Valor CIF del producto.

j. Valor FOB del producto.

k. Peso Neto.

l. Peso Bruto.

m. Valor en Kilogramos del producto.

n. Valor del Flete y Seguro.

5.2.1.4 Modelado de la información

Como parte del proyecto de análisis de nuevas variables para el modelo de importaciones de Grupo Transmerquim, una herramienta nueva es Azure Machine Learning, la cual nos permite procesar los grupos importantes de datos y consolidarlo con otras fuentes de forma transparente, además permite utilizar herramientas comunes para el usuario final y que con poca curva de aprendizaje llega a dominar.

Como parte de este proceso se llega al análisis de los algoritmos que se ajustan al aprendizaje supervisado propuesto. Para este caso se va a tomar como base el siguiente modelo de aprendizaje supervisado.

- Regresión lineal
- Regresión en redes neuronales.

- Regresión en arboles de decisión.
- Regresión lineal bayesiana

A continuación, se detalla los pasos por seguir en el proceso de modelado de los datos por los algoritmos de Azure Machine Learning,

8. Subir Información a Azure Machine Learning. El siguiente paso es importar esta información a Azure Machine Learning, para lo cual se procede a exportar los archivos para que se puedan subir a Azure Machine Learning por medio de archivos separados por comas, los mismo se importan desde la herramienta web.
9. Aplicación de Algoritmos Supervisados. Con la información requerida se procede con la ejecución de procesos de modelación de datos que supone una nueva limpieza y acomodo de datos que nos permita eliminar sesgo de información, luego de esto se aplica los algoritmos seleccionados para este proceso. La información debe de procesarse según sea la necesidad, por país o por región vinculante, por ejemplo, Centroamérica o Zona Andina.
10. Patrones y Tendencias. Se analiza la información según sea la necesidad. Se seleccionan diferentes modelos para poder cotejarlos, analizar sus diferencias y saber si los algoritmos seleccionados pueden tener resultados similares o son respuestas opuestas según el trato que se le da a los datos. Para luego de esto seleccionar los modelos adecuados al análisis.

Figura 20. Flujo de datos de Data Quality hacia Azure Machine Learning

En la imagen anterior se muestra el flujo de datos desde los archivos creados por Data Quality hacia Azure Machine Learning.

5.2.1.5 Evaluación y despliegue de la información

11. Análisis de Datos. Con la información procesada, el analista de importaciones se encargado de cotejar la información en una hoja de Microsoft Excel 2013, por medio del Web Services que Azure Machine Learning pone a disposición del usuario final y con la cual el usuario tiene la posibilidad de trabajar datos procesados y descargados a una carpeta compartida o nuevos datos que tenga como parámetro las columnas propuestas en el modelo de aprendizaje supervisado.

En la siguiente imagen se detallan los pasos del proceso que finaliza con datos analizados por medio de Machine Learning.

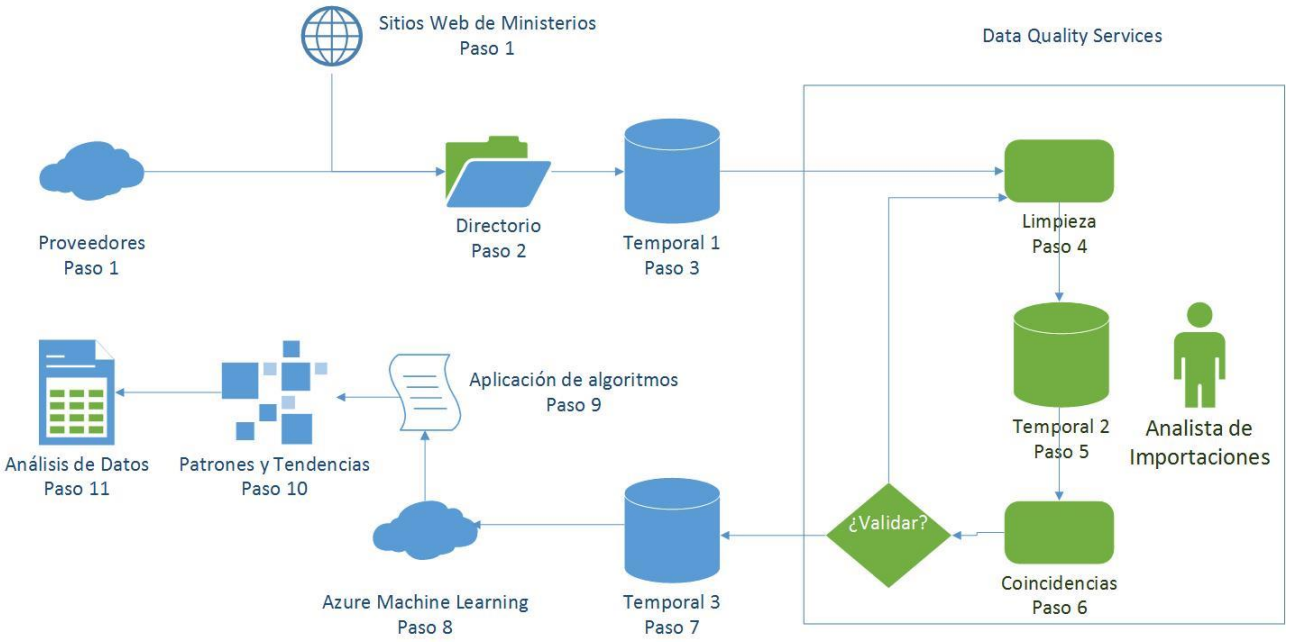


Figura 21. Flujo de información en el proceso de predicción de variables.

5.2.1.6 Azure Machine Learning

5.2.1.6.1 Preparación de los Datos

5.2.1.6.1.1 Conjunto de datos de entrenamiento

Como primer paso del análisis. la escogencia del conjunto de entrenamiento nos permite dar el primer paso en el análisis de la información y en el aprendizaje y ajuste de los algoritmos seleccionados. Para esto, se debe de tener presente las consultas que el algoritmo debe de responder. Como resultado se desea conocer,

I. ¿Cuál será el valor CIF Y en un producto, cuando se tenga un valor de kilogramos X?

II. ¿Cuál será el valor CIF Y de un producto, cuando el producto se importe de un país X?

III. ¿Cuál será el valor CIF Y de un producto, cuando el producto tenga un proveedor X?

IV. ¿Cuál será el valor CIF Y de un producto, en una fecha determinada X?

V. ¿Cuál será el Valor CIF Y de un producto, cuando el producto tenga un país de destino X?

Basados en estas incógnitas iniciales es que se empieza a tener nuestra función de regresión lineal, la cual se detalla en el capítulo de marco teórico del proyecto, la cual debería de verse de la siguiente forma,

$$\mathcal{F}(x) = \beta_0 + \beta_1 X$$

La función detalla la necesidad de calcular o predecir un valor X dados unos valores de entrada. Para esto el primer acercamiento con los datos es el cumplimiento de esta fórmula en los datos por evaluar. Para esto se procede a graficar las cinco familias que comprende el estudio,

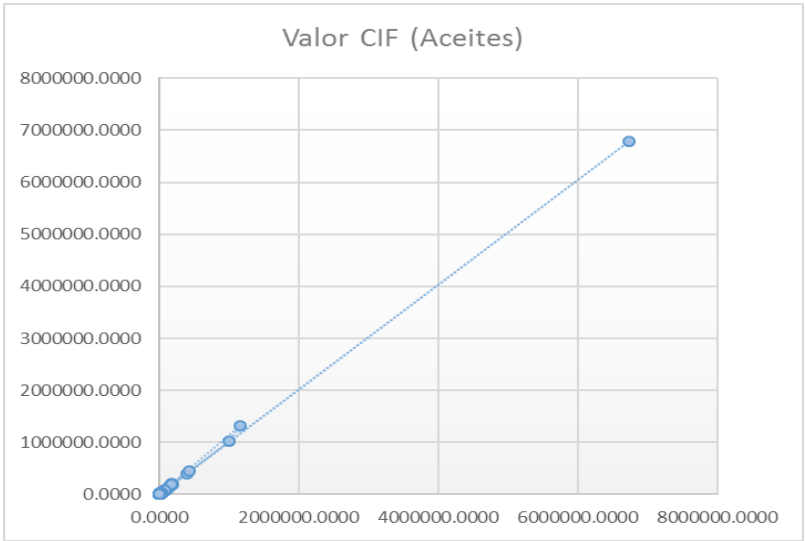


Figura 22. Relación del valor FOB y el valor CIF de aceites.

En la imagen anterior se muestra la relación del valor FOB con respecto del valor CIF del producto de la familia de aceites.

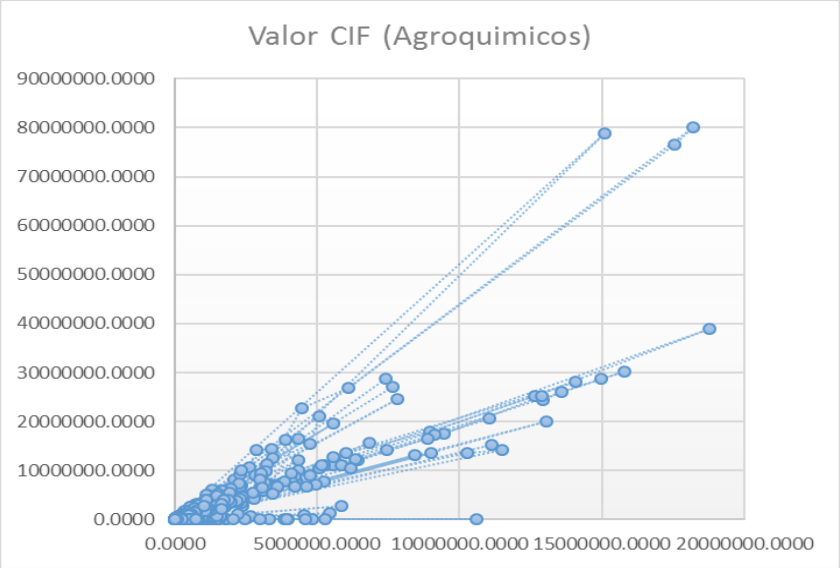


Figura 23. Relación del valor FOB y el valor FOB de agroquímicos.

En la imagen anterior se muestra la relación del valor FOB con respecto del valor CIF del producto de la familia de agroquímicos.

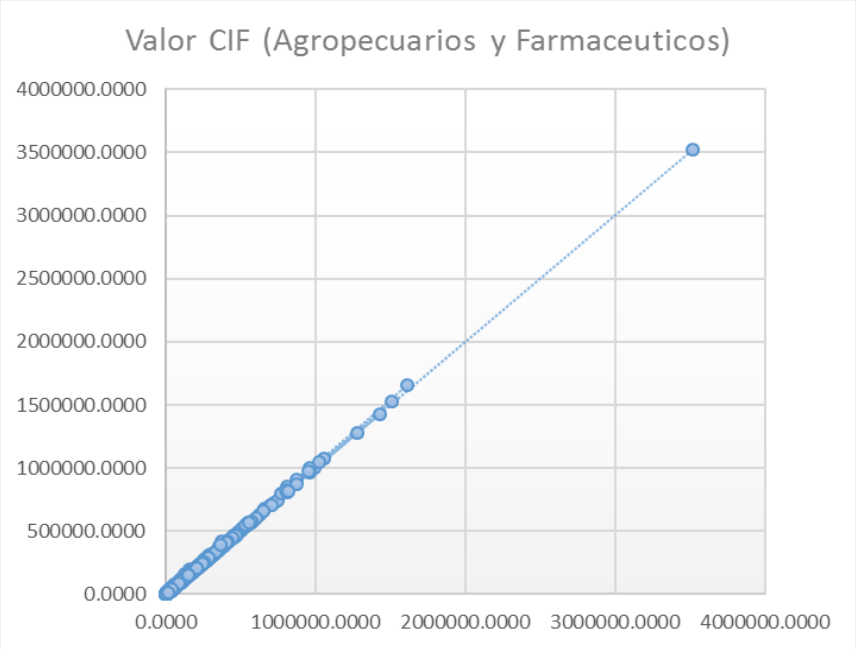


Figura 24. Relación de valor FOB y el valor CIF de agropecuarios y farmacéuticos.

En la imagen anterior se muestra la relación del valor CIF con respecto del valor FOB del producto de la familia de agropecuarios y farmacéuticos.

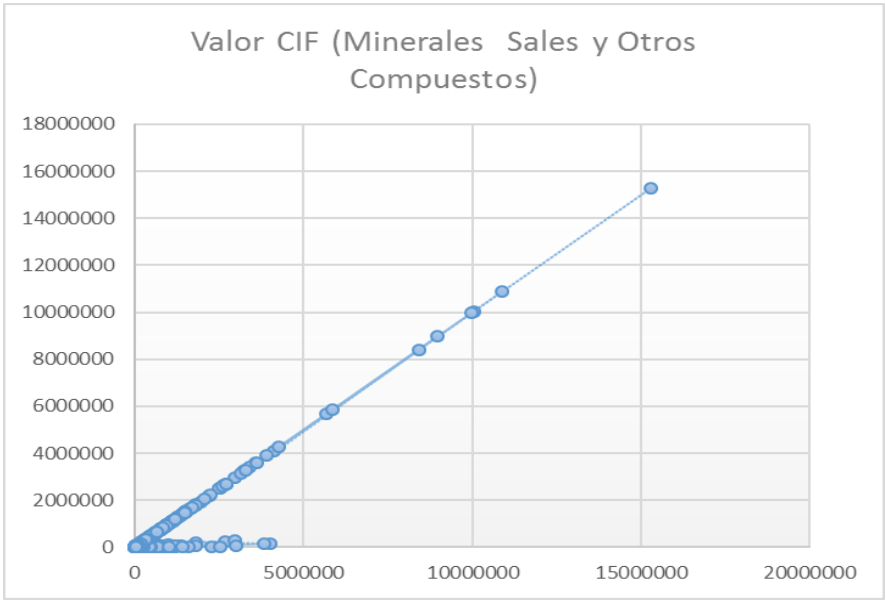


Figura 25. Relación de valor FOB y valor CIF de minerales, sales y otros compuestos.

En la imagen anterior se muestra la relación del valor CIF con respecto del valor FOB del producto de la familia de minerales, sales y otros compuestos.

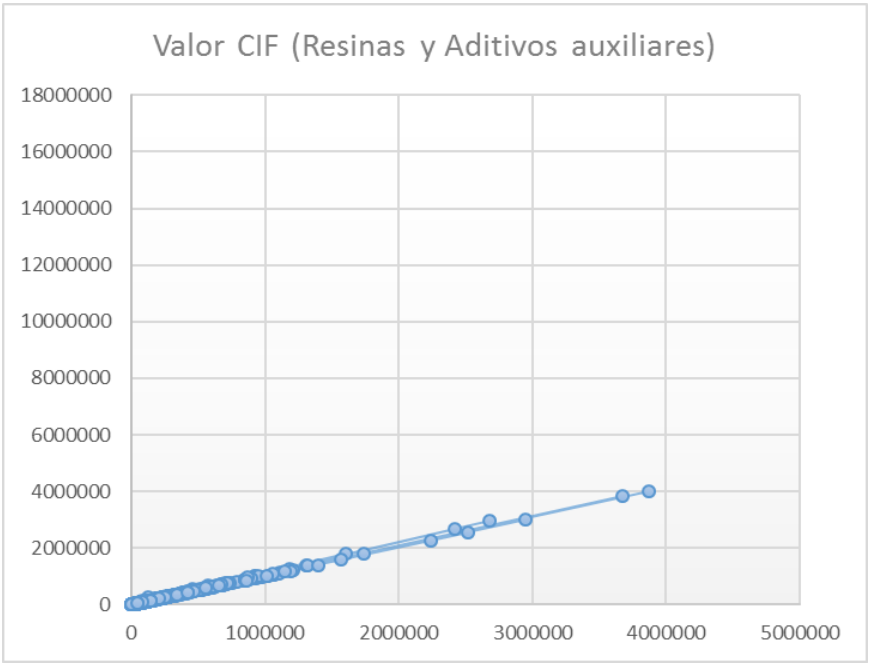


Figura 26. Relación de valor FOB y valor CIF de resinas y aditivos auxiliares.

En la imagen anterior se muestra la relación del valor CIF con respecto del valor FOB del producto de la familia de resinas y aditivos auxiliares.

Para todos los modelos anteriores se aplicará un modelo de regresión para el cálculo de la variable CIF, esto es muy importante para el negocio ya que puede facilitar el ingreso del producto una vez llega a puerto sin esperar que lleguen por aparte las facturas correspondientes al flete y al seguro del producto, pudiendo así realizar el pre-costeo del producto y su puesta a la venta en un tiempo menor al que se tiene actualmente.

5.2.1.6.1.2 Validación de datos de entrenamiento

Como parte del análisis requerido se hace primeramente una vista de la muestra de datos para esto se elige una muestra de un mes de los países involucrados teniendo la siguiente información en mente para su preparación,

- **Datos desbalanceados.** Son aquellos conjuntos de datos que naturalmente se tienden a agrupar hacia un solo grupo del conjunto total de datos por analizar y provoca un sesgo que hace que los datos no se puedan analizar si no es con alguna limpieza previa.
- **Valores extremos y errores.** Valores que pueden sesgar un análisis de información los cuales se deben de aislar en el análisis de información o tomar las medidas necesarias para que no afecte su estudio.
- **Datos perdidos y valores repetidos.** Son los valores que una determinada dupla no aparecen y hace que el análisis se vea afectado por ese efecto de no encontrar el valor. Los valores que se repiten, pero por error humano también son analizados ya que tienen una misma estructura de información.

- **Escalamiento de los datos o normalización de datos.** Es colocar los datos e integrar y analizar en una escala similar para que no exista un dato dominante dentro del conjunto de datos. Para lo cual se busca normalizar los datos.

- **Filtrado de datos.** Como se mencionó en la parte de limpieza de datos los mismo se cotejan contra un grupo de artículos de interés en el estudio, los cuales son grupos de artículos o derivados de un determinado artículo que haga ver los diferentes mercados directos o indirectos del giro de negocio de Grupo Transmerquim.

El conjunto de datos de la muestra tomará en cuenta un grupo de países del grupo total, en total se tomará 3 países, en donde se evaluará el comportamiento de enero en 3 años diferentes. Con una población total de 165.053 líneas y en 5 familias diferentes de productos.

Con esto se tiene las siguientes familias de productos por analizar.

- Aceites.
- Agropecuarios y Farmacéuticos.
- Minerales, Sales y Otros Compuestos.
- Resinas y Aditivos Auxiliares.
- Agroquímicos.

Para esto se inicia con la importación de los datos en Azure Machine Learning y se valida los datos que la herramienta de Data Quality ayudó a eliminar los datos que no eran útiles para el análisis.

El primer análisis que se realiza son los datos perdidos o sin valor sobre las columnas que se referencian, se procede a configurar un bloque de acción de Machine Learning para la limpieza de estos registros.

Figura 27. Vista preliminar de datos.

Como se muestra en la imagen anterior se tiene un total de 7897 líneas sin el valor de CIF, que nos podría desequilibrar el modelo. Para esto se procede a colocar un módulo que elimine las líneas que presentan esta característica.

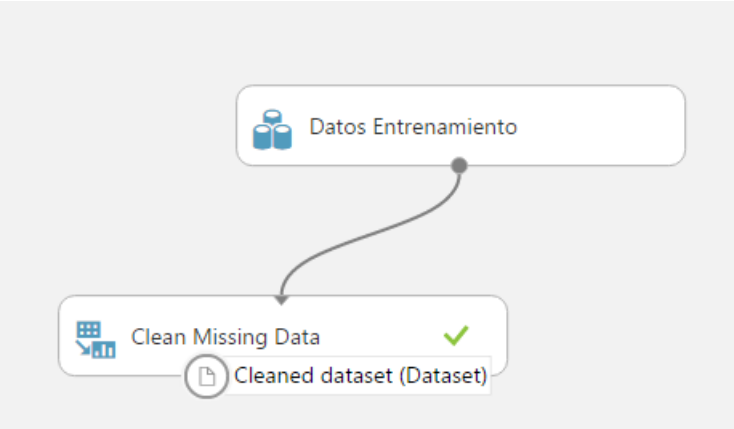


Figura 28. Módulos de Azure Machine Learning para limpieza de datos.

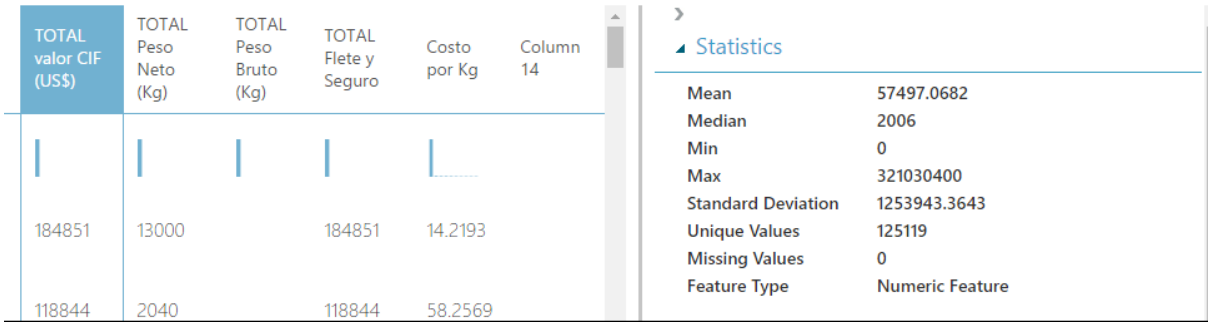


Figura 29. Datos preliminares luego de la aplicación del módulo de limpieza.

En las imágenes anteriores se muestra la aplicación de esta regla y su resultado en el conjunto de datos.

El siguiente paso es la validación de duplicados en el conjunto de datos. Para esto se procede a agregar el módulo de duplicados.

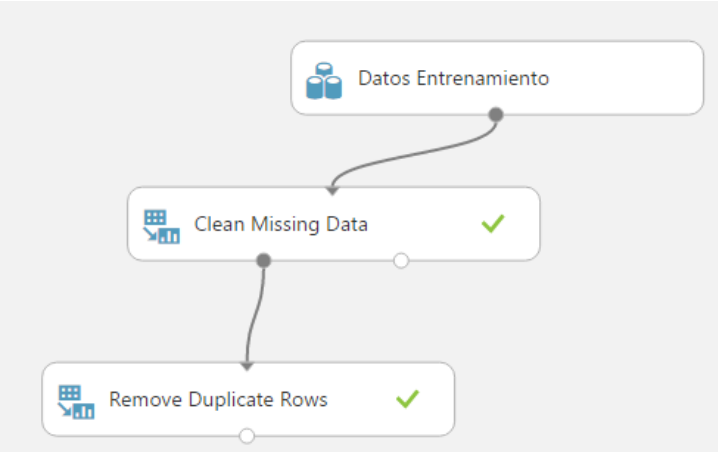


Figura 30.Inclusion del módulo que remueve duplicados al modelo.

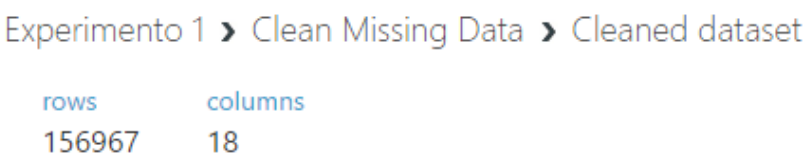


Figura 31. Valores previos a la aplicación del proceso.

En la imagen de arriba se muestra el valor de líneas antes de aplicar el proceso.

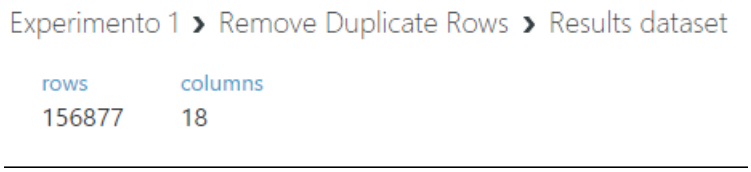


Figura 32. Valores luego de la aplicación del proceso.

En la imagen de arriba se muestra el valor de líneas luego de aplicar el proceso. Este mismo paso se realiza para las 5 diferentes familias de productos.

El siguiente paso es validar los valores extremos y los errores en la información. Para esto se procede con la agrupación de las líneas según su familia de producto para que los valores no sean extrapolados por su costo según el tipo de producto.

Como parte de la limpieza de los datos de entrenamiento se buscan los valores extremos como se muestran las imágenes siguientes de diagramas de dispersión. Sin embargo, antes de construir un modelo se debe de buscar la relación entre las columnas y la columna por predecir y así retirar los valores aislados que puedan sesgar el análisis. En este momento por la experiencia que el usuario de importación que aporta se determina que el valor por predecir es el valor CIF. Para esto se tiene como variables relacionadas las siguientes,

- Valor FOB del producto.
- Peso neto del producto.
- Peso bruto del producto.
- Flete y seguro de la carga
- Costo por kilogramo

Se procede a hacer el análisis por familia ya que los modelos, no necesariamente aplican para todas las familias de igual forma.

5.2.1.6.2 Interpretación y Análisis de resultados

En esta etapa se persigue la exploración de los datos con visualización, entender las relaciones de los datos, crear múltiples vistas de los datos, usar condiciones y comprender la fuente de los errores.

Para evitar sesgos de información, los análisis e interpretaciones se harán por familia de productos.

5.2.1.6.2.1 Aceites

5.2.1.6.2.1.1 Análisis de Valores Iniciales y Relaciones

Se muestra el valor CIF del grupo de productos seleccionado y sus estadísticas.

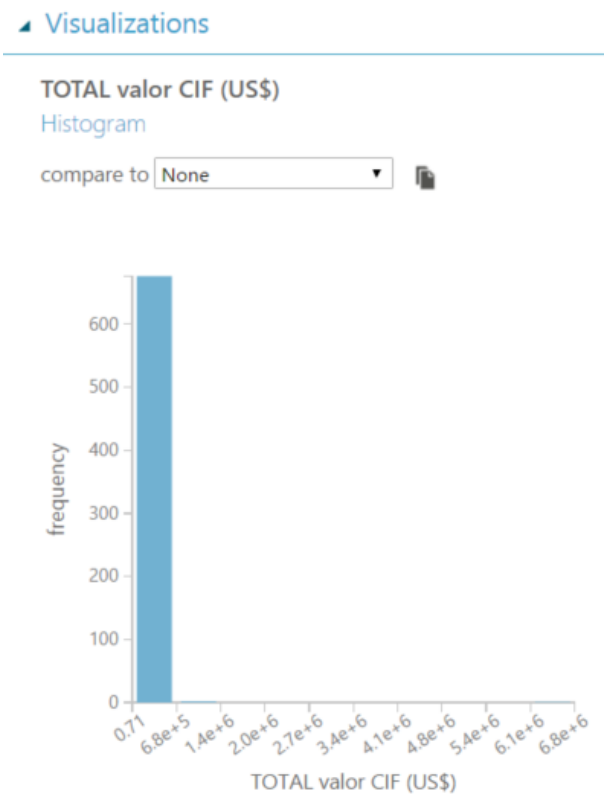


Figura 33. Valores iniciales en aceites

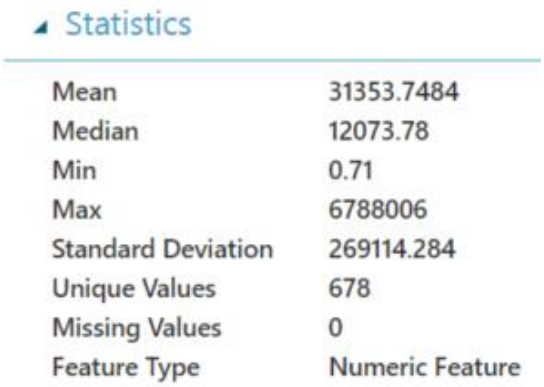


Figura 34. Estadísticas iniciales en aceites.

Las imágenes muestran como los valores extremos hacen que la información se muestre sesgada y sin sentido alguno. El valor mínimo está muy por debajo de la media y la mediana del grupo de datos y el valor máximo muy por encima de los mismos valores.

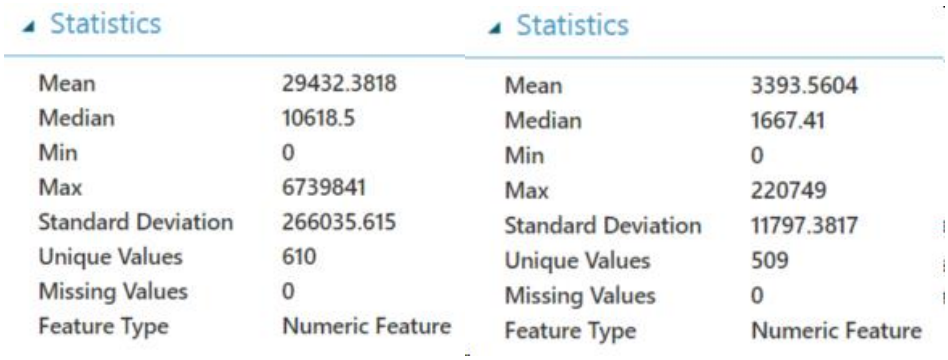


Figura 35. Valores iniciales de FOB y kilogramos

En las imágenes se muestran las estadísticas de las otras dos variables, se hacen visibles valores muy similares de sesgo en ambos extremos del conjunto de datos. Los datos de la izquierda corresponden la variable de FOB y la imagen de la derecha a la variable de peso en kilogramos.

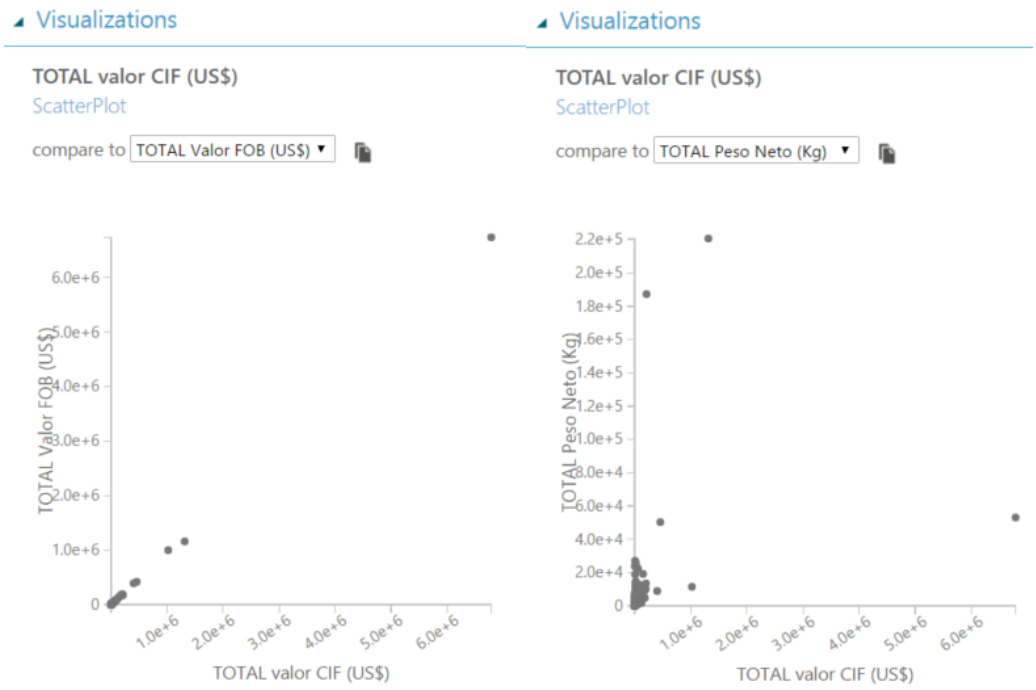


Figura 36. Relación de datos entre valor CIF y el valor FOB, así como el valor CIF y kilogramos.

La imagen de la derecha muestra la relación del CIF con la variable FOB, los datos se agrupan en un solo conjunto muy definido, pero con un registro fuera del grupo, que podría ser un error o un dato extremo. De igual forma el gráfico de dispersión de la derecha muestra la relación entre el valor CIF y la variable de peso en kilogramos, con un grupo definido y valores extremos fuera del mismo.

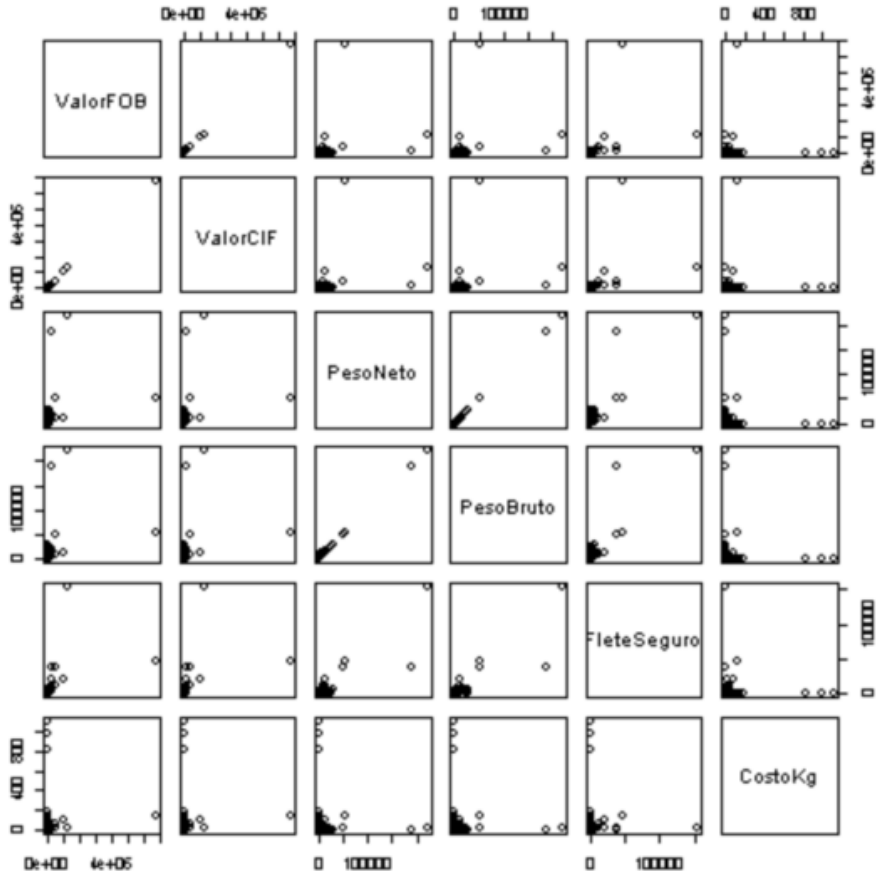


Figura 37.Intercección de variables en la familia de aceites.

En la imagen se muestra los valores y sus intersecciones con las demás variables previo a la limpieza de los valores extremos.

5.2.1.6.2.1.2 Limpieza de Valores Extremos

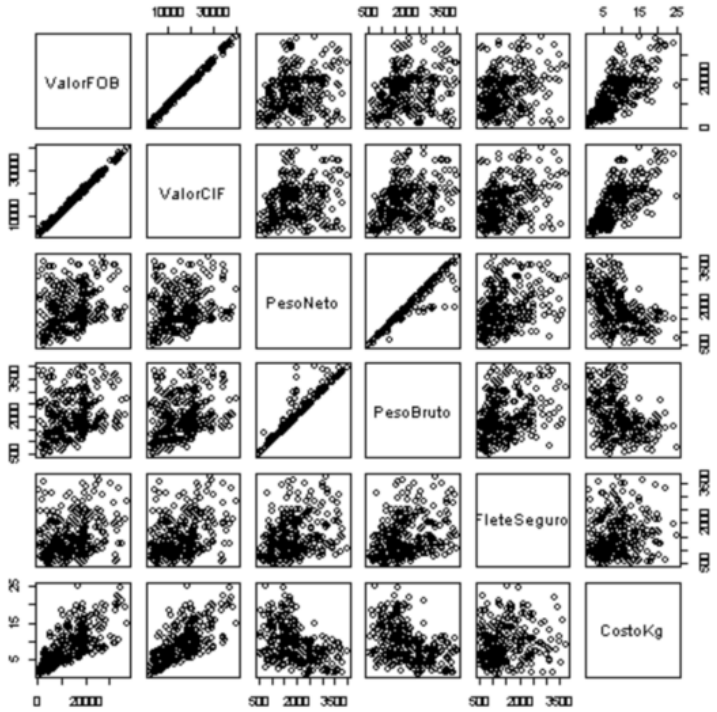


Figura 38. Intercesión de variables luego de la limpieza.

La imagen muestra los gráficos de dispersión luego de la limpieza de valores extremos, módulo ejecutado con código R en Azure Machine Learning.

En el Anexo 6, se muestra el código en R utilizado para el filtrado de los datos extremos y sus estadísticas luego de la normalización. Además del código utilizado para visualizar datos y sus sesgos.

5.2.1.6.2.1.3 Normalización de los Datos

El siguiente paso es la normalización de los datos al utilizar el método de mínimos y máximos.

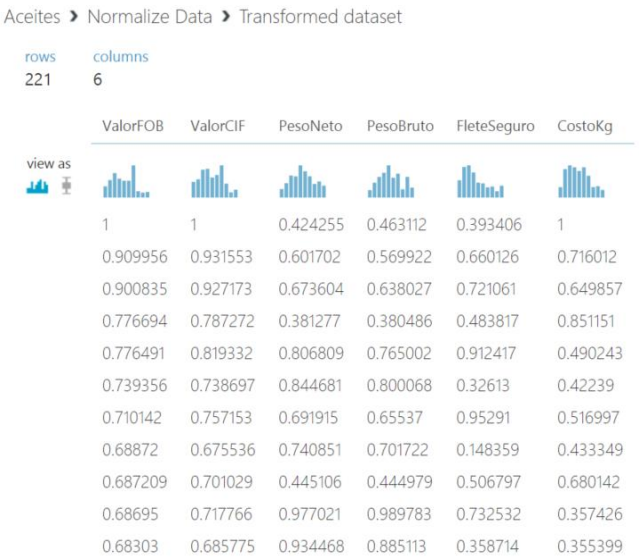


Figura 39. Normalización de datos.

En la imagen se muestra las columnas luego de la normalización de los datos al aplicar el método de mínimos y máximos. Proceso indispensable en la creación de los modelos de predicción.

Una de las partes más importantes del análisis son las relaciones de las variables con respecto del objetivo por predecir, para esto se ejecutan en el proyecto diferentes módulos de R para poder comprender mejor la información.

Figura 40. Módulos de Azure machine Learning para la visualización de datos.

En la imagen se muestra parte de los módulos creados en Azure Machine Learning para la visualización de los datos y su mejor comprensión. La salida del último script de R nos muestra información básica con relación de los datos.

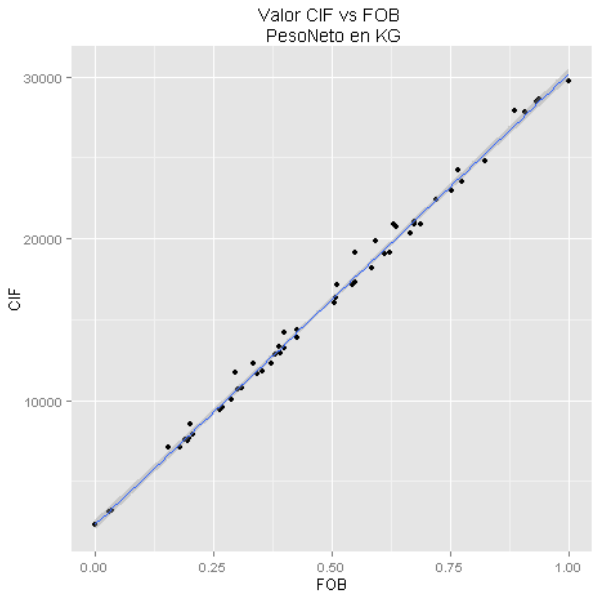


Figura 41. Relación de CIF y FOB en un gráfico de dispersión.

La imagen muestra la relación entre la variable CIF y la variable FOB en un gráfico de dispersión.

5.2.1.6.2.1.4 Técnicas de Observación

Se procedió a crear una categoría del peso neto y categorizarlo en cuatro categorías, las que se pueden llamar,

- Carga baja.
- Carga media baja.
- Carga media alta.
- Carga alta.

Tomando en cuenta esta relación de categorías de peso con respecto del valor de la carga, se pueden encontrar cosas interesantes del análisis de la familia de los aceites.

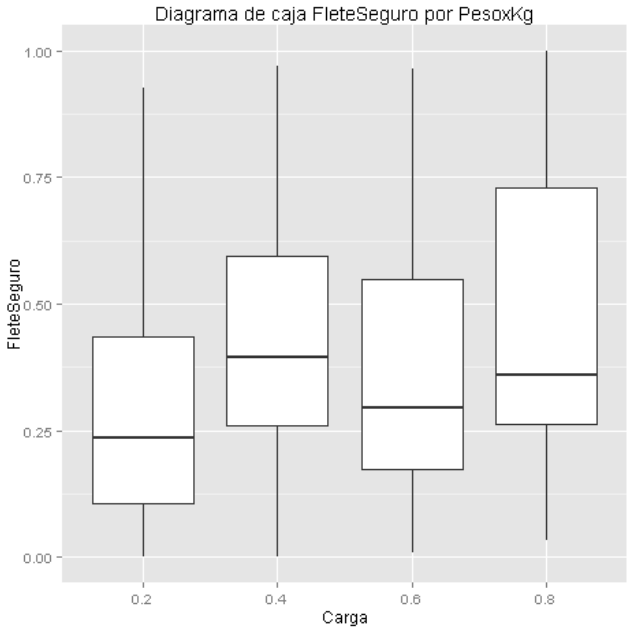


Figura 42. Diagrama de caja de flete y seguro por peso en kilogramos.

En la imagen anterior se muestra como el costo por flete y seguro de la cuarta categoría, llamada carga alta, puede tomar montos bajos de flete y seguro, esta zona es la que se llamara la ideal para hacer el traslado del producto. Y en la que como parte del desarrollo del proyecto debe ir arrojando luz sobre el entendimiento y relaciones de datos.

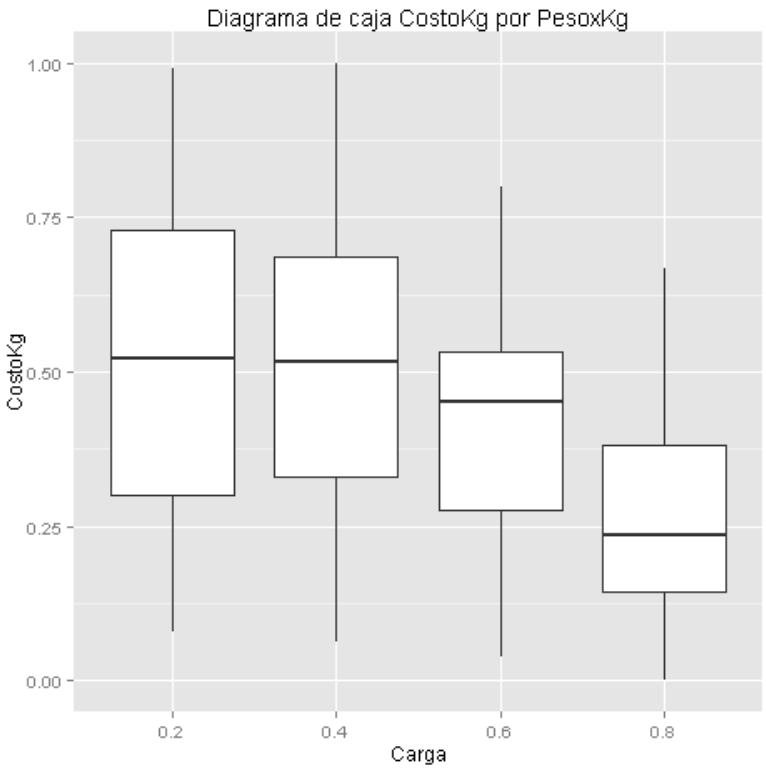


Figura 43. Relación de carga y el costo por kilogramo.

En la imagen se muestra como la carga es inversamente proporcional al costo por kilogramos, algo que no era nuevo lo que sí es nuevo es que con esta categorización se puede observar que existe una porción de carga baja que se entrelaza con el costo de carga alta, esto quiere decir que en los casos que la carga sea aun baja, se puede obtener precios por kilogramo bajos.

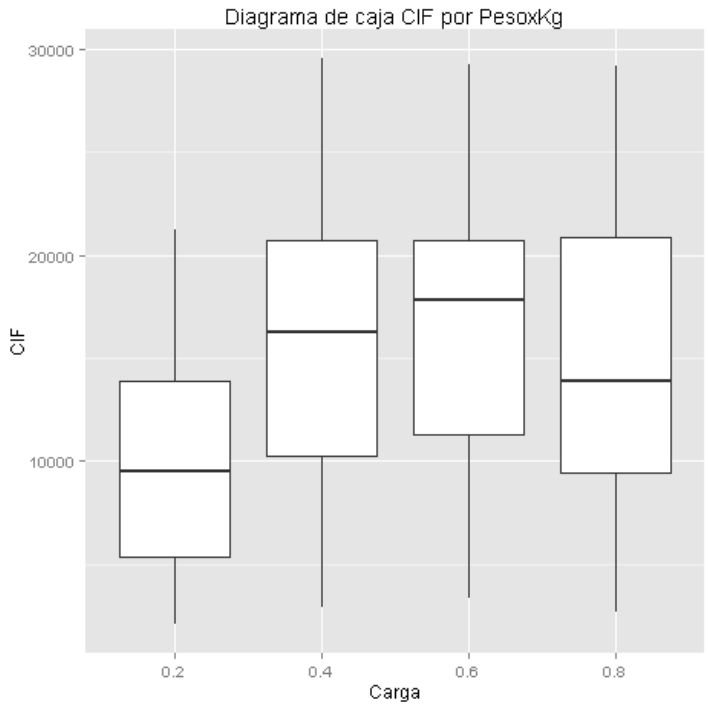


Figura 44. Relación de la carga con respecto al valor CIF.

Por último, se observa en la imagen anterior la relación entre el peso de la carga y el valor CIF del grupo de familias, la misma muestra como un volumen alto de productos pueden tener un costo bajo, esto muestra que aún se puede buscar mejores precios de productos y que la media aún se puede mejorar con volúmenes altos, o sea que la concentración de pedidos del grupo de empresas es algo que aún puede dar mejores resultados en cuanto al precio total de la carga en esta familia de productos.

En el Anexo 7 se puede observar el total de relaciones de las variables con respecto de la variable por predecir en esta familia de productos.

5.2.1.6.2.1.5 Regresión Polinomial de las Variables

Al continuar con el análisis de la dependencia de las variables en la predicción del nuevo valor, para esto se recurre a la regresión polinomial de las

variables. Sin querer entrar en la discusión de aplicación de los diferentes métodos se procede a una aplicación polinomial en las variables dependientes, con respecto de la variable por predecir. Cuando se vuelve a la formulación inicial se tiene que las variables dependientes se puedan comportar de la siguiente forma,

$$\mathcal{F}(x) = \beta_0 + \beta_1 X$$

Aunque en nuestro caso el ajuste de la forma polinomial se debe de ver de la siguiente forma,

$$\mathcal{F}(x) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots\ldots \beta_n X^n$$

Esta técnica se aplica para mejorar el proceso de cálculo de la variable predictible, no es el fin del estudio explicar este proceso más allá de un tema de cálculo de variables y su relación y de cómo estas nuevas columnas cúbicas y al cuadrado ayudan a hacer la distribución más útil de aquellas variables que en su función base no se ajusta a la variable por predecir. Este proceso nos ayuda a reducir el error en el conjunto de entrenamiento y la complejidad del modelo.

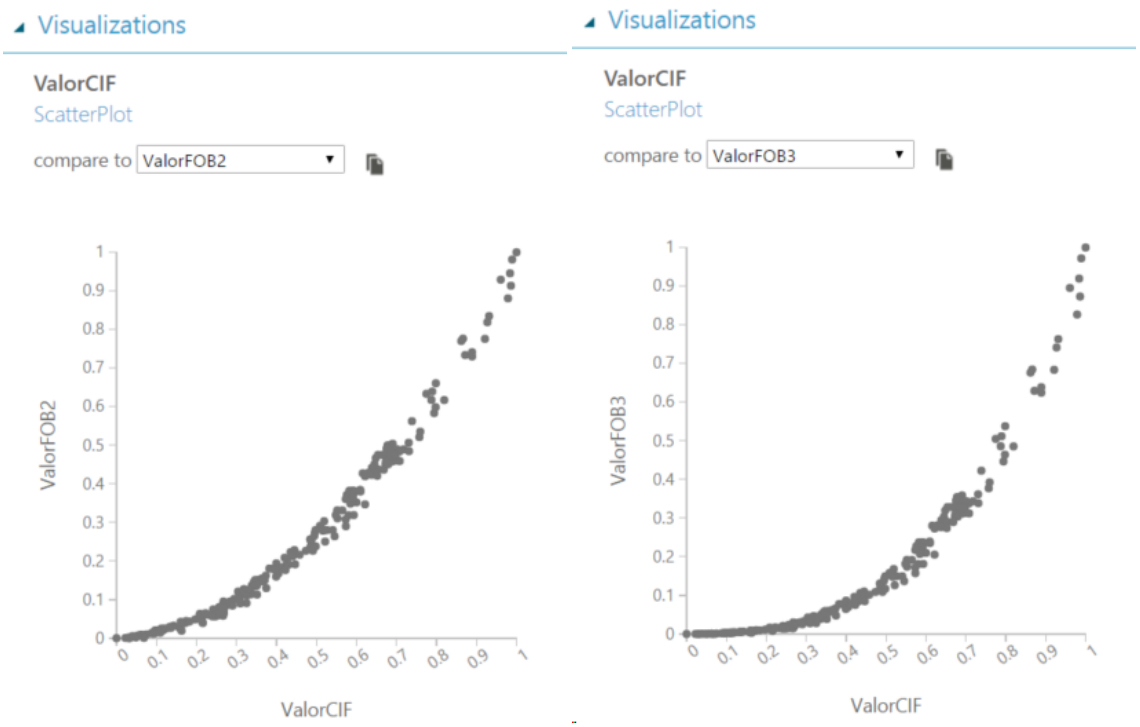


Figura 45. Intersección de valores polinomiales

En las imágenes se observan los efectos de las varíales polinómicas sobre la variable por predecir. En el Anexo 8 se puede observar el código usado para crear esta función polinomial.

5.2.1.6.2.1.6 Evaluación del Modelo

Luego de esto, se está listo para la creación del modelo de entrenamiento para lo cual se agregan los respectivos módulos en Azure Machine Learning. Como se muestra en la siguiente imagen.

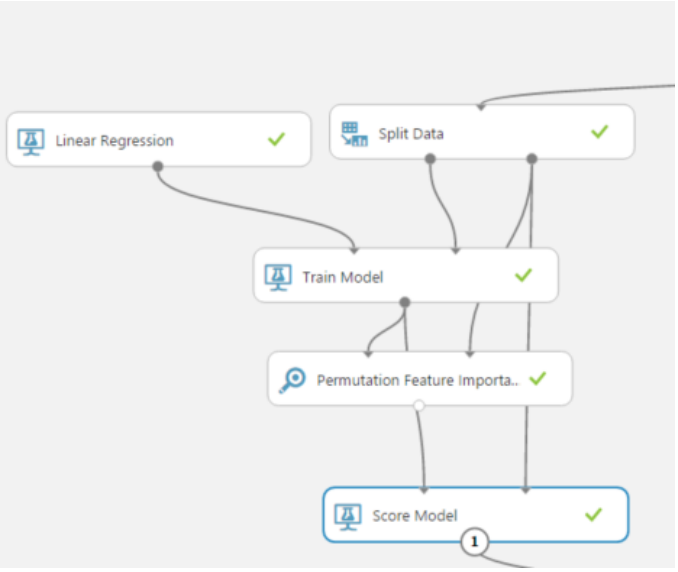


Figura 46. Módulos de Azure Machine Learning para regresión lineal.

La primera validación en la ejecución del modelo es el análisis de las variables escogidas para la predicción. Para esto en el módulo de “Train Model” se puede observar las estadísticas de correlación en las variables dependientes y la variable a predecir.

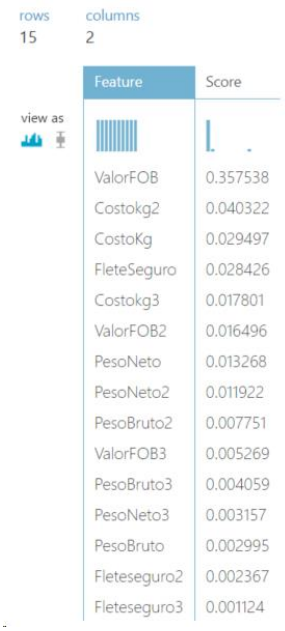
Feature Weights

Feature	Weight
ValorFOB	1.03674
Costokg2	0.105465
CostoKg	-0.0806078
FleteSeguro	0.0738581
ValorFOB2	-0.0510822
Costokg3	-0.0505551
PesoNeto	-0.0417112
PesoNeto2	0.0379241
PesoBruto2	0.0255489
ValorFOB3	0.0183138
PesoBruto3	-0.0148739
PesoNeto3	-0.0121213
PesoBruto	-0.0112699
Fleteseguro2	0.00734558
Fleteseguro3	-0.00478659

Figura 47. Peso de las variables en la predicción de la variable.

En la imagen anterior se puede observar los pesos de las diferentes variables en la predicción de la variable.

Se puede observar de una forma más sencilla en el módulo “Permutation Feature Importance” de Azure, como lo muestra la imagen siguiente.


rows 15 columns 2

Feature	Score
ValorFOB	0.357538
Costokg2	0.040322
CostoKg	0.029497
FleteSeguro	0.028426
Costokg3	0.017801
ValorFOB2	0.016496
PesoNeto	0.013268
PesoNeto2	0.011922
PesoBruto2	0.007751
ValorFOB3	0.005269
PesoBruto3	0.004059
PesoNeto3	0.003157
PesoBruto	0.002995
Fleteseguro2	0.002367
Fleteseguro3	0.001124

Figura 48. Peso de las variables en el cálculo de variable dependiente.

Por último, se valida la salida del algoritmo seleccionado y los resultados de predicción de la variable respectiva, parte esencial del desarrollo de esta segunda etapa y del proyecto. Para esto se evalúa a través del módulo “Score Model” de Azure Machine Learning al tener como resultado lo siguiente. Se seleccionan las dos columnas solamente para su análisis más cómodamente.

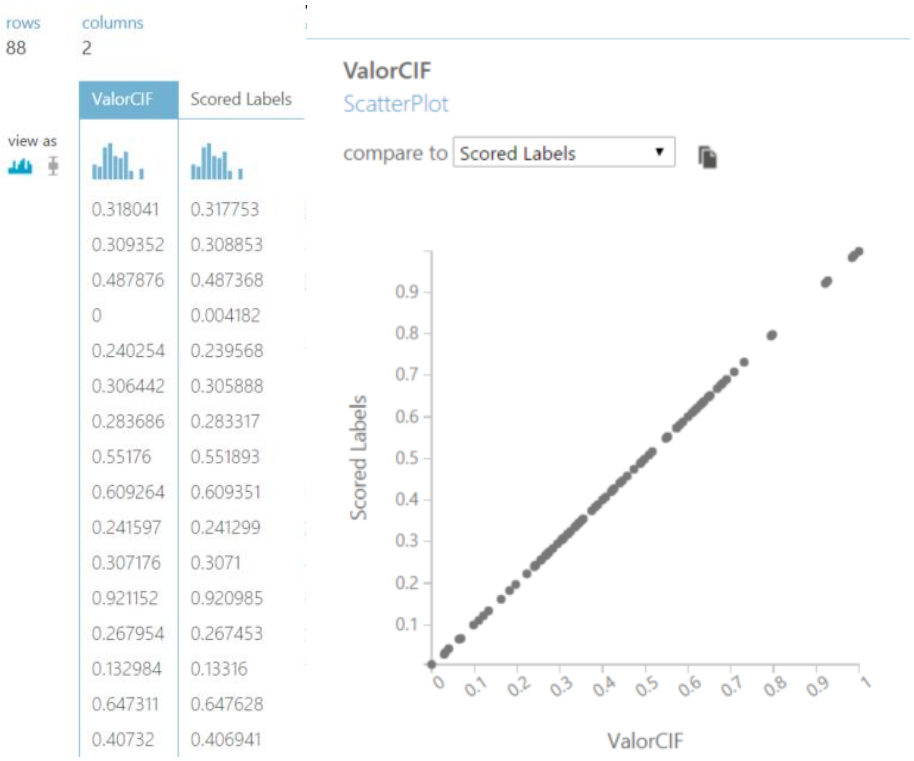


Figura 49. Resultados de la variable predicha y el valor original.

Como se observa en la imagen la columna “Scored Labels” tiende a ser muy parecida a la variable por predecir Valor CIF, lo que nos demuestra que la selección de variables es la acertada y el algoritmo respectivo responde a lo requerido de forma correcta. También el gráfico de dispersión nos muestra la correspondencia de los valores predichos con el valor de prueba. Una forma más tangible de evaluar el modelo es colocar el módulo de Azure machine Learning, “Evaluate Model”,

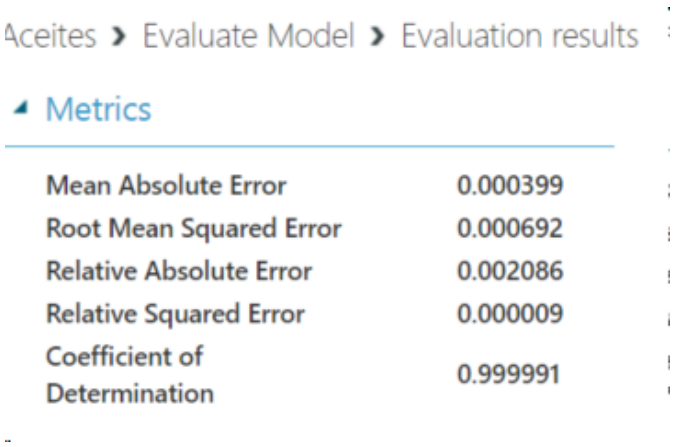


Figura 50. Estadísticas de predicción

En general las estadísticas son muy buenas sobre el modelo y se nota principalmente en dos estadísticas, el coeficiente de determinación y el error cuadrático relativo. El coeficiente de determinación es una medida de la reducción de la variación, entre la variable por predecir y el error del modelo. Esta estática se refiere a menudo como R^2. Un modelo perfecto tendría un coeficiente de determinación de 1.0. Mientras que el error cuadrático relativo es el cociente de la varianza o error cuadrado del modelo, dividido por la varianza de los datos. Un modelo perfecto tendría un error cuadrático relativo de 0.0. Con esto se puede decir que el modelo cumple con lo requerido. La imagen nos muestra en la parte inferior estos dos valores en el primer caso el valor de coeficiente de determinación es muy cercano a 1 y en el segundo caso el error cuadrático relativo tiende a cero.

5.2.1.6.2.2 Agropecuarios y Farmacéuticos

5.2.1.6.2.2.1 Análisis de Valores Iniciales y Relaciones

Se selecciona el valor CIF del conjunto de entrenamiento y se analizan los valores, luego de realizar la limpieza de valores perdidos y duplicados dentro de la herramienta.

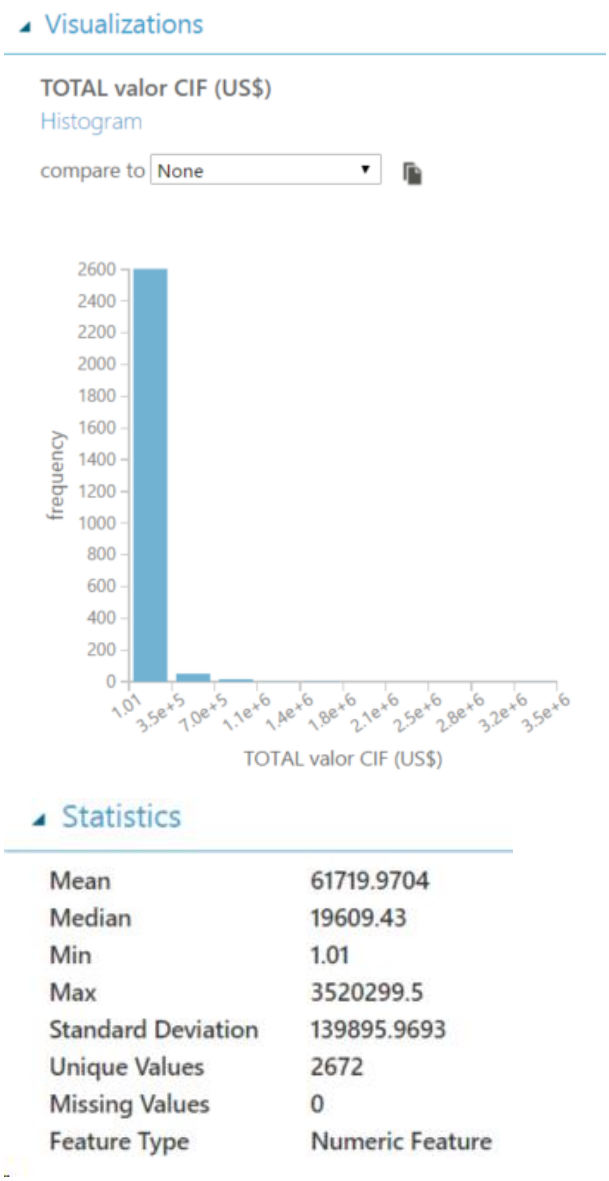


Figura 51. Valores iniciales de los datos de agropecuarios.

La imagen muestra valores máximos y mínimos muy por encima de la media y la mediana lo que provoca una desviación estándar muy elevada para el grupo de artículos en análisis. Se determinará un rango de valores máximos y mínimos y se eliminarán los extremos para este estudio.

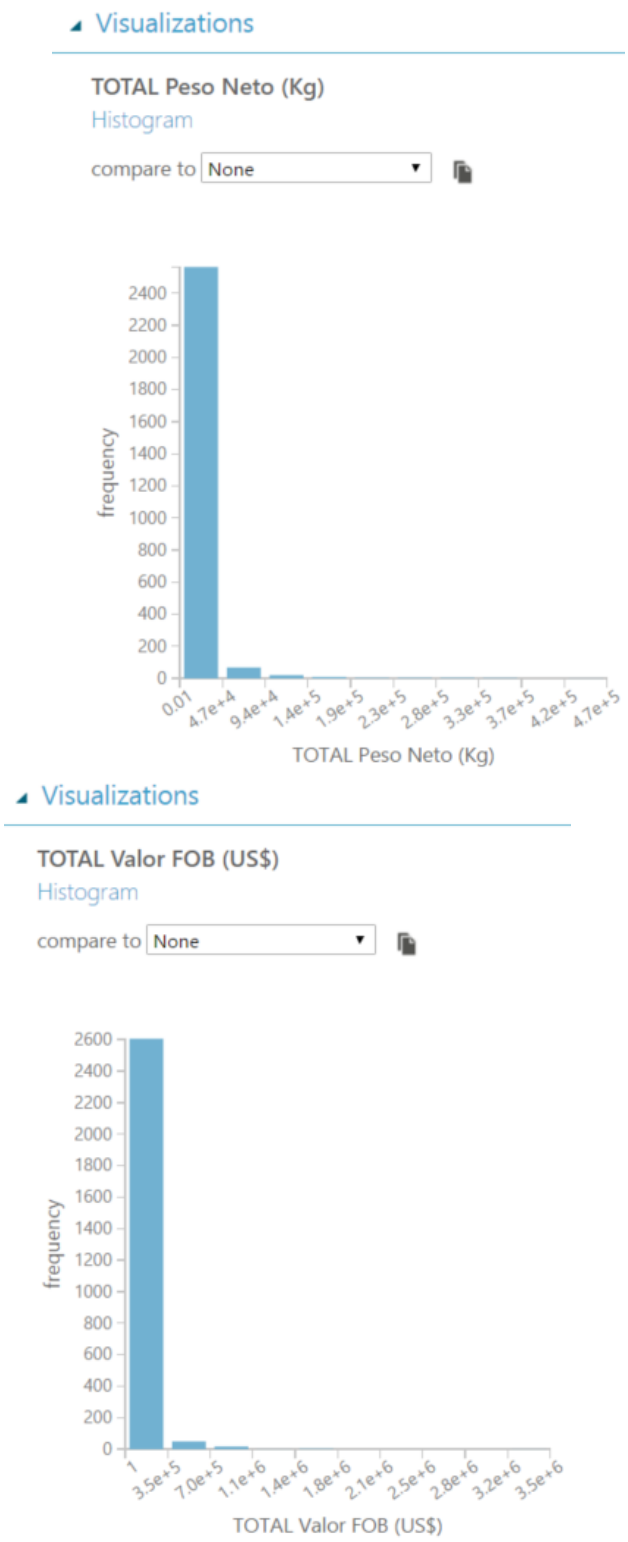


Figura 52. Muestreo de datos iniciales de agropecuarios.

En la imagen de la izquierda se muestra el comportamiento similar de la variable FOB, así como la imagen de la derecha muestra un comportamiento similar del total del peso en kilogramos.

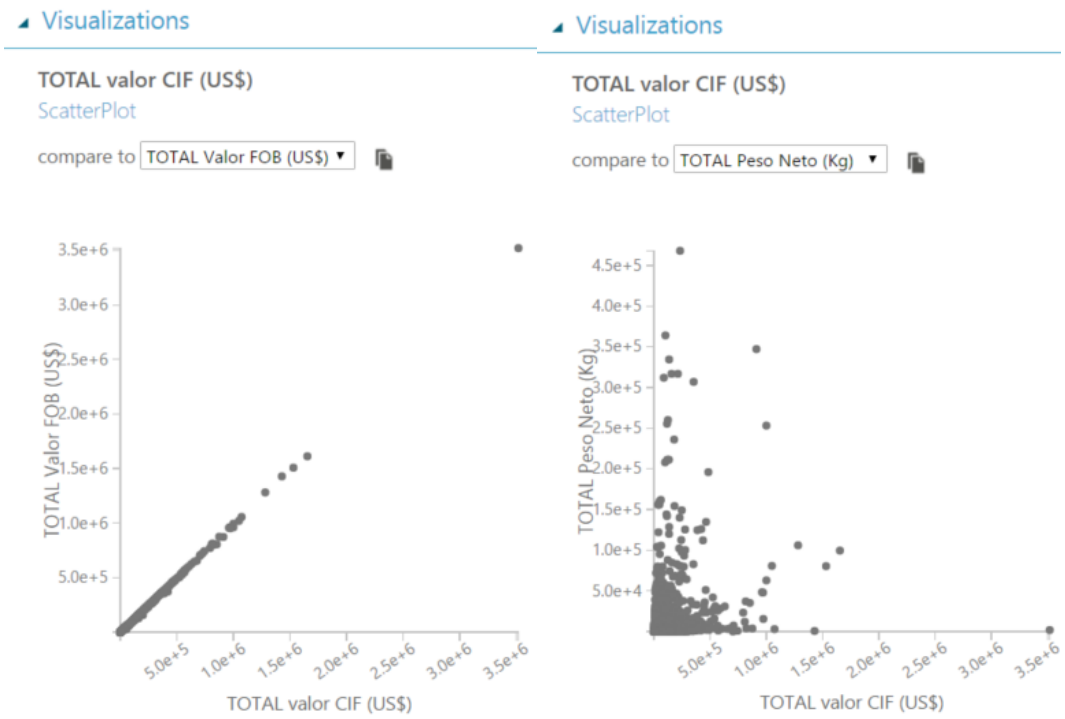


Figura 53. Relaciones de CIF y el valor FOB y peso en kilogramos.

Las imágenes anteriores muestran la relación que existen entre el valor FOB y el peso en kilogramos con el valor CIF a predecir.

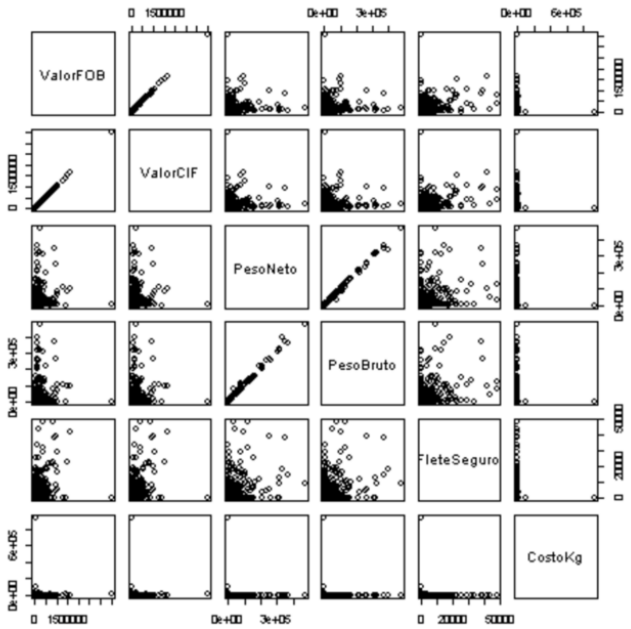


Figura 54. Relaciones de las variables en un gráfico de dispersión.

La imagen anterior muestra los gráficos de dispersión previos a la ejecución de los procesos de limpieza de los valores extremos.

5.2.1.6.2.2.2 Limpieza de Valores Extremos

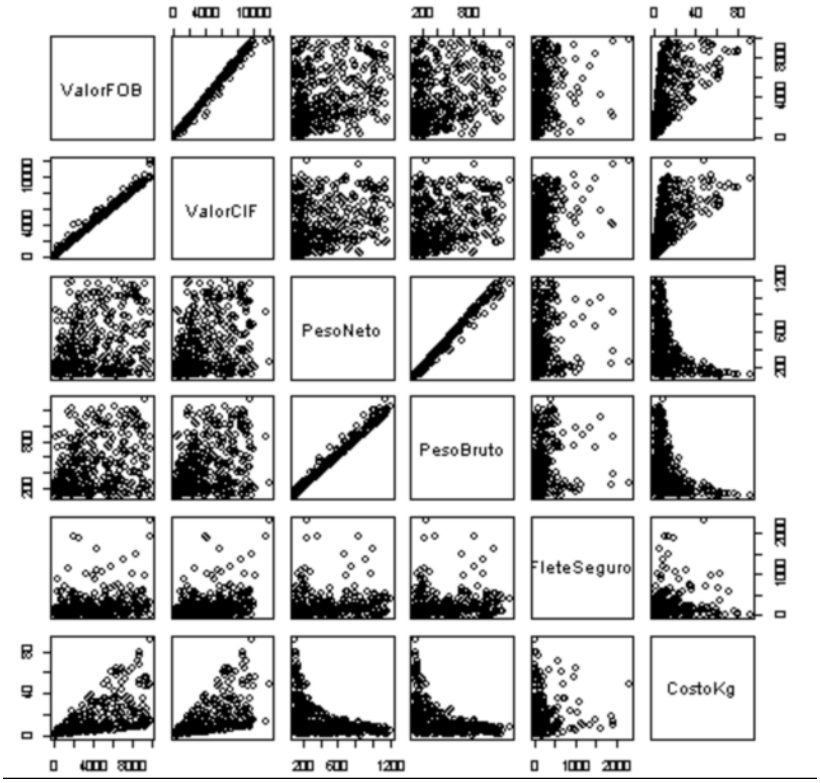


Figura 55. Relación de variables luego de aplicar limpieza de datos.

La imagen muestra los resultados luego de la ejecución del módulo en lenguaje R, y con esto poder cotejar los resultados con respecto de la anterior imagen.

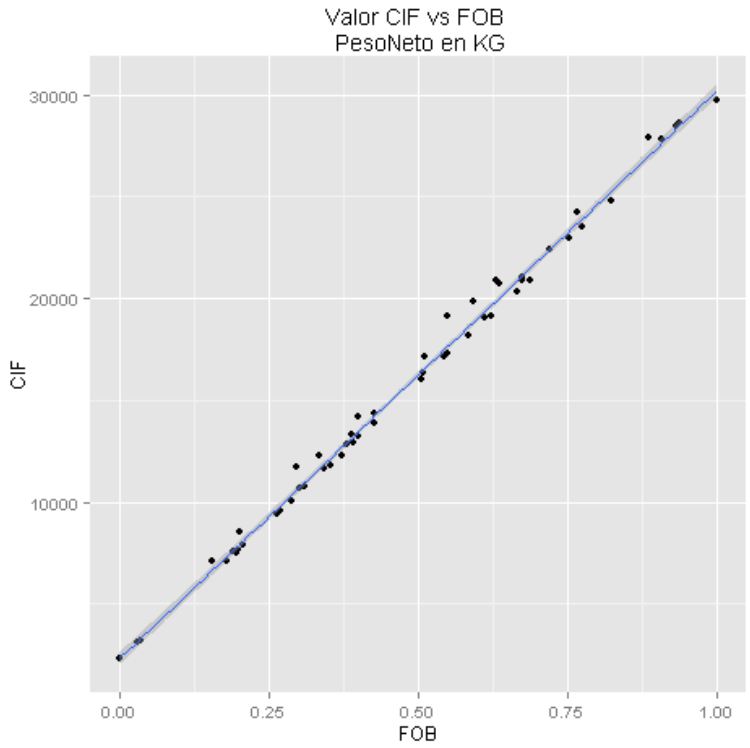


Figura 56. Relación de CIF y FOB luego de la limpieza.

La imagen muestra la relación entre el valor CIF y el valor FOB en la familia de agropecuarios y farmacéuticos.

5.2.1.6.2.2.3 Normalización de los Datos

Se realiza la normalización de los datos al utilizar el método de mínimos y máximos.

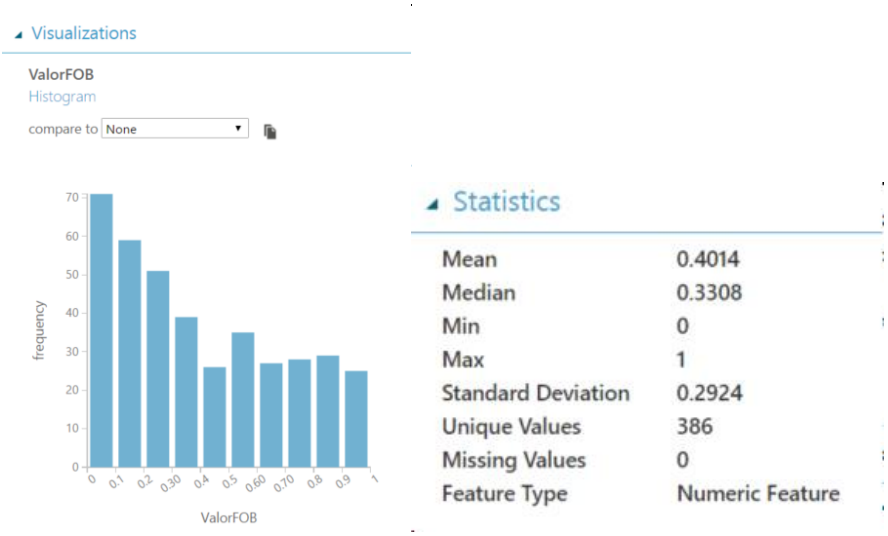


Figura 57. Resultados de la normalización de los datos.

En las imágenes de arriba se muestran el valor FOB luego de la normalización de los datos.

5.2.1.6.2.2.4 Técnicas de Observación

Se procedió a crear una categoría del peso neto y categorizarlo en cuatro categorías, las que se pueden llamar,

- Carga baja.
- Carga media baja.
- Carga media alta.
- Carga alta.

Tomando en cuenta esta relación de categorías de peso con respecto del valor de la carga, se puede encontrar cosas interesantes del análisis de la familia de los agropecuarios y farmacéuticos.

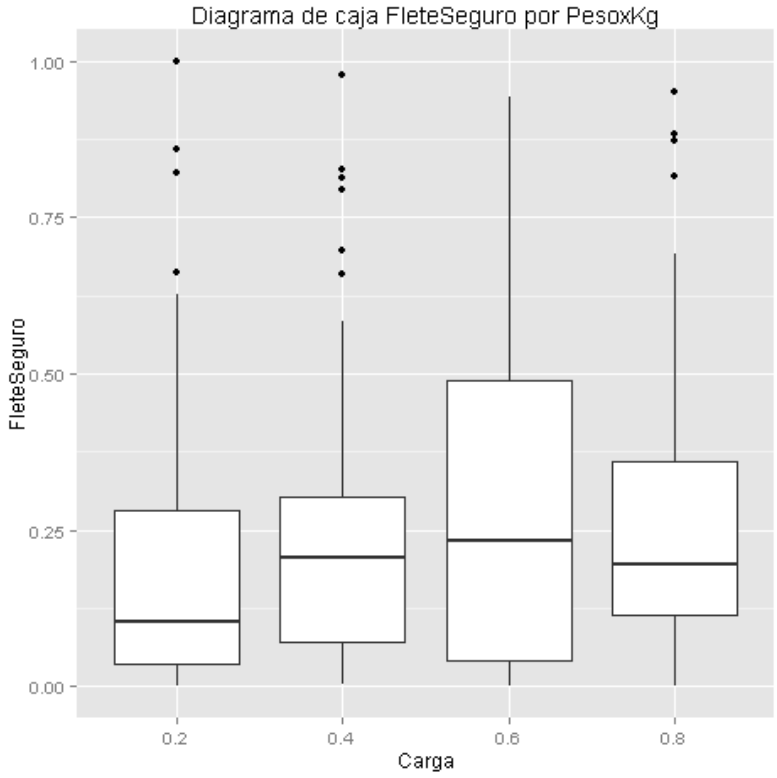


Figura 58. Diagrama de caja, relación de flete y seguro con la carga.

La imagen muestra como un porcentaje de carga alta se ubica dentro del rango de montos de fletes medios bajos, o que pueden distinguir como el valor por kilogramo puede mejorar, aun en cargas altas.

5.2.1.6.2.2.5 Regresión Polinomial de las Variables

Como se explicó en la anterior familia esta técnica ayuda a una distribución normal de las variables de entrada o dependientes.

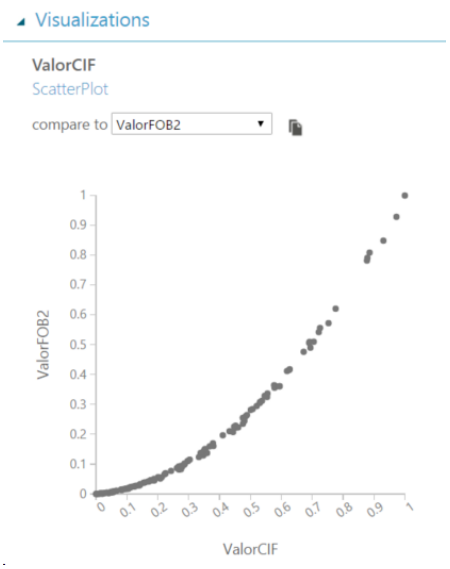


Figura 59. Relación entre los valores polinomiales y la variable por predecir.

En la imagen anterior se muestra el efecto que produce la normalización en la variable dependiente de Valor FOB.

5.2.1.6.2.2.6 Evaluación del Modelo

En este punto se limitará a hacer un resumen de los resultados de los modelos seleccionados y aplicados al grupo de artículos.

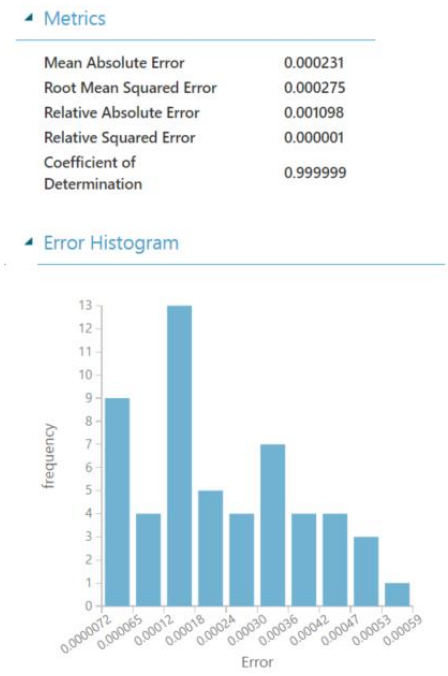


Figura 60. Resultado de la aplicación de algoritmo de regresión.

En la imagen se muestra las estadísticas luego de la ejecución del algoritmo de regresión lineal, modelo que se ajusta a la perfección para este grupo de datos. Lo que continúa afirmando que este tipo de predicción para estos conjuntos de datos es el que mejor resultados da.

5.2.1.6.2.3 Agroquímicos

5.2.1.6.2.3.1 Análisis de Valores Iniciales y Relaciones

Se selecciona el valor CIF del conjunto de entrenamiento y se analiza los valores, luego de realizar la limpieza de valores perdidos y duplicados dentro de la herramienta.

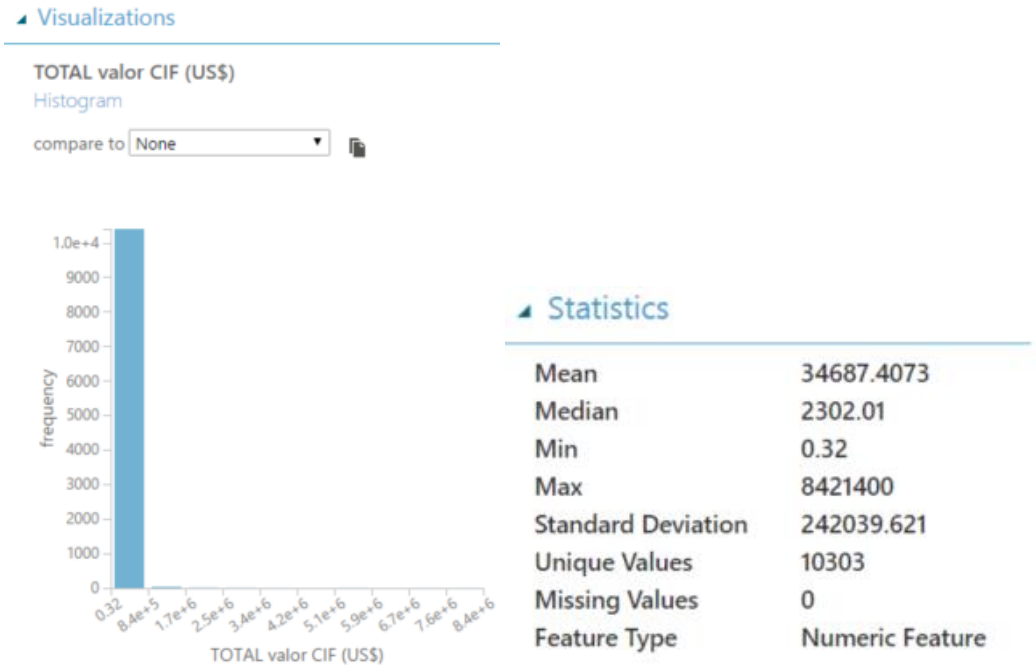


Figura 61. Valores máximos y mínimos del grupo de datos.

La imagen muestra valores máximos y mínimos muy por encima de la media y la mediana lo que provoca una desviación estándar muy elevada.

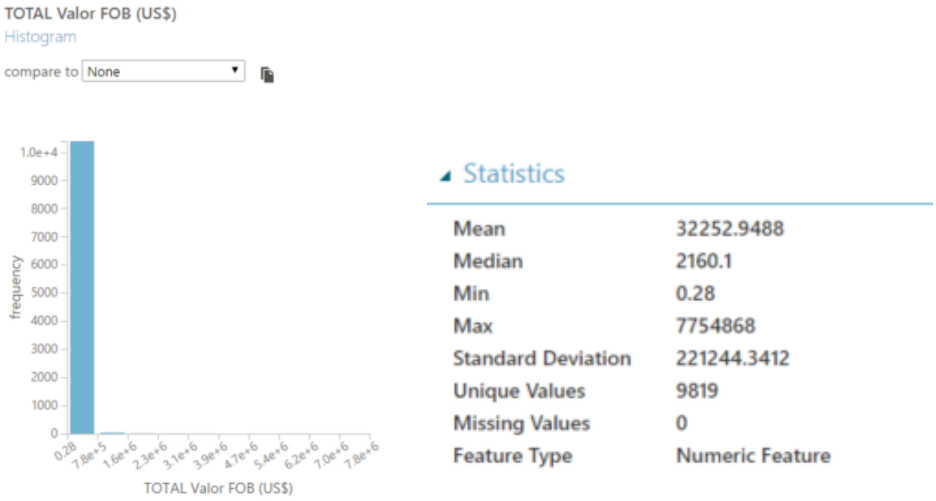


Figura 62. Estadísticas de los valores iniciales.

En la imagen ose observa como la gran mayoría de valores se agrupan en un solo tracto cerca del 99%, pero existe un grupo de menos del 1% que se salen de ese tracto. A estos valores se designarán valores extremos que serán eliminados del análisis. El siguiente paso es validar las estadísticas del grupo de datos que se nos presenta y los máximos y mínimos muy por encima y debajo de la media y la mediana del grupo.

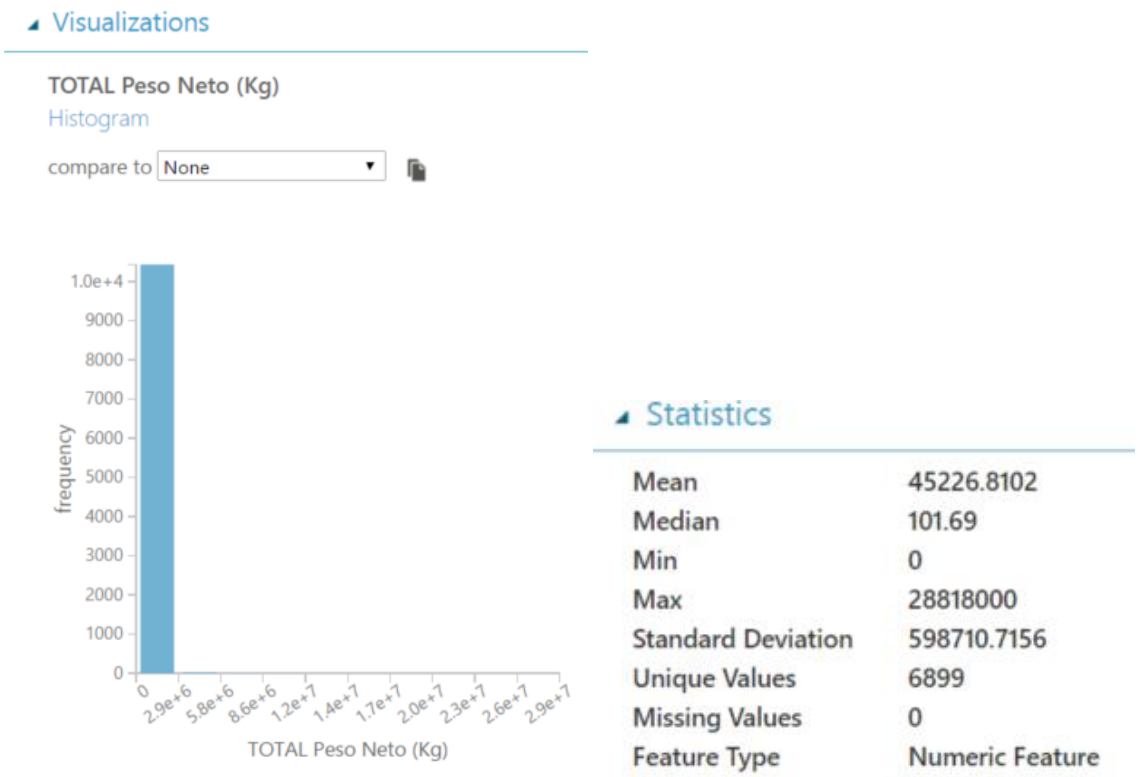


Figura 63. Valores y estadísticas del peso neto.

En cuanto al peso total de este grupo de productos, se eliminan los valores que están fuera del rango del análisis, para esto al igual se cotejan los valores mínimos y máximos y los valores de la media y la mediana. Como lo muestra la imagen anterior.

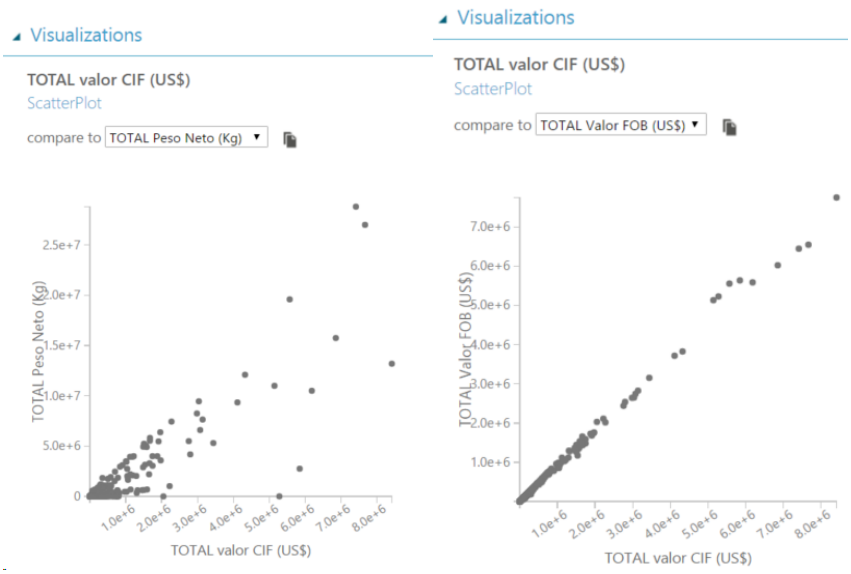


Figura 64. Relación del valor CIF y los valores FOB y peso neto.

En la imagen de la derecha se muestran como los valores CIF junto con los valores peso total se agrupan en el sector izquierdo inferior, sin embargo, algunos registros se distancian hacia la derecha, que como se menciona pude que sean correctos, pero para nuestro análisis este sesgo se trata de evitar. En la imagen de la izquierda se muestra la comparación del valor CIF junto con el valor FOB lo que muestra es un incremento casi lineal, pero de igual manera valores extremos que se eliminan para el estudio.

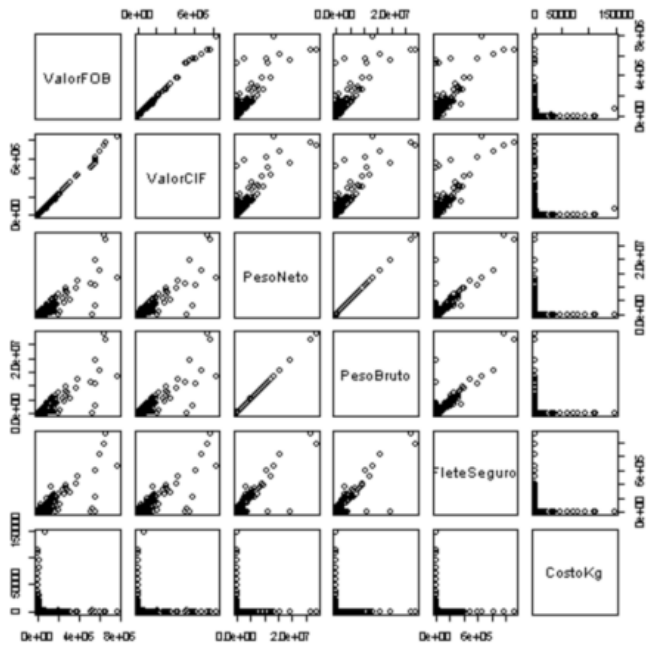


Figura 65. Relación de las diferentes variables.

La imagen muestra la interacción de totas las variables entre si y muestra los valores extremos que se presentan también.

5.2.1.6.2.3.2 Limpieza de Valores Extremos

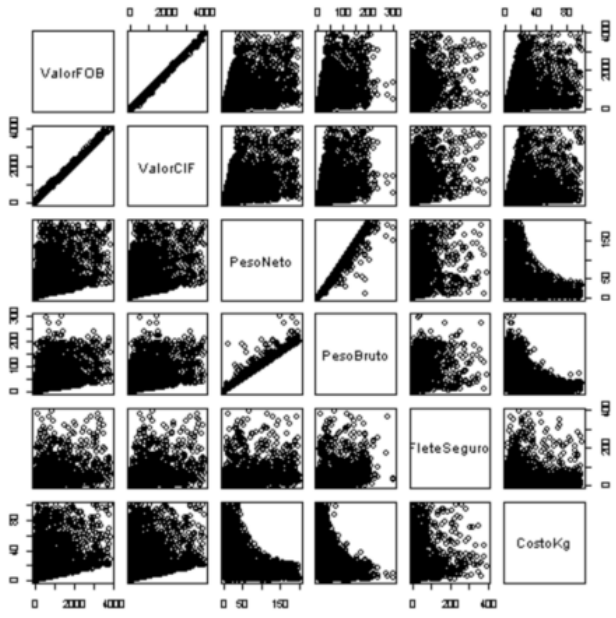


Figura 66. Valores luego de aplicar limpieza de datos.

La imagen muestra los resultados luego de la ejecución del módulo en lenguaje R, y con esto para poder cotejar los resultados con respecto a la anterior imagen. Donde se muestra más densa la imagen y los registros que se mostraban lejos del grupo principal han desaparecido.

5.2.1.6.2.3.3 Normalización de los Datos

Se realiza la normalización de los datos utilizando el método de mínimos y máximos.

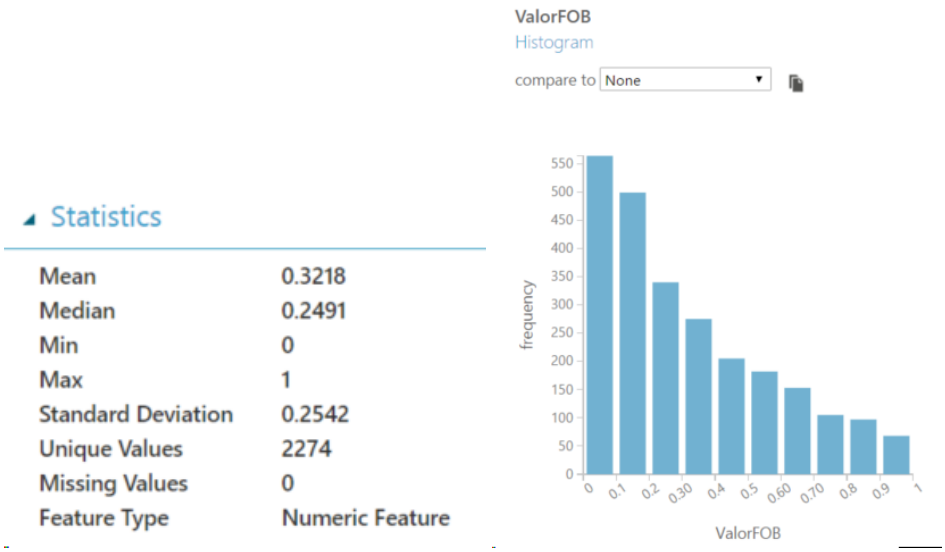


Figura 67.Normalizacion de datos, familia agroquímicos.

En la imagen se muestra el proceso que resulta de la normalización de las variables. Los resultados son similares a las demás familias de productos.

5.2.1.6.2.3.4 Evaluación del Modelo

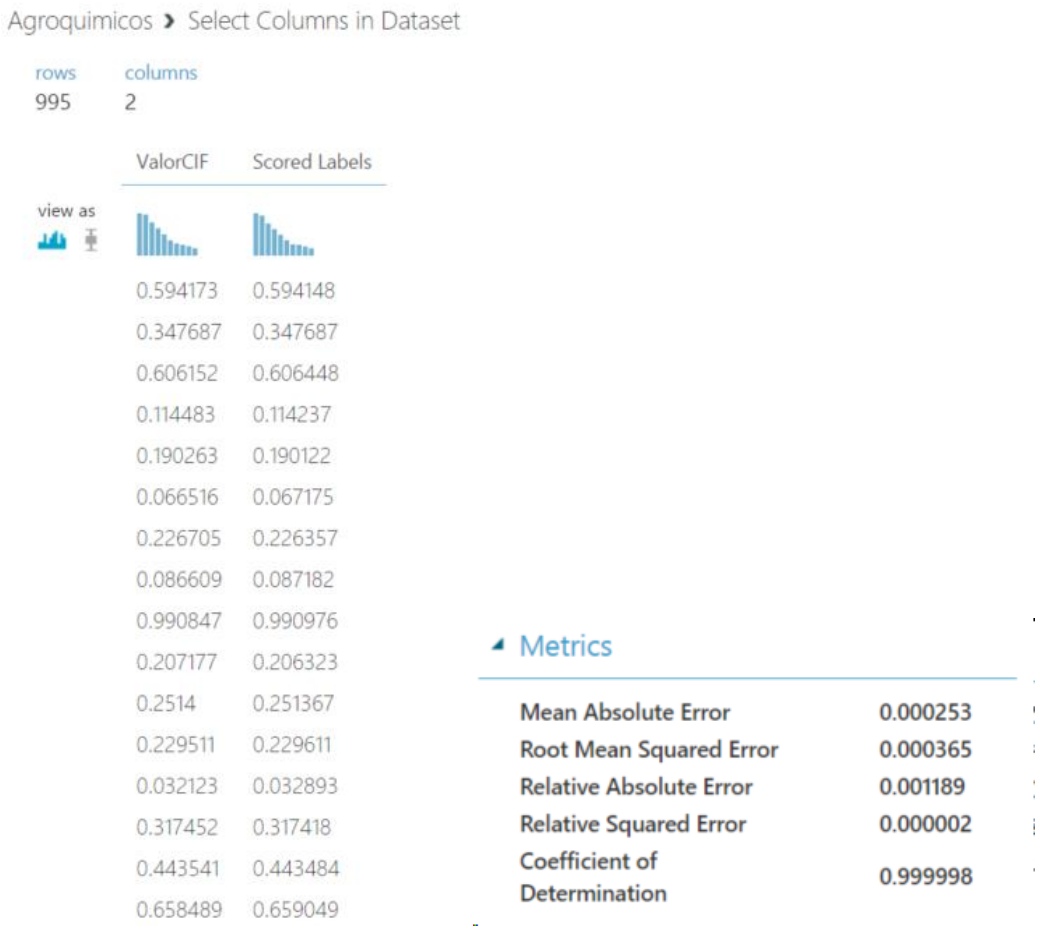


Figura 68. Estadísticas resultantes de agroquímicos.

En la imagen se muestra lo efectivo de la parametrización y la elección del modelo de regresión lineal para este grupo de datos. Las estadísticas del modelo confirman que el coeficiente de determinación tiende a 1.0 y el error cuadrático relativo tiende a 0.0.

5.2.1.6.2.4 Minerales, Sales y Otros Compuestos.

5.2.1.6.2.4.1 Análisis de Valores Iniciales y Relaciones

Se muestra el valor CIF en la muestra de este grupo de productos.

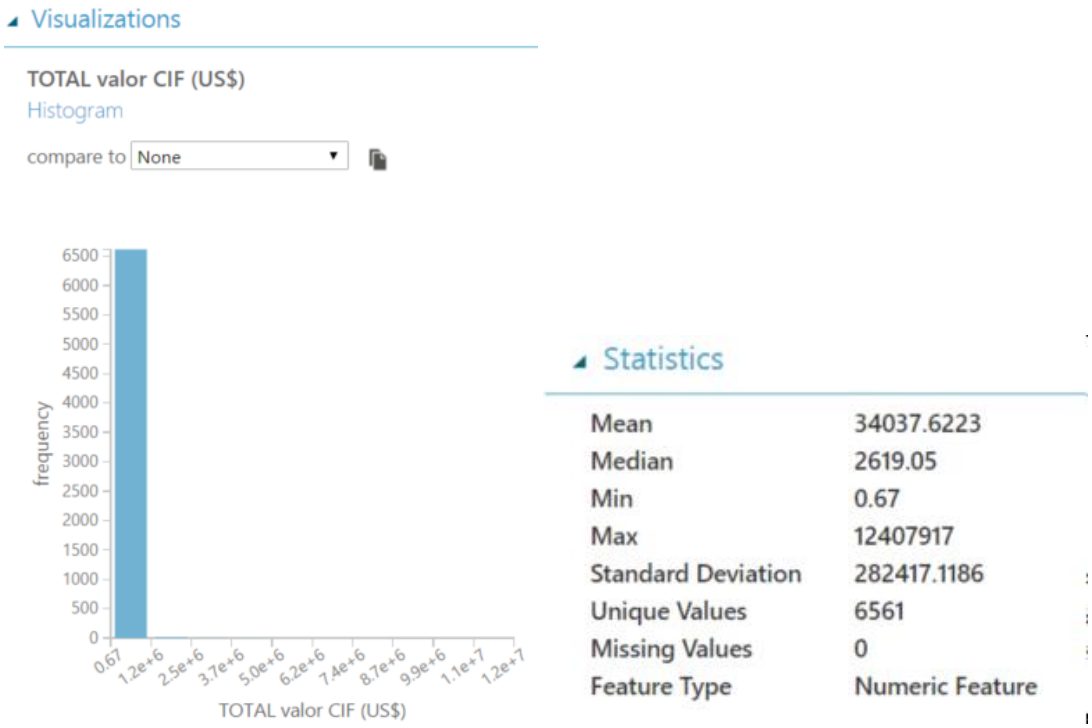


Figura 69. Valores iniciales del valor CIF en minerales, sales y otros.

Las imágenes muestran como los valores extremos perturban el conjunto de datos, del total del grupo existen cerca de 17 valores extremos, lo que provoca una desviación estándar tan alta. Esto se corrobora cuando se validan los valores mínimos y máximos con la media y la mediana.

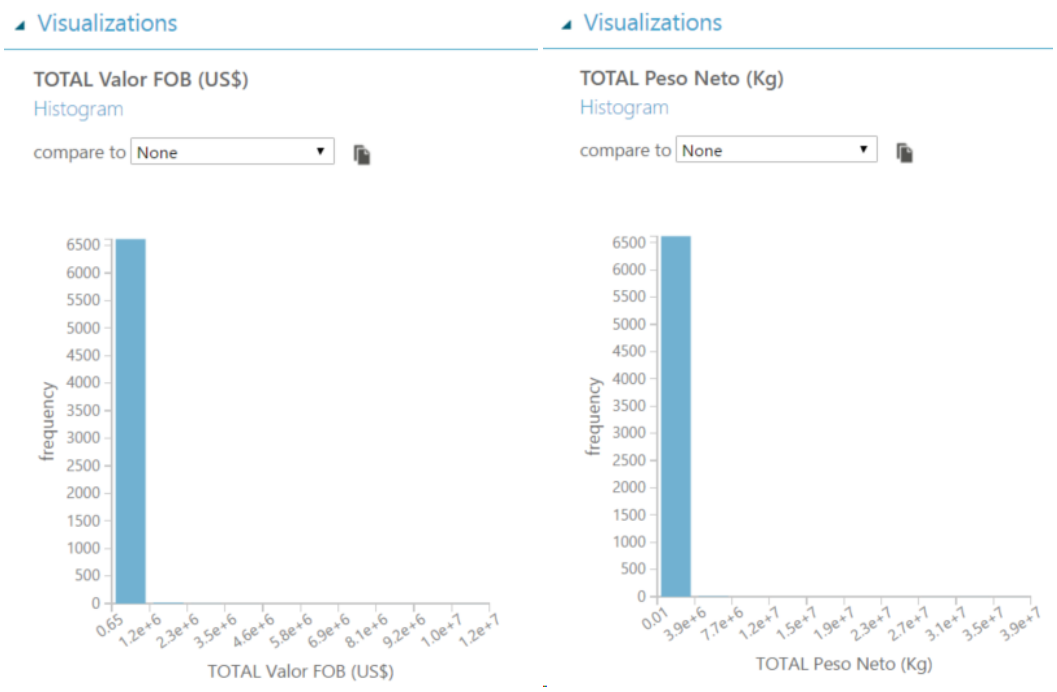


Figura 70. Estadísticas iniciales de FOB y peso neto.

La imagen de la izquierda muestra la distribución del valor FOB en donde se encuentra que del total 17 registros se salen de este grupo, mientras que la imagen de la derecha muestra como la distribución similar, pero en este caso la variable peso en kilogramos tienen un total de 10 registros fuera del rango principal.

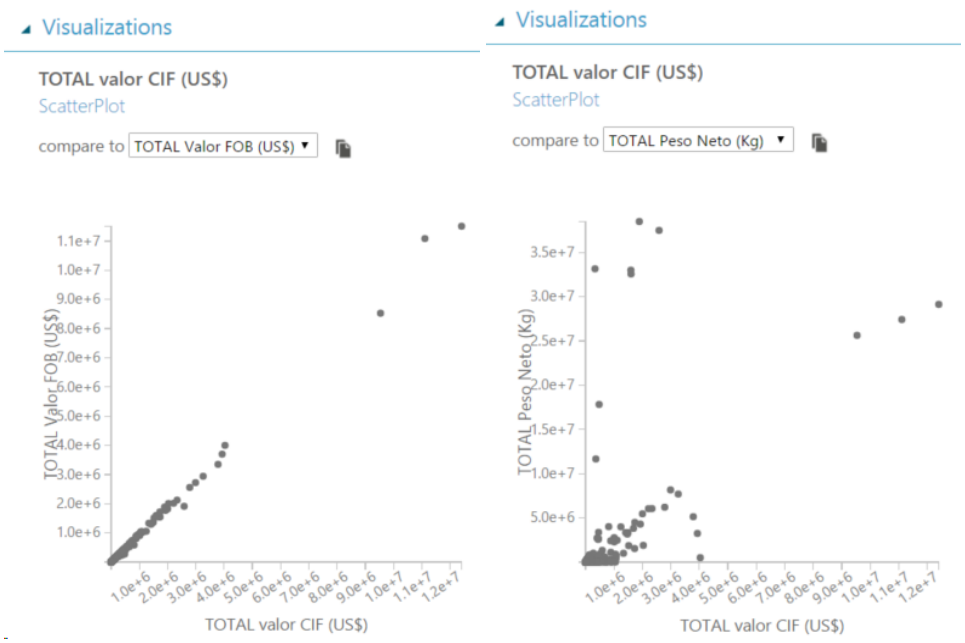


Figura 71. Relación de los valores CIF y los valores de FOB y peso neto.

La imagen de la izquierda muestra la relación de la variable CIF con la variable FOB, se pueden observar los valores extremos que acompañan esta relación. En la imagen de la derecha, la relación entre el valor CIF y el valor de peso en kilogramos se observa una concentración en la parte inferior izquierda y varios registros fuera de este grupo, que serán eliminados para este estudio.

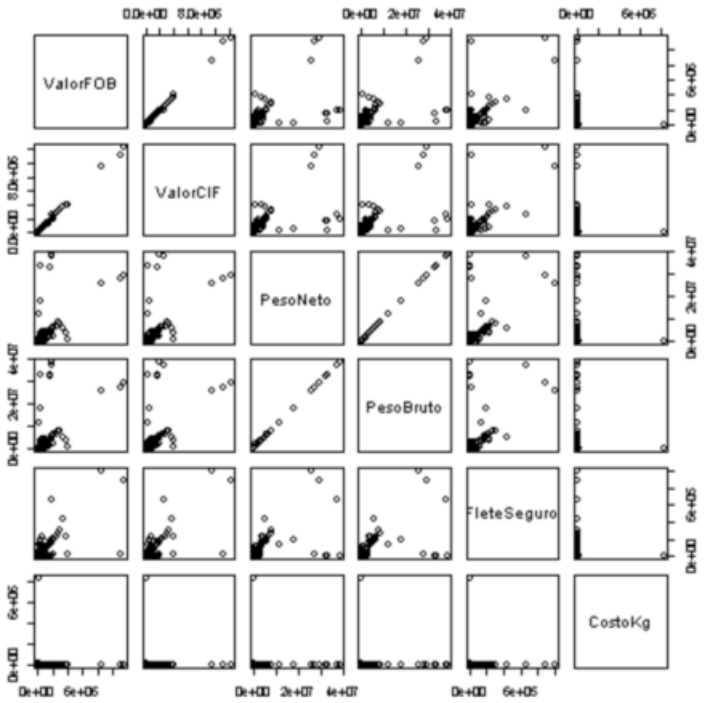


Figura 72. Relación de las variables en un gráfico de dispersión.

La imagen muestra como los valores extremos se comportan de forma diversa en las diferentes comparaciones de las variables. El grafico de dispersión muestra como esto podría provocar un sesgo en el análisis de la información.

5.2.1.6.2.4.2 Limpieza de Valores Extremos

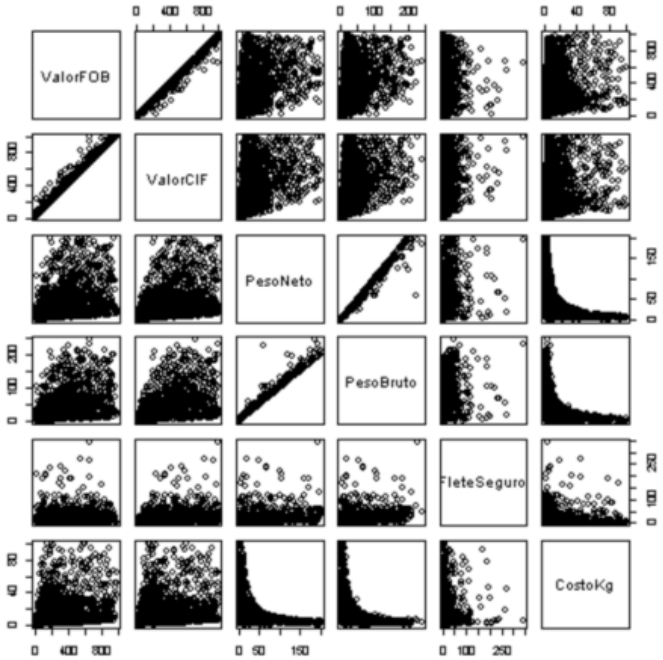


Figura 73. Relación de las variables luego de aplicar limpieza de datos.

La imagen muestra como luego de aplicar el código R en los datos los extremos prácticamente desaparecen y la información está lista para el siguiente paso.

5.2.1.6.2.4.3 Normalización de los Datos

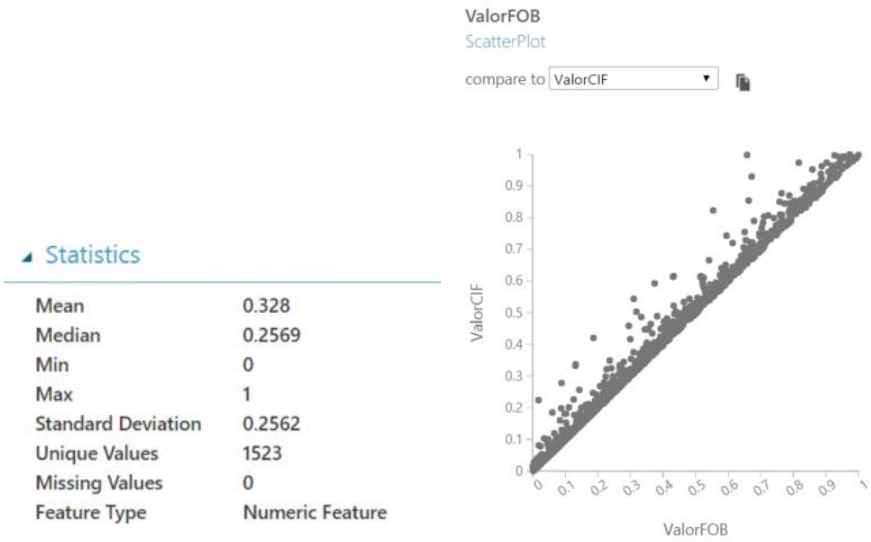


Figura 74. Estadísticas de la variable CIF luego de la normalización.

Las imágenes muestran como luego de la normalización las estadísticas de la columna FOB se vuelven más manejables para el análisis y el grafico muestra como el aplanamiento de la relación entre la variable a predecir se torna más clara y favorable para su análisis.

5.2.1.6.2.4.4 Evaluación del Modelo

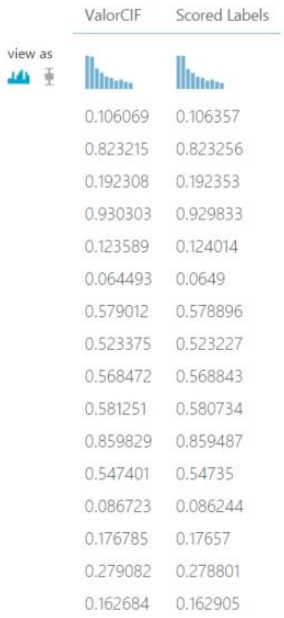


Figura 75. Resultados de aplicar el algoritmo de regresión lineal.

La imagen muestra como los valores resultantes se apegan bastante al valor real, teniendo por enterado que el modelo también resulta bastante apropiado para este grupo de productos.

5.2.1.6.2.5 Resinas y Aditivos Auxiliares

5.2.1.6.2.5.1 Análisis de Valores Iniciales y Relaciones

Se selecciona el valor CIF del conjunto de entrenamiento y se analiza los valores, luego de realizar la limpieza de valores perdidos y duplicados dentro de la herramienta.

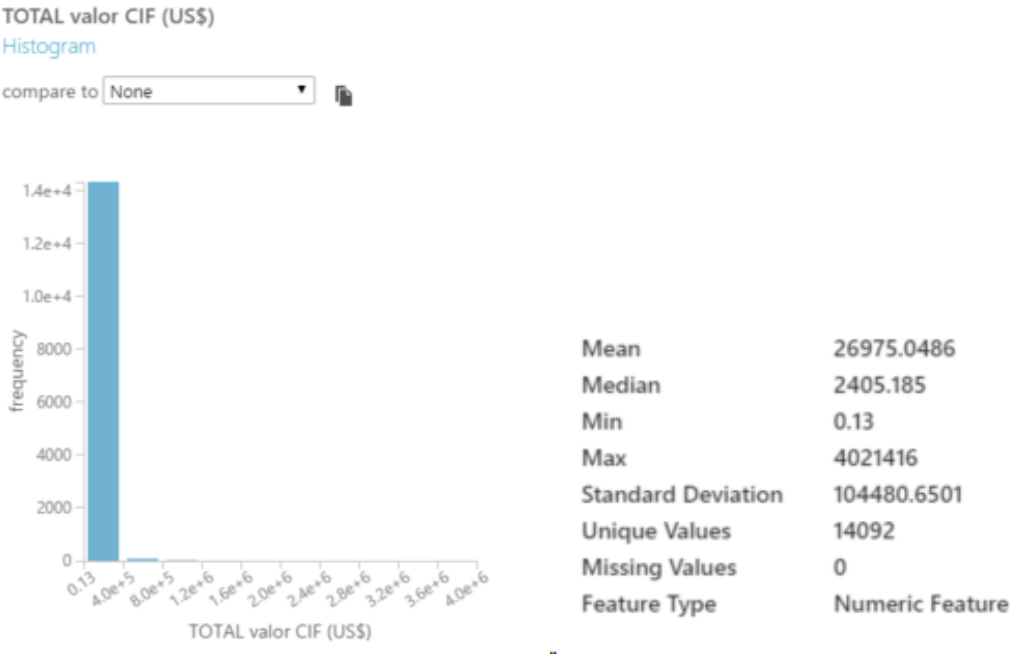


Figura 76. Análisis inicial de valores CIF.

En la imagen se observa como la gran mayoría de valores se agrupan en un solo tracto cerca del 99%, pero existe un grupo de menos del 1% que se salen de ese tracto. Se corrobora en las estadísticas, en donde si se observa los valores extremos están muy por encima y por debajo de la media y de la mediana, lo que se refleja en la desviación estándar. Para nuestro conjunto de prueba estos datos se eliminarán para no sesgar el estudio.

En los valores de FOB y de peso total se tiene comportamientos similares a los anteriores como se muestran a continuación.

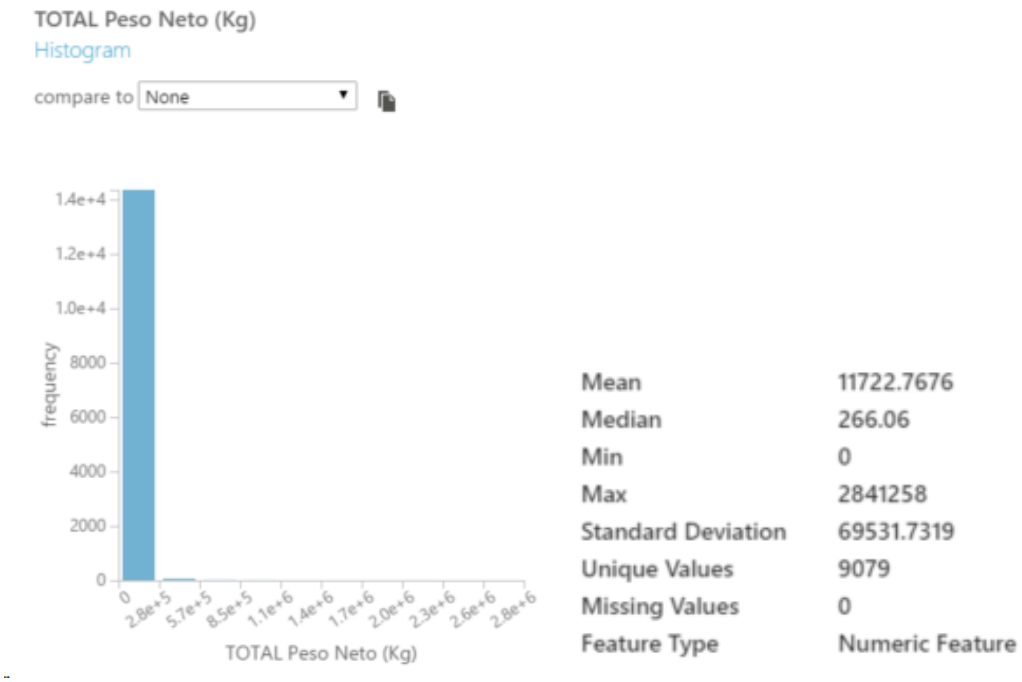


Figura 77 Máximos y mínimos de datos de resinas y aditivos auxiliares.

En la imagen se muestra como los valores máximo y mínimo se encuentran lejos del valor medio y de la mediana del grupo de datos.

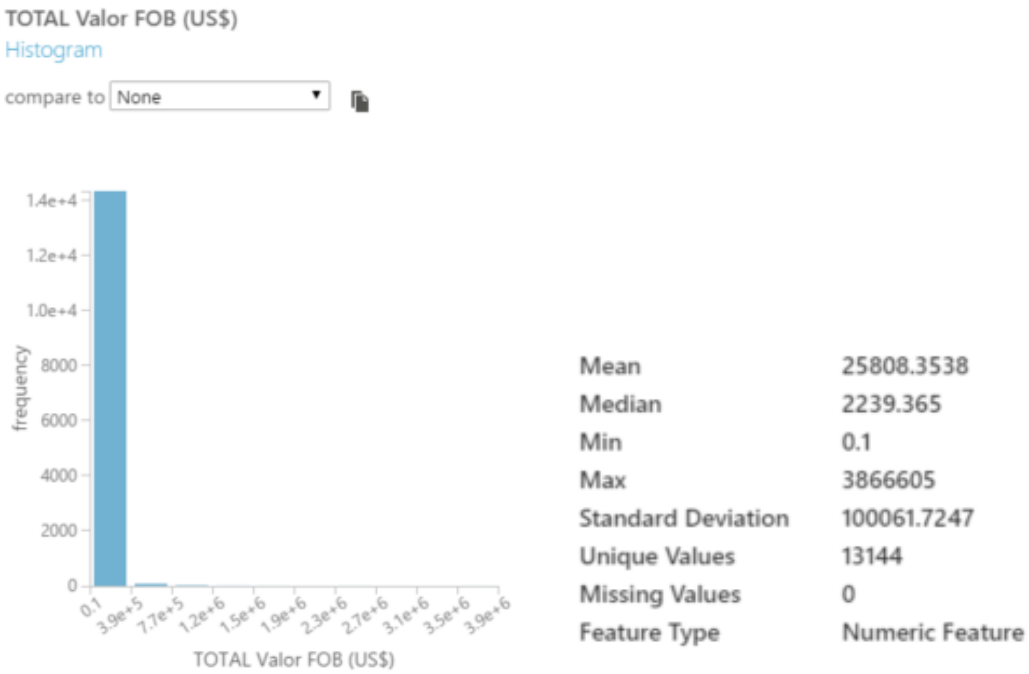


Figura 78. Estadísticas del valor FOB.

En la imagen se muestra el comportamiento del valor FOB y como los valores mínimos y máximos distan de la media y la mediana, lo que provoca que la desviación estándar se muestre con ese valor.

En este análisis también se debe de ver la comparación de variables y así corroborar los sesgos posibles.

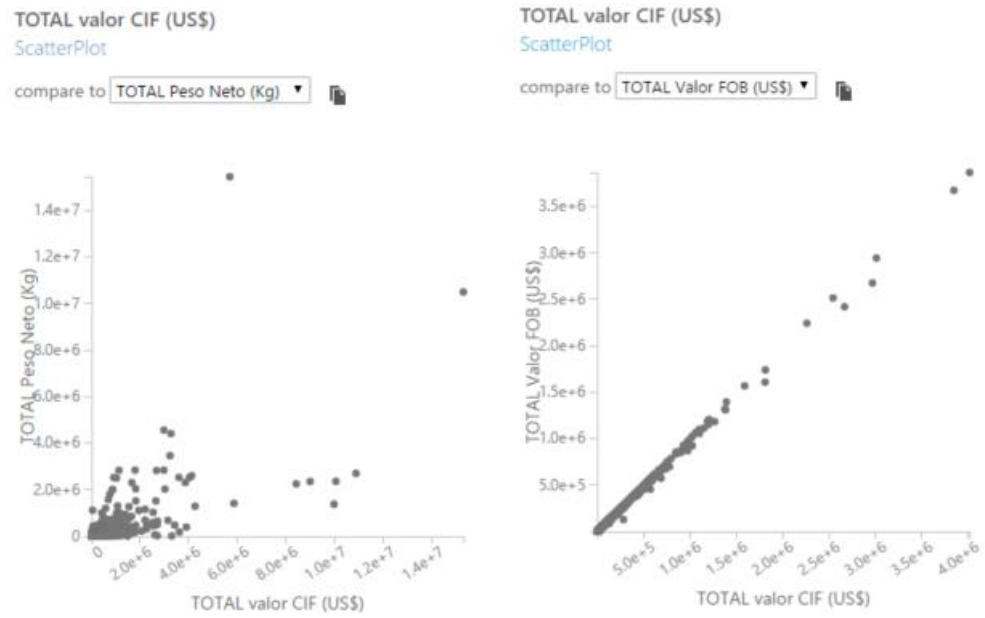


Figura 79. Relación de la variable CIF, con respecto a FOB y peso neto.

En la imagen de la derecha se muestran como los valores CIF junto con los valores FOB se agrupan en el sector izquierdo, sin embargo, algunos registros se distancian hacia la derecha, que como se menciona pude que sean correctos, pero para nuestro análisis este sesgo se trata de evitar. En la imagen de la izquierda se muestra la comparación del valor CIF junto con el valor de peso en kilogramos lo que muestra es un incremento casi lineal, pero de igual manera valores extremos que se eliminan para el estudio.

Para el análisis, el valor de FOB y peso total en kilogramos serán las variables que nos permitirá saber el valor CIF de un determinado grupo de productos o familia.

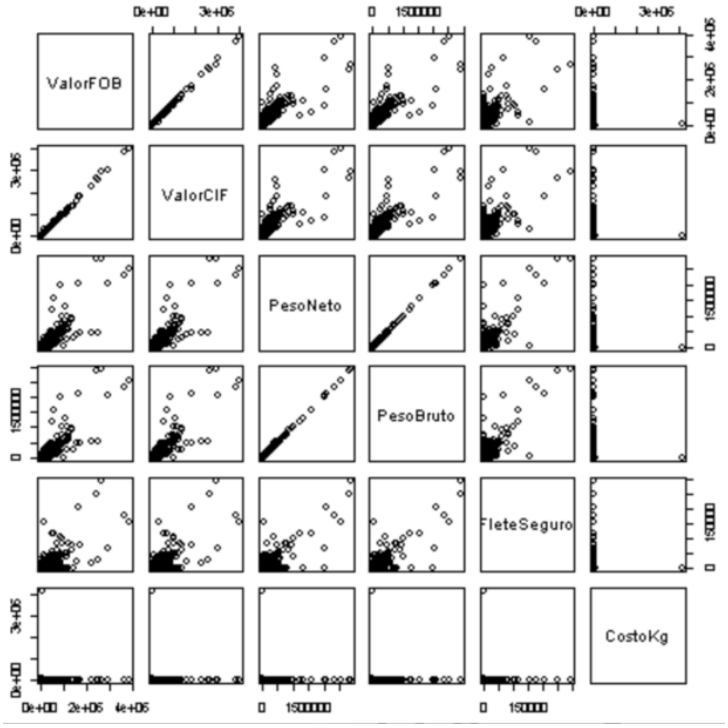


Figura 80. Relación de variables y sus valores extremos.

La imagen se muestra los gráficos de dispersión que se ejecutan en Machine Learning con lenguaje R previo a la ejecución de la limpieza.

5.2.1.6.2.5.2 Limpieza de Valores Extremos

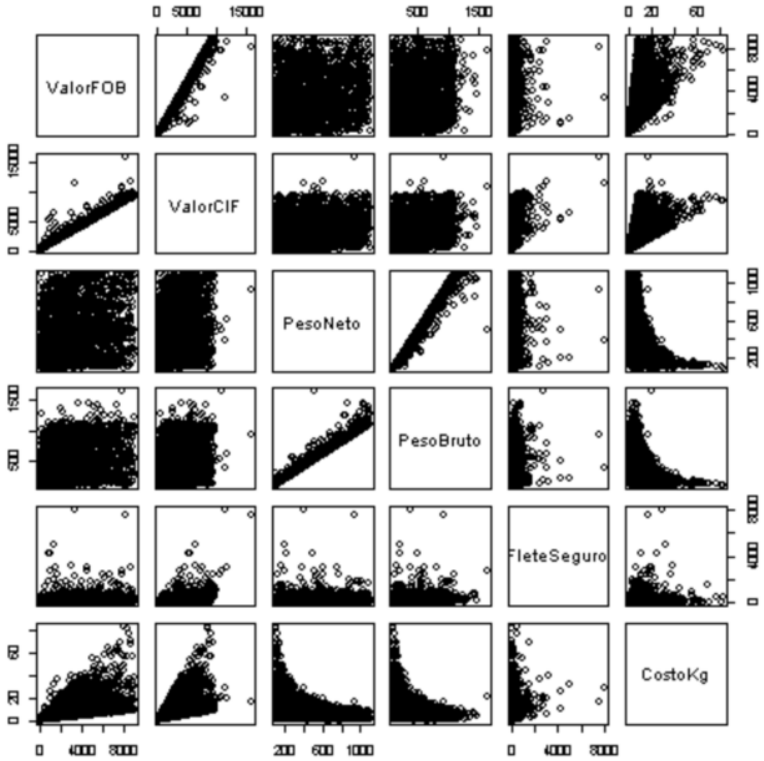


Figura 81. Relaciones de variables luego de aplicar métodos de limpieza.

La imagen muestra luego de la ejecución de los procesos de limpieza de valores extremos en el grupo de valores. Los gráficos que intersecan los valores CIF con los valores FOB y peso en kilogramos ya no muestran los valores fuera del grupo principal. Este tipo de trama permite examinar las relaciones entre muchas variables en una sola vista.

5.2.1.6.2.5.3 Normalización de los Datos

Luego de esto se normaliza los datos para que la distribución sea de una curva normal o Gaussiana.

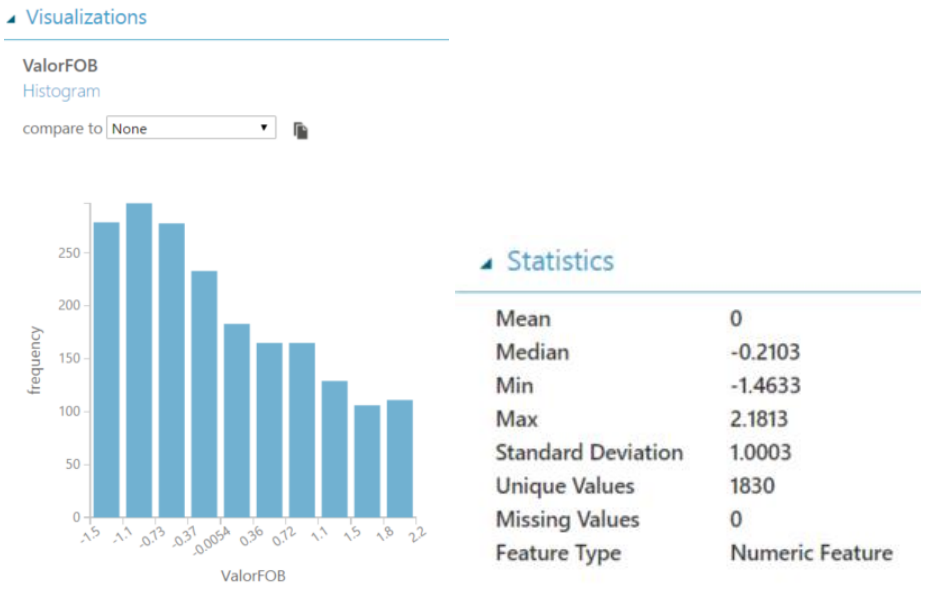


Figura 82. Estadísticas luego de la normalización de los datos.

La imagen muestra la normalización de los datos para el grupo correspondiente a resinas y aditivos auxiliares.

5.2.1.6.2.5.4 Evaluación del Modelo

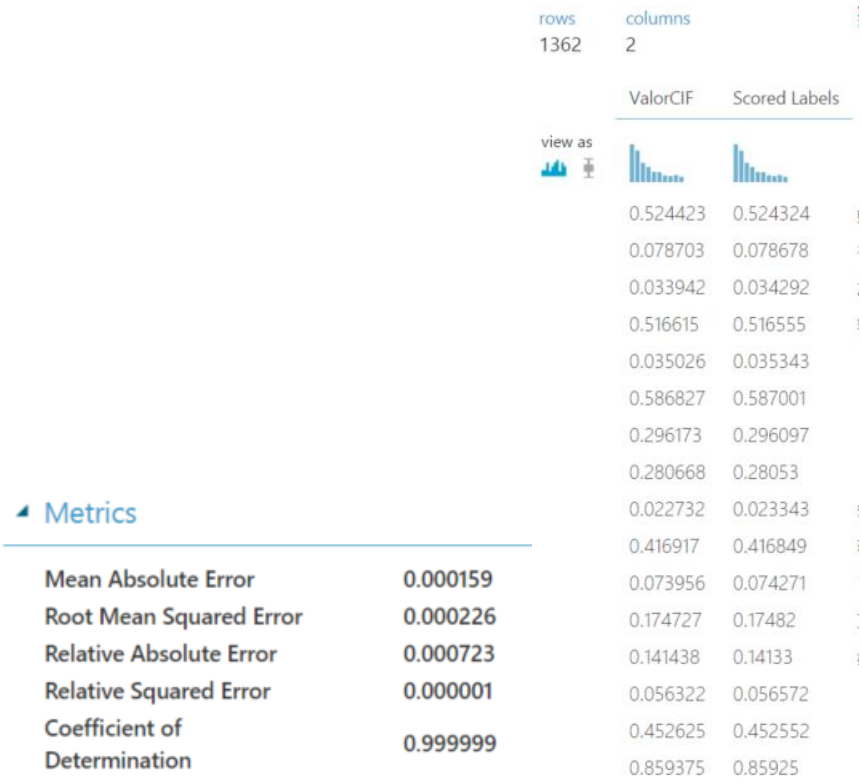


Figura 83. Estadísticas resultantes luego de aplicar regresión lineal.

La imagen de la izquierda muestra los valores resultantes la primera columna el valor original y la segunda el cálculo hecho por el modelo de regresión lineal. La segunda imagen muestra los valores estadísticos correspondientes a este grupo de familia de productos.

5.2.1.6.3 Evaluación de los Modelos Predictivos

El proceso de validación del mejor prospecto de algoritmo se debe de llevar a cabo teniendo en mente que el modelo debe de pasar por una serie de pasos antes de ser considerado como un modelo en producción, lo primero que se debe de hacer es crear un modelo de entrenamiento, con un grupo significativo de registros. Validar, como se ha hecho en el trascurso del proyecto, los registros, ejecutar la limpieza de los mismo y crear visualizaciones de los mismo para tomar la mejor decisión en cuanto al modelo. Luego de esto el modelo pasa a un proceso de pruebas, en donde se envían diferentes parámetros se cambian los datos, se ejecutan diferentes modelos sobre los diferentes grupos de datos. Y finalmente el proceso de evaluación en donde se entrega a un usuario final para que lo evalué desde su óptica y haga sus aportes al modelo previo de colocarlo en producción.

Es por esto que en esta parte del proyecto se desarrollaron varios modelos y se evaluaron entre sí, para así poder dar con el algoritmo adecuado a la predicción de la variable Valor CIF. Para esto se toman los siguientes algoritmos y se evalúan con el mismo conjunto de datos.

- Regresión lineal, este algoritmo se evalúa en con dos variantes una con las columnas polinomiales y otra con solo las variables base del conjunto. Teniendo los siguientes resultados,

- Función polinomial: Los valores de referencia coeficiente de determinación y error cuadrático relativo se dan muy favorables y casi perfectos en el cálculo de la variante

- Función base: Si se compara con el análisis anterior se puede observar un desmejoramiento en las estadísticas y por ende una menor aproximación a la variable predictible con respecto a su grupo de entrenamiento.

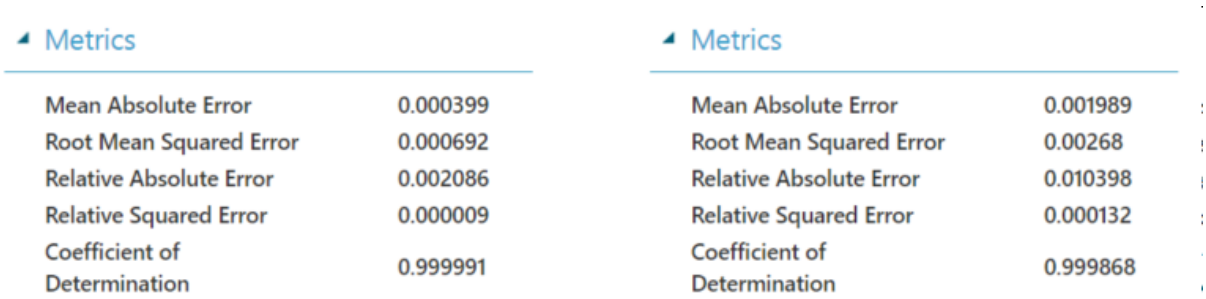


Figura 84. Comparación de función base y la polinomial.

En la imagen se muestra la comparación de ambas salidas de algoritmos, a la derecha con la aplicación de las columnas polinomiales y a la derecha sin ellas, se nota un mejor desempeño con las columnas polinomiales.

- Regresión de redes neuronales. Si se compara con la regresión lineal el algoritmo es menos preciso, a continuación, se muestran las comparaciones entre ambos algoritmos.

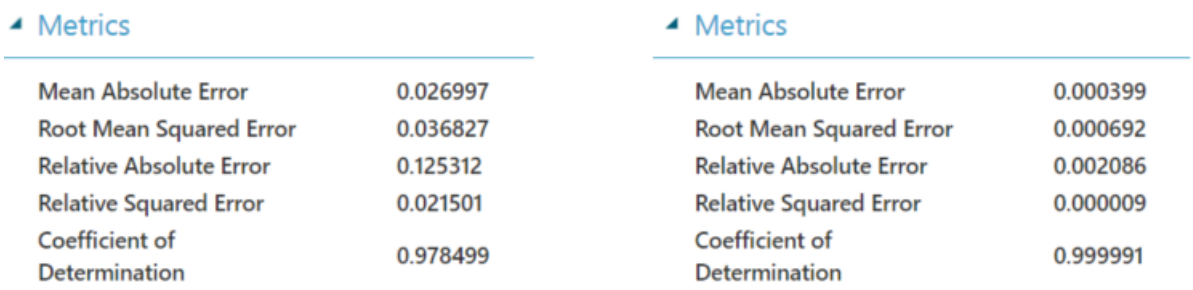


Figura 85. Comparación entre los algoritmos de regresión lineal y regresión de redes neuronales.

En la imagen se muestra la comparación de ambos algoritmos, a la izquierda los valores estadísticos del procesamiento por redes neuronales y la derecha por regresión lineal. Se muestra que la regresión lineal para este conjunto de datos sigue siendo la mejor opción.

- Arboles de regresión. Este tipo de algoritmo se distancia del valor presentado en el algoritmo de regresión lineal.

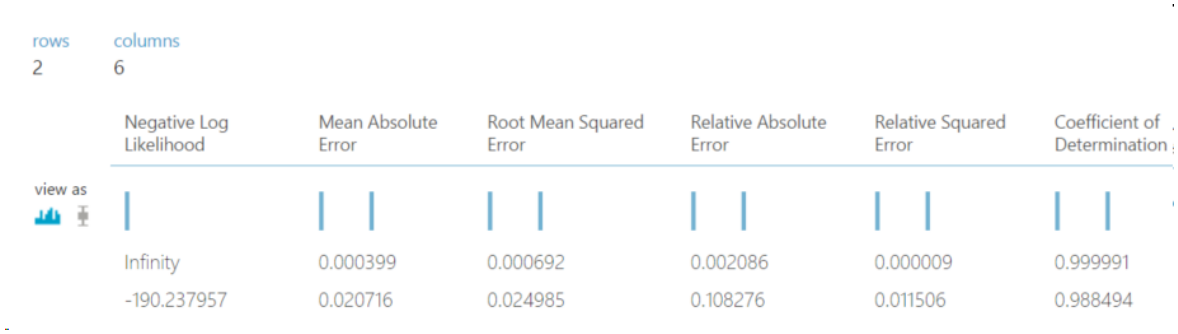
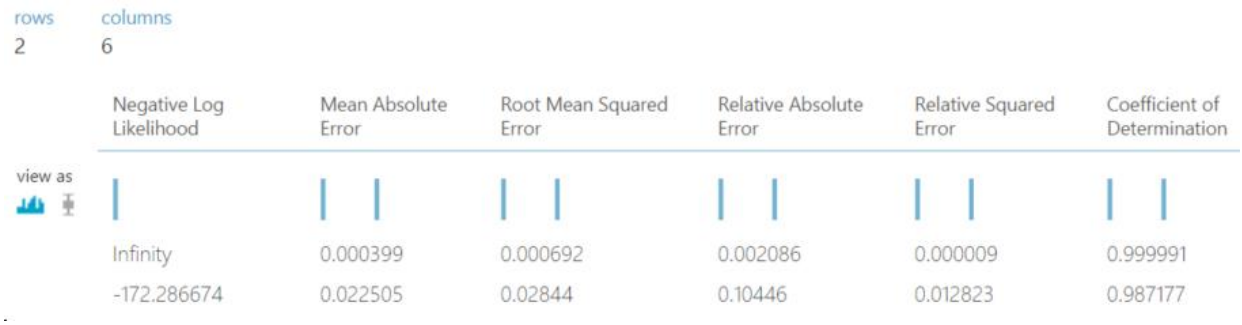


Figura 86. Comparación entre arboles de regresión y regresión lineal.

Como se muestra en la imagen, las estadísticas presentan una diferencia en el coeficiente de determinación, aunque mínimo, pero diferencia con fin que sigue teniendo en favor del algoritmo de regresión lineal como algoritmo con un mejor desempeño en los datos.

- Regresión lineal bayesiana. Este es otro algoritmo que se presenta como alternativa de predicción de datos, sin embargo, como se muestra en las estadísticas para el conjunto de datos en estudio sigue presentando mejor rendimiento el algoritmo de regresión lineal.



rows 2 columns 6

	Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
view as						
	Infinity	0.000399	0.000692	0.002086	0.000009	0.999991
	-172.286674	0.022505	0.02844	0.10446	0.012823	0.987177

Figura 87. Comparación entre regresión lineal bayesiana y regresión lineal.

En la imagen se muestra las estadísticas de ambos algoritmos, la línea de arriba la regresión lineal y en la línea de abajo la regresión lineal bayesiana.

Como resultado de esta fase el algoritmo que mejor desempeñen los datos es la regresión lineal, por lo que este modelo será el que se llevara a cabo como parte del desarrollo del proyecto.

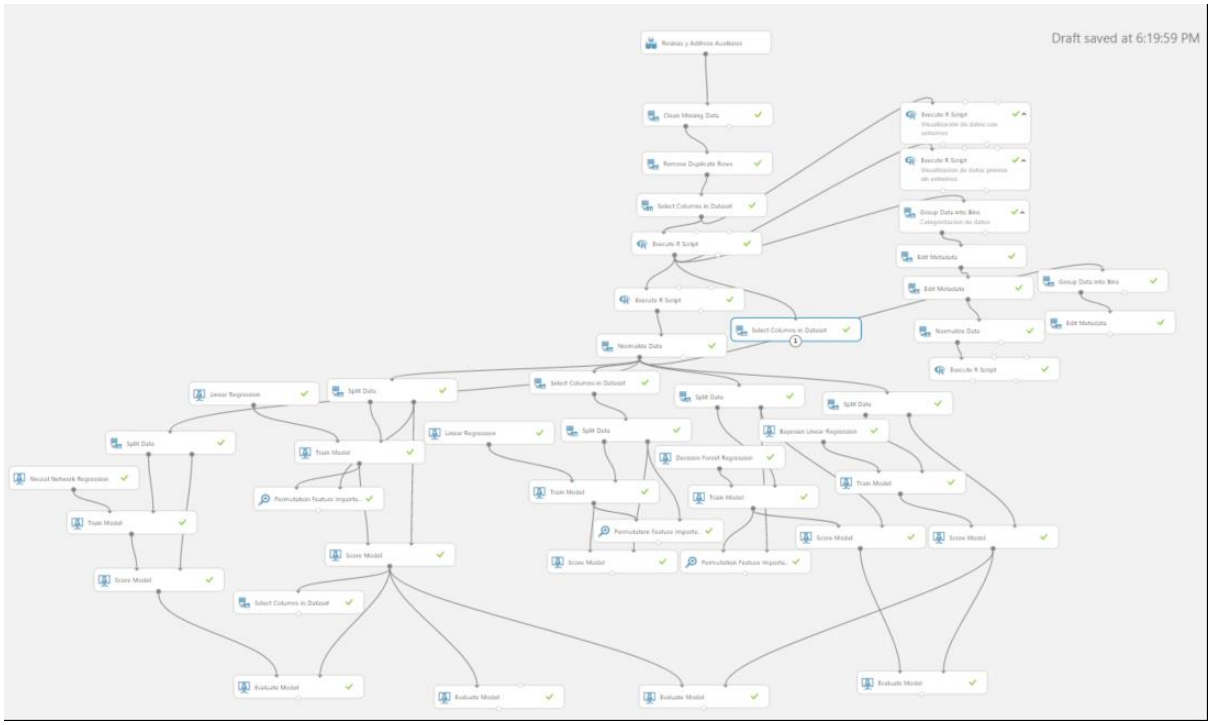


Figura 88. Panorámica del modelo general en Azure Machine Learning.

La imagen muestra una panorámica del modelo creado en Azure Machine Learning para los diferentes modelos, incluyendo módulos de visualización de datos, normalización, selección y evaluación de los diferentes algoritmos propuestos en la solución.

6 CAPÍTULO 6. CONCLUSIONES Y RECOMENDACIONES

6.1 CONCLUSIONES

- En este proyecto se llegaron a identificar los procesos correctos de extracción y monitoreo de datos para que el proceso sea más automático y no dependa de tanto tiempo de preparación de los usuarios de importaciones. Lo que conlleva a tener un mejoramiento en la preparación de la información final, su análisis y así comprensión del proceso que ha sucedido en cada empresa.

- Se ha podido establecer mecanismos de limpieza de datos desde los cuales se automatice el proceso, de tal forma que el usuario pueda ir creando reglas de negocio que enriquecen las bases de conocimiento del proceso y afinar limpieza y por ende mejorar el producto final.

- Se han podido encontrar las variables de entradas óptimas del proceso de modelaje en Machine Learning, muestra de esto son los valores de entrenamiento resultantes de los modelos para la predicción de flete y seguro, en los cuales se demuestra como algunas variables tienen un peso mayor que otras y como algunas se pueden obviar en este proceso.

- Se obtuvo el mejor modelo posible para cada una de las diferentes familias de productos que el proyecto contemplaba, modelo de regresión lineal, basado en las variables disponibles para este estudio, por lo cual también se cumple el objetivo de un análisis predictivo de variables de importaciones al iniciar por las que el estudio indicaba.

- En este proyecto se han descrito diferentes técnicas de análisis de datos, que se engloban en técnicas de procesamiento de datos pasados con herramientas de inteligencia de negocios como cubos multidimensionales y técnicas de regresión que se alinean más a la analítica de datos. Estas técnicas se desarrollan en dos fases distintas del proceso de análisis de datos de importaciones como un todo. La primera parte se demuestra que la aplicación de nuevas tecnologías como la extracción automática de diferentes fuentes de datos ayudan no solo en mejorar el tiempo de recaudación de la información, sino que también ayuda en mejorar el empleo del tiempo de los usuarios en analizar con mayor detenimiento los datos procesados.

- En este proyecto se ha podido llegar a implementar las dos fases iniciales de los requerimientos que se planteó como objetivo general del proyecto. La primera parte que contemplaba una herramienta multidimensional que sustituyera a la actual herramienta y la segunda el análisis predictivo de variables de importaciones flete y seguro para el desarrollo de una política más ágil de pre-costeo de los productos.

- Con este proyecto se pone en evidencia la falta de control en la carga de datos de la primera fase del proyecto y en el poco control que se llevaba en la carga de los datos que a la postre servían de base para la toma de decisiones, ya que la herramienta anterior no contaba con un sistema integrado de acceso, como la actual si lo tiene. Es por esto que ahora la extracción de las diferentes fuentes de datos se hace en forma automática y sin la intervención del usuario final al provocar que los

errores se minimicen a un tema de ingreso de los mismo que aun carga de información errónea.

- La herramienta de inteligencia de negocios provee un lugar en donde el usuario final puede ir desglosando el comportamiento de ventas, inventarios y prospectos de ventas que diferentes áreas llevan en su día a día, haciendo esto un concentrador de datos con un fin, determinar posibles compras a futuro. Como tal la herramienta ha demostrado que cumple los objetivos por los que fueron creadas.

- Al tener un cubo dimensional de las diferentes fuentes de datos se pueden hacer comparaciones más fáciles y con un menor grado de inconsistencias, ya que las métricas disponibles se pueden analizar por diferentes visiones y así poder encontrar esa información oculta que la anterior herramienta no podía suministrar por lo poco manejable.

- La actualización diaria de los datos de las fuentes transaccionales hace posible un seguimiento con poco esfuerzo por parte del usuario final, que anteriormente solo podía realizar este proceso una vez al mes por lo costoso en tiempo que le resultaba. Además, que con el procesamiento actual se adhieren nuevos usuarios que pueden encontrar una nueva fuente de datos que antes no se tenía.

- Los usuarios expertos han podido dar un mejor seguimiento a sus propias áreas y externar ciertas mejoras de otros procesos que impactan directamente la recolección de los datos en los sistemas transaccionales, esto trae como consecuencia que la calidad de los datos a procesar sea

paulatinamente mejor y con un mayor sentido, ya que al final los mismos usuarios serán los filtros de esta mejoría.

- Desde el punto de vista predictivo, la segunda fase busca ajustar de manera eficiente el proceso de pre-costeo del producto entrante, con esto se persigue que el proceso sea más expedito y se logre bajar los tiempos muertos del producto almacenado en puerto sin poder ser movido para su consiguiente comercialización.

- La comprensión de los datos y su consiguiente visualización por medio de técnicas de programación, incluyendo modelaje en R, ayudan a una comprensión integral de las variables dependientes y su relación con la variable por predecir, esta parte es fundamental, ya que permite hacer los diferentes controles sobre cada una de las variables y nos permiten ir comprendiendo cuál será el mejor algoritmo que procese la información.

- La definición de rangos de datos en los conjuntos de pruebas ayuda en la determinación de los valores extremos, que producen el sesgo de la información y su consecuente distorsión en la salida de los datos en la variable por predecir, este tema debe ser manejado con sumo cuidado en los modelos.

- Los modelos de predicción creados en el proyecto, muestran como los datos se ajustan a una función sencilla de regresión lineal y las variables, aunque algunas con un peso mayor que otras ayudan en la elaboración de un modelo robusto, con márgenes de errores o residuales muy por debajo de lo estándar.

- El manejo de rangos en las variables permite tener un análisis de información previo muy profundo y que luego ayuda a poder dar sugerencias de análisis de la información al usuario final que sin estas técnicas se podría dejar pasar por alto esta información. Como la encontrada en los rangos de carga con respecto a las demás variables como las de filete y seguro o las del monto de costo por kilogramo.

- El análisis por familia de producto surge luego de un estudio de datos que como consecuencia hace que el modelo se pueda ajustar de una mejor manera y encuentre valores extremos diferentes en las distintas familias de productos, con esto se puede hacer un modelo predictivo más preciso y adecuado a los productos que se encuentran agrupados en las distintas familias. Permitiendo que las variables relacionales se ajusten a la realidad del grupo de datos y se tenga un resultado muy similar al esperado.

- El proceso de creación de los modelos en Azure permite tener herramientas que se pueden crear desde cero, pero que requieren un conocimiento previo desde el punto de vista de programación y estadístico, además de entendimiento de relaciones de los datos. Lo que da como resultado que la herramienta tenga la elasticidad de poder integrar diferentes herramientas en un solo escenario, lo que la convierte en una herramienta robusta para la implementación de proyectos de predicción de datos dentro de las compañías.

6.2 RECOMENDACIONES

- Dentro de una nueva versión del sistema se puede implementar una política de construcción de la información en la que se pueda determinar los pesos de las variables según el sentido que se le quiere dar, por ejemplo, en cierto grupo de artículos por evaluar las tendencias de búsquedas podrían indicar posibles aumentos en el consumo de un determinado artículo, este tipo de definiciones son las que se pueden implementar en una nueva actualización de la herramienta. Cuando la salida es predictiva la política de construcción debe de ser prescriptiva.

- Ampliar la gama de variables a ser analizadas en Machine Learning enriquecerían los resultados y los análisis ya que permitirían poder hacer relaciones directas con variables del proceso de transporte de los productos a ser importados y con esto la mejora en la relación del costo y el beneficio del uso de la herramienta en el grupo de empresas.

- Tener en mente la siguiente fase de este proyecto es siempre poder adecuar la herramienta para que la misma pueda ser aplicada a las herramientas transaccionales, como lo son el ERP, CRM y en la toma de decisiones, no solo involucre a unos cuantos sino que la herramienta pueda expandirse en el tiempo para nuevas tomas de decisiones del día a día del usuario final, desde el simple hecho de crear un cliente y colocarlo en el grupo de adecuado de crédito, a 30, 60 o 90 días hasta para el movimiento de un determinado producto entre bodegas y su mejor costo/beneficio de esa acción. Y que por supuesto permeé las decisiones de alto nivel.

- Con la finalización de este proyecto la información es subida a Azure Machine Learning de una forma manual, en el futuro se espera subir la información en tiempo real a través de los mismos Web Services que la herramienta tiene a disposición y así poder hacer más interactiva la toma de decisiones con el soporte de Machine Learning.

- Agregar una parte multidimensional la segunda fase ayudaría a procesar información pasada y cotejarla con las predicciones de Azure Machine Learning. Ayudaría a poder entender el pasado de los datos y lo que se podría estar esperando del análisis de Azure Machine Learning.

- Cuando ya se coloque en producción el proyecto se debe de cambiar la cuenta actual de una gratuita a una por cobro por utilización, ya que el proceso de cálculo se tardaría más de lo debido cuando se trabajó más de un usuario con la herramienta de Azure Machine Learning.

- Utilizar nuevas herramientas de extracción de datos, como el lenguaje R o Python, incluye la opción de poder agregar nuevas fuentes de datos estructurados y los no estructurados. Inclusive la siguiente fase podría incluir fuentes no tradicionales de datos o fuentes que estén alojadas en la misma nube de Microsoft como Cortana Analytic.

- Utilizar aquella data que en este momento se deshecha para análisis de nuevos mercados y no solo enfocarse en los productos que actualmente se comercializan y ver mucho más allá de una comparación contra los mercados actuales, expandir esa visión de ventas hacia nuevos productos complementarios y que hagan sentido con el giro del negocio y

que esté en consonancia con lo que cada país demanda en ciertas épocas del año.

- Realizar la normalización de los datos una vez realizada la limpieza de los valores extremos para que estos no causen una distorsión en el procesamiento de los datos.
- Los encabezados de las columnas deben de ir sin espacio y sin algún símbolo como las tildes, además que las columnas que sean numéricas no incluyan leyendas producto del cálculo de una fórmula.
- Explorar nuevas utilizaciones de la herramienta de decisiones, por ejemplo, una decisión de los días de crédito que un cliente puede tener, se podría realizar con modelos de clasificación en donde cada tracto de pago se podría colocar como una clasificación que la herramienta podría sugerir en cual tracto debe de ir el cliente nuevo.

7 BIBLIOGRÁFIA

- Carbajal S., Julio Manuel. Introducción a la Clasificación Arancelaria de las Mercancías. Coedición Universo Arancelario 4a Edición, 2012.
- Ríos Diaz, Felipe Ezequiel. Sistema armonizado, nomenclatura común del Mercosur y estructura de la clasificación arancelaria en los países del Mercosur.
- Exploratory Data Analysis, John Tukey, 1977 (Addison-Wesley)
- Ríos Diaz, Felipe Ezequiel. El tratamiento de las mercaderías en los convenios internacionales en Argentina y el Mercosur. Centro de Estudios Sudamericanos. Agosto 2013.
- Lichman, M. (2013). UCI Machine Learning Repository.
- Grupo de Estadística Aplicada. 2006. Universidad de Salamanca. “Regresión y correlación”. Introducción a la Estadística.
- Data Science and Machine Learning Essentials. Microsoft Virtual Academy. Graeme Malcolm and Stephen Elston.
- Flach, Peter (2012) Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press.
- Mitchell, T. (1997). Machine Learning, McGraw Hill.
- Redes de Neuronas Artificiales - RAI - UC3M 2011/2012.
- Alba Castro, José Luis. «Máquinas de Vectores Soporte (SVM).
- Rousseeuw, P.J.; Kaufman, L. (1990). Finding Groups in Data: An Introduction to Clúster Analysis. Wiley.

- CRISP-DM 1.0, Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler)
- Massachusetts Instititute of Technology.
- Stanford University.
- Systematic Review in Software Engineering. Article · Jan 2005 · Information and Software Technology. Jorge Biolchini.
- S. Kotsiantis, Aprendizaje Automático: Una Revisión de la Clasificación de las técnicas de Informática (2007).
- https://www.python.org/doc
- https://www.rstudio.com/online-learning/

8 APÉNDICES

8.1 APÉNDICE 1

Diagrama de ventas de Microsoft Dynamics AX 2009

Compras de Microsoft Dynamics AX 2009

Diagrama de Inventarios de Microsoft Dynamics AX 2009

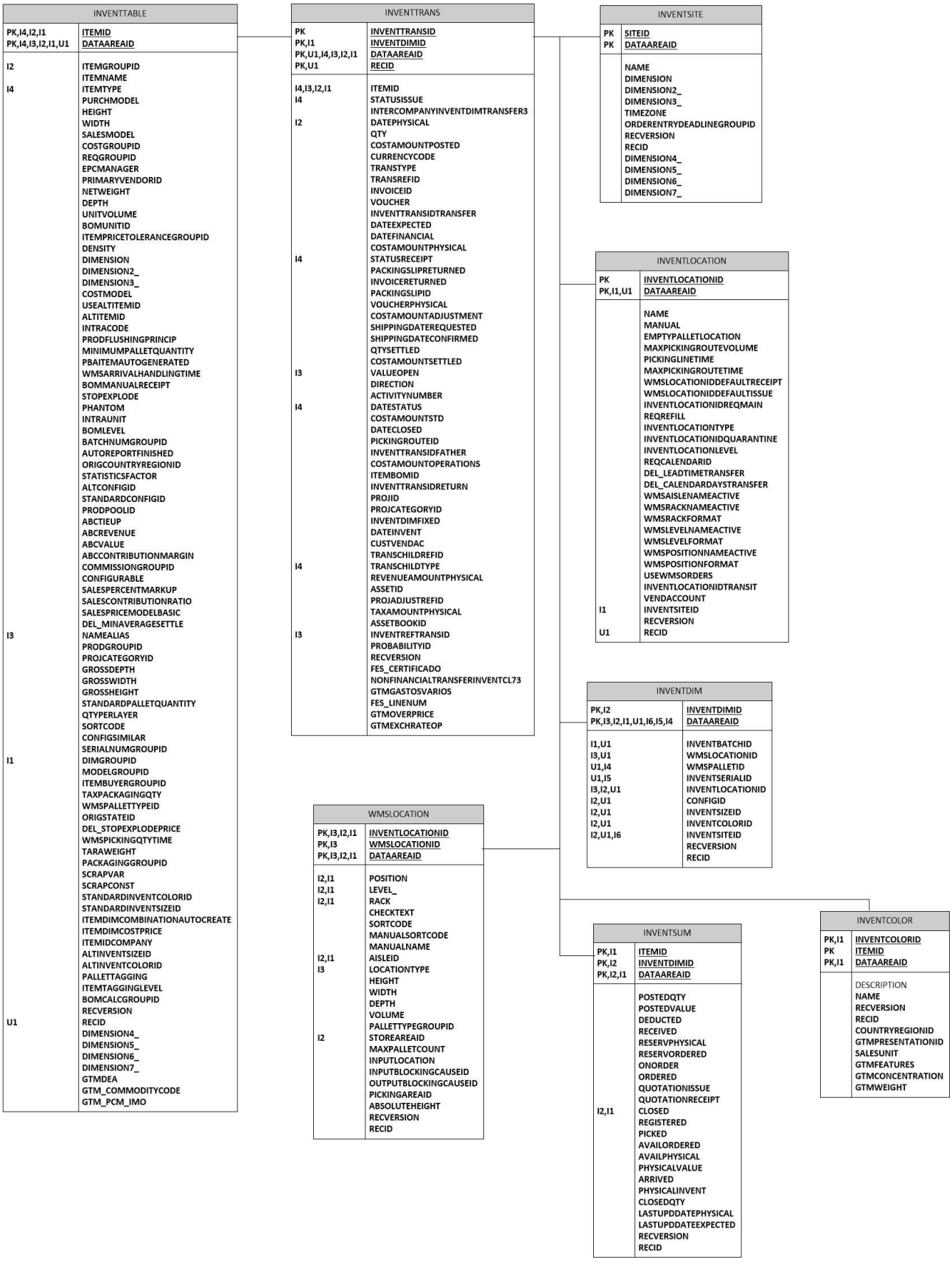
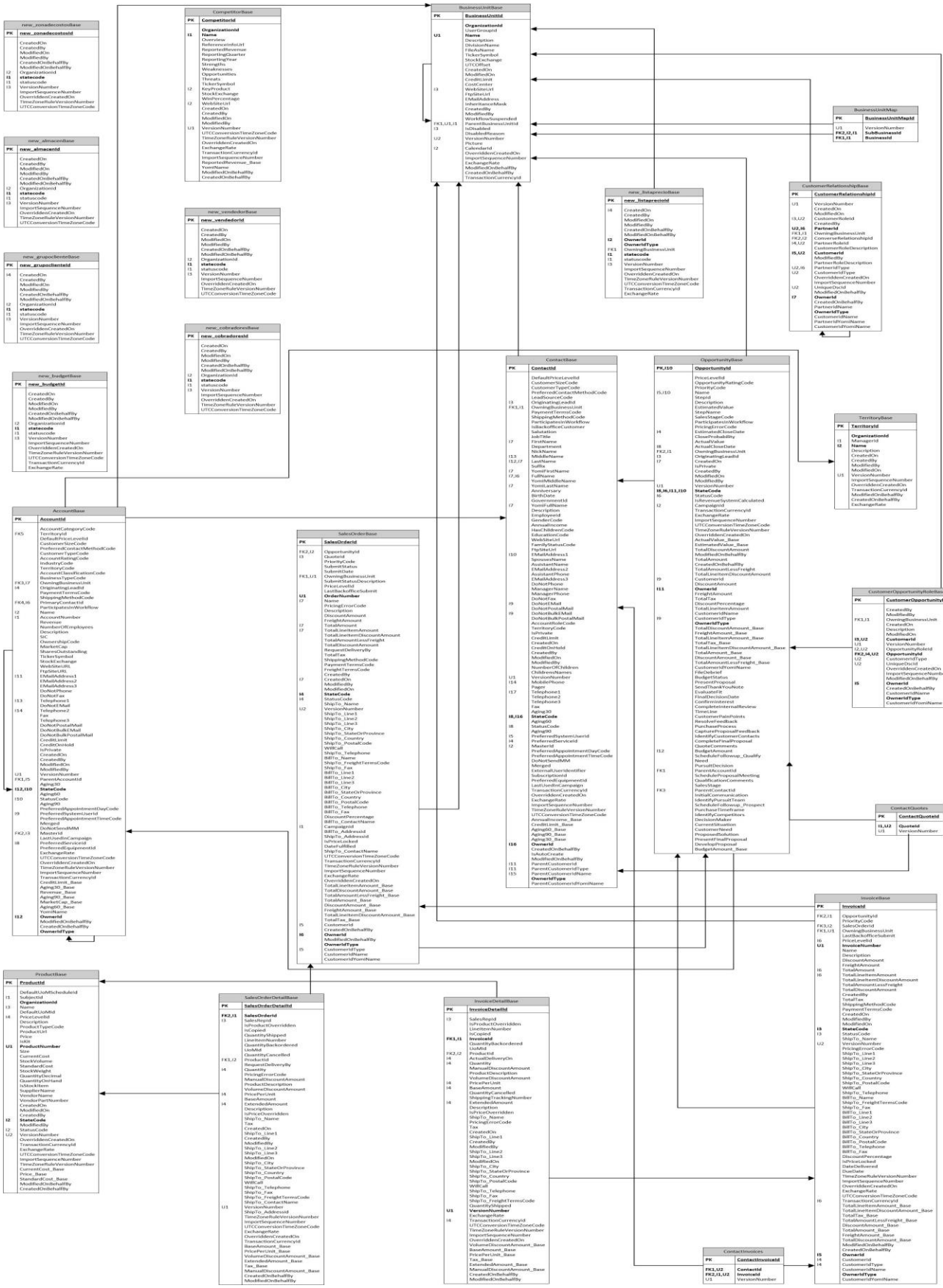
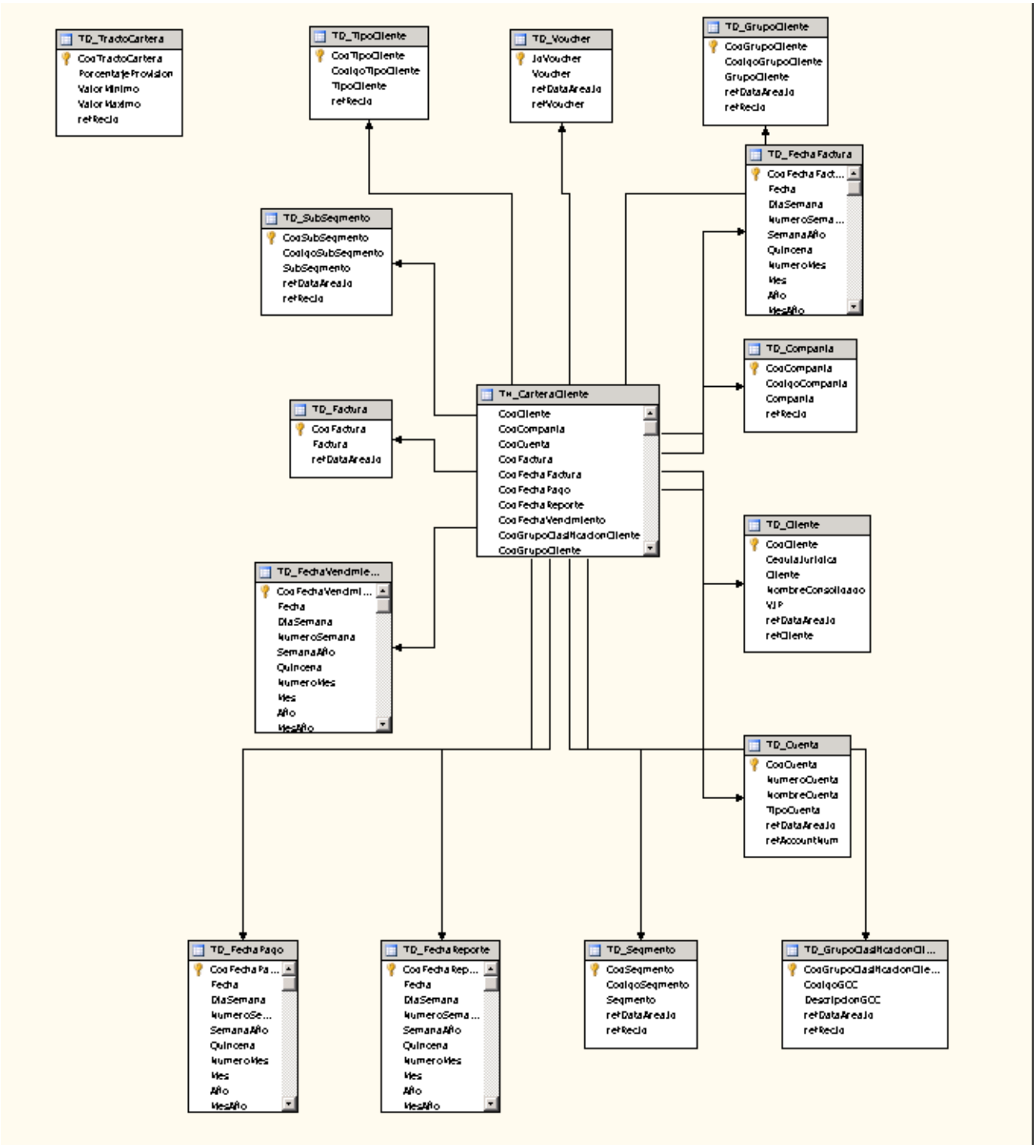


Diagrama de Microsoft Dynamics CRM 2011



8.2 APÉNDICE 2

Diagrama de la vista multidimensional del cubo de importaciones.



8.3 APÉNDICE 3

Consolidación de campos

Campo Homologado	Argentina	Brasil	Colombia	Costa Rica	Ecuador	El Salvado	Guatemal	Honduras	Nicaragua	Panama	Peru	Mexico
Partida Arancelaria	1	1	1	1	1	1	1	1	1	1	1	1
Valor CIF U$S	1	1	1	1	1	1	1	1	1	1	1	1
Peso Kg. Netos	1	1	1	1	1	1	1	1	1	1	1	1
Descripcion Comercial	1	1	1	1	1	1	1	1		1	1	1
Fecha	1	1	1	1	1	1	1	1	1	1	1	1
Pais de Origen	1	1	1	1	1	1	1	1	1	1	1	1
Unidad de Medida	1	1	1	1	1			1	1	1	1	1
Importador	1		1	1	1			1	1	1	1	
Porcentaje	1		1	1								
Aduana	1		1	1	1			1	1	1	1	1
Cantidad	1			1	1			1	1	1	1	1
Valor FOB /Unidad U$S	1	1	1	1	1					1	1	1
Via o Transporte	1		1	1	1			1		1	1	1
Proveedor			1	1	1			1	1	1	1	
Valor de Flete U$S	1		1	1	1			1		1	1	
Valor FOB U$S	1		1	1	1			1		1	1	
Peso Kg. Brutos	1		1	1	1			1		1	1	
Pais de Procedencia	1		1	1	1			1			1	
Regimen			1	1	1			1		1	1	
Numero de DUA	1		1	1	1					1	1	
Valor Seguro U$S	1		1	1	1			1		1		
Item	1			1	1						1	
Codigo Armonizado		1	1			1	1					1
Empresa de Transporte			1		1			1			1	
Descripcon Codigo Armonizado en I		1	1			1	1					1
Descripcion Arancel		1				1		1	1			1
Bultos			1	1	1					1	1	
Fecha 1		1	1			1						1
Agente Aduanero			1	1	1							
Otros Impuestos U$S				1				1		1		
Pais de Compra			1	1				1				
NIT			1					1	1		1	
Marca	1			1	1						1	
CIF/ Unidad U$S					1					1	1	
Base Arancel Pagado			1		1					1		
Estado / Ciudad	1	1	1									
Arancel Pagado	1		1		1							
Cantidad Estandar	1										1	
Valor Aduana U$S			1	1								
Unidad de Medida Estanc	1							1				
Manifiesto					1						1	
Poliza								1	1			
Condiciones de Venta	1							1				
Descripción Arancel			1		1							
Modalidad de la Importacion			1	1								
Cantidad		1	1									
Moneda Divisa	1							1				
Dirección del Importador			1		1							
Mes								1	1			
IVA U$S			1	1								
Tipo de Aforo					1							
Descripcion CUODE			1									
CIIU			1									
Fecha de Embarque					1							
Refrendo					1							
Fecha de Llegada					1							
Unidad de Medida 1											1	
Fecha de Pago					1							
Fecha de Manifiesto			1									
Fecha de Salida de Almacen					1							
Dia								1				
Fecha Factura			1									
Subtotal IVA Moneda Loca			1									
Flujo Comercial								1				
Total IVA en Moneda Local			1									
FOB Moneda Local	1											
Documento de Transporte					1							
Forma de Pago			1									
Estado de la Mercaderia					1							

8.4 APÉNDICE 4

Homologación de Campos Indispensables.

Pais	Campos de informacion Disponibles	Campo Homologado	Importancia del Campo
Argentina	Descripcion Arancel	Descripcion Comercial	Indispensable
Argentina	Fecha	Fecha	Indispensable
Argentina	IMPORTADOR PROBABLE	Importador	Indispensable
Argentina	Pais Origen	Pais de Origen	Indispensable
Argentina	NCE-SIM	Partida Arancelaria	Indispensable
Argentina	Kg. Netos	Peso Kg. Netos	Indispensable
Argentina	CIF U$S	Valor CIF U$S	Indispensable
Brasil	Descripcion Comercial del Producto	Descripcion Comercial	Indispensable
Brasil	Fecha	Fecha	Indispensable
Brasil	Pais Origen	Pais de Origen	Indispensable
Brasil	Posición	Partida Arancelaria	Indispensable
Brasil	TOTAL Peso Neto (Kg)	Peso Kg. Netos	Indispensable
Brasil	TOTAL Valor FOB (US$)	Valor CIF U$S	Indispensable
Colombia	Descripcion Comercial del Producto	Descripcion Comercial	Indispensable
Colombia	Fecha	Fecha	Indispensable
Colombia	Importador	Importador	Indispensable
Colombia	Pais Origen	Pais de Origen	Indispensable
Colombia	Posición	Partida Arancelaria	Indispensable
Colombia	TOTAL Peso Neto (Kg)	Peso Kg. Netos	Indispensable
Colombia	Proveedor	Proveedor	Indispensable
Colombia	TOTAL valor CIF (US$)	Valor CIF U$S	Indispensable
Costa Rica	DETALLE DE MERCADERIA	Descripcion Comercial	Indispensable
Costa Rica	FECHA	Fecha	Indispensable
Costa Rica	IMPORTADOR	Importador	Indispensable
Costa Rica	PAIS ORIGEN	Pais de Origen	Indispensable
Costa Rica	ARANCEL	Partida Arancelaria	Indispensable
Costa Rica	KG. NETOS	Peso Kg. Netos	Indispensable
Costa Rica	PROVEEDOR	Proveedor	Indispensable
Costa Rica	CIF U$S	Valor CIF U$S	Indispensable
Ecuador	DESCRIPCION DE MERCADERIA	Descripcion Comercial	Indispensable
Ecuador	FECHA	Fecha	Indispensable
Ecuador	IMPORTADOR	Importador	Indispensable
Ecuador	ORIGEN	Pais de Origen	Indispensable
Ecuador	ARANCEL	Partida Arancelaria	Indispensable
Ecuador	KG. NETOS	Peso Kg. Netos	Indispensable
Ecuador	PROVEEDOR	Proveedor	Indispensable
Ecuador	CIF U$S	Valor CIF U$S	Indispensable
El Salvador	Descripcion Comercial del Producto	Descripcion Comercial	Indispensable
El Salvador	Fecha	Fecha	Indispensable
El Salvador	Pais Origen	Pais de Origen	Indispensable
El Salvador	Posición	Partida Arancelaria	Indispensable
El Salvador	TOTAL Peso Neto (Kg)	Peso Kg. Netos	Indispensable
El Salvador	TOTAL valor CIF (US$)	Valor CIF U$S	Indispensable
Guatemala	Descripción Arancel	Descripcion Comercial	Indispensable
Guatemala	Fecha	Fecha	Indispensable
Guatemala	Pais Origen	Pais de Origen	Indispensable
Guatemala	Posición	Partida Arancelaria	Indispensable
Guatemala	TOTAL Peso Bruto (Kg)	Peso Kg. Netos	Indispensable
Guatemala	TOTAL Valor FOB (US$)	Valor CIF U$S	Indispensable
Honduras	Descripcion Comercial del Producto	Descripcion Comercial	Indispensable
Honduras	Fecha	Fecha	Indispensable
Honduras	Importador	Importador	Indispensable
Honduras	Pais Origen	Pais de Origen	Indispensable
Honduras	producto	Partida Arancelaria	Indispensable
Honduras	Peso Neto KG	Peso Kg. Netos	Indispensable
Honduras	proveedor	Proveedor	Indispensable
Honduras	Valor Aduana	Valor CIF U$S	Indispensable
Mexico	Descripcion Comercial del Producto	Descripcion Comercial	Indispensable
Mexico	Fecha	Fecha	Indispensable
Mexico	Pais Origen	Pais de Origen	Indispensable
Mexico	Posición	Partida Arancelaria	Indispensable
Mexico	TOTAL Peso Bruto (Kg)	Peso Kg. Netos	Indispensable
Mexico	TOTAL Valor FOB (US$)	Valor CIF U$S	Indispensable
Nicaragua	Fecha	Fecha	Indispensable
Nicaragua	consignatario	Importador	Indispensable
Nicaragua	nombre del importador	Importador	Indispensable
Nicaragua	Pais Origen	Pais de Origen	Indispensable
Nicaragua	sac	Partida Arancelaria	Indispensable
Nicaragua	peso_bruto	Peso Kg. Netos	Indispensable
Nicaragua	exportador	Proveedor	Indispensable
Nicaragua	valor_cif	Valor CIF U$S	Indispensable
Panama	DETALLE DE MERCADERIA	Descripcion Comercial	Indispensable
Panama	FECHA	Fecha	Indispensable
Panama	IMPORTADOR	Importador	Indispensable
Panama	PAIS PROCEDENCIA	Pais de Origen	Indispensable
Panama	ARANCEL	Partida Arancelaria	Indispensable
Panama	KG. NETOS	Peso Kg. Netos	Indispensable
Panama	SHIPPER	Proveedor	Indispensable
Panama	CIF U$S	Valor CIF U$S	Indispensable
Peru	DESCRIPCION PARA FILTRO	Descripcion Comercial	Indispensable
Peru	DETALLE DE MERCADERIA	Descripcion Comercial	Indispensable
Peru	FECHA	Fecha	Indispensable
Peru	IMPORTADOR	Importador	Indispensable
Peru	PAIS ORIGEN	Pais de Origen	Indispensable
Peru	NANDINA	Partida Arancelaria	Indispensable
Peru	KG. LIQUIDOS	Peso Kg. Netos	Indispensable
Peru	PROVEEDOR - SHIPPER	Proveedor	Indispensable
Peru	CIF U$S	Valor CIF U$S	Indispensable

8.5 APÉNDICE 5

Entrevista al Encargado de la Cadena de Abastecimiento y Operaciones de Grupo Transmerquim

1. ¿Cuál es el estado del análisis de la información del área, en la actualidad?

Actualmente la herramienta que se tiene para este trabajo está desarrollada en Access y la información se extrae de reportes del ERP y del CRM, no integra información de presupuestos. Este proceso se lleva cerca de 4 días en su producción y luego entra en desuso el resto del mes y la información se toma solo para la reunión mensual de compras.

2. ¿Cuál es la propuesta que se usted como encargado del área puede hacer con respecto al análisis de la información?

Por lo laborioso del tema, propone poder tener la herramienta más a la mano y poder consultad diariamente, sin la necesidad de estar generando tanto reporte y poder ejecutar la información casi en tiempo real o con un día de diferencia. Además, que se integre nuevas variables de importaciones que nos permita ver más allá del simple análisis de la información pasada y más bien pueda proyectar de forma inteligente las operaciones de importaciones de productos o sugiera nuevos nichos de negocios sin cubrir.

3. ¿Cuál es el objetivo primordial que debe de cumplir la nueva herramienta?

Debe de ser ágil, automática, disponible cuando se requiera y que no demande tanta operación humana. Y que pueda sugerir información a futuro basado en datos del pasado.

4. ¿Cuáles deben de ser las variables de análisis de la información de la nueva herramienta?

Por el momento las mismas que se manejan actualmente, pero debe poder adaptarse a nuevas y a grupo de información mucho más extensa o de diferentes fuentes, desde redes sociales hasta información de archivos planos.

5. ¿Cuál debe de ser la herramienta que por excelencia debe de integrar la información resultante?

Los usuarios de las diferentes afiliadas son un poco reticentes a utilizar herramientas muy complicadas, por lo mismo la herramienta debe ser sencilla pero que aun así pueda hacer el análisis poderoso, lo que se sugiere es que se siga analizando en Office Excel de Microsoft.

6. ¿El costo versus el beneficio es determinante en el proyecto?

Se considera que con el hecho de automatización se podría ayudar a liberar los recursos de muchas de las compañías que se encargan de alimentar la información, por lo que el costo siendo alto justifica el cambio.

7. ¿Se considera un proyecto por fases en la que se pueda ir cambiando y adecuando nuevas herramientas de análisis?

La idea sería poder avanzar primero con el cambio de la herramienta actual y luego de esto incorporar nuevas herramientas de análisis que consoliden en Excel.

8. ¿Se considera la precisión como un punto insalvable del proyecto?

Ahí se podría ver de dos maneras, la primera del análisis del transaccional del ERP y del CRM los datos son muy precisos y por lo mismo esta fase debe de

dar análisis de igual forma precisos. La segunda fase debe de ser una herramienta que pueda comprender la información de muchas variables y pueda hacer sugerencias y proyecte esto al futuro.

9. ¿Quién definiría los resultados como aceptables o incorrectos en el proyecto?

Un grupo que se encargaría de dar apoyo a la parte técnica y que tiene el pleno conocimiento del análisis de la información y sus patrones.

10. ¿Cuál sería el grupo que integraría el proyecto?

El grupo estaría integrado por el Encargado de Abastecimiento, un grupo de encargados de importaciones de los diferentes países y el apoyo del departamento de TI del grupo.

8.6 APÉNDICE 6

Filtro de datos extremos y código de visualización de sesgos y comparaciones.

Execute R Script

R Script

```r
frame1 <- maml.mapInputPort(1)
## Remove outliers
library(dplyr)
frame1 <- frame1 %>% filter(ValorFOB < 30000) %>%
filter(ValorFOB > 1000)  %>%
filter(ValorCIF < 30000) %>%
filter(ValorCIF > 2000)  %>%
filter(PesoNeto < 2900)  %>%
filter(PesoNeto >500)    %>%
filter(PesoBruto < 3000)  %>%
filter(PesoBruto >500)    %>%
filter(FleteSeguro < 2600)  %>%
filter(FleteSeguro >500)    %>%
filter(CostoKg < 20)  %>%
filter(CostoKg >0.001)
## Output the data frame
maml.mapOutputPort('frame1')
```

Properties Project

Execute R Script

R Script

```r
# Map 1-based optional input ports to variables
dataset1 <- maml.mapInputPort(1) # class: data.frame
## Create a pairs plot.
pairs(dataset1)
```

```r

# Install ggplot2 (use a personal library if  prompted!)
## install.packages('ggplot2', dep = TRUE)

## Use basic R graphics to create a pair-wise scatter plot
Azure = TRUE
if(Azure){
  eeframe <- maml.mapInputPort(1)
  maml.mapOutputPort('eeframe')
}
pairs(~ ., data = eeframe)
## Use ggplot2 to create conditioned scatter plots
library(ggplot2)
plotCols <- c("FOB",
              "Neto",
              "Bruto",
              "FleteSeguro",
              "CostoKg",
              "FOB2",
              "Neto2",
              "Bruto2",
              "FleteSeguro2",
              "CostoKg2",
              "FOB3",
              "Neto3",
              "Bruto3",
              "FleteSeguro3",
              "CostoKg3")
plotEE <- function(x){
  title <- paste("Valor CIF vs", x, "\n PesoNeto en KG")
  ggplot(eeframe, aes_string(x, "CIF")) +
    geom_point() +
    ##  facet_grid(Carga) +
    ggtitle(title) +
    stat_smooth(method = "lm")
}
lapply(plotCols, plotEE)


## Crear  histogramas
plotCols4 <- c("FOB",
               "Neto",
               "Bruto",
               "FleteSeguro",
               "CostoKg",
               "FOB2",
               "Neto2",
               "Bruto2",
               "FleteSeguro2",
               "CostoKg2",
               "FOB3",
               "Neto3",
               "Bruto3",
               "FleteSeguro3",
               "CostoKg3")
library(gridExtra)
eeHist <- function(x) {
  title <- paste("Histograma de", x, "condicional en PesoxKg")
  ggplot(eeframe, aes_string(x)) +
    geom_histogram(aes(y = ..density..)) +
    ## facet_grid(. ~ Carga) +
    ggtitle(title) +
    geom_density()
}
lapply(plotCols4, eeHist)

## Create box plots
eebox <- function(x) {
  title <- paste("Diagrama de caja", x, "por PesoxKg")
  ggplot(eeframe, aes_string('Carga', x)) +
    geom_boxplot() +
    ggtitle(title)
}
lapply(plotCols4, eebox)
```

8.7 APÉNDICE 7

Estadísticas luego de la normalización de datos de la familia de aceites.

Statistics

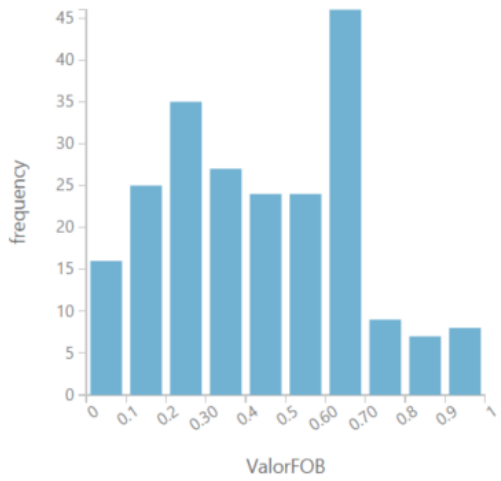
Mean	0.4391
Median	0.4311
Min	0
Max	1
Standard Deviation	0.2446
Unique Values	201
Missing Values	0
Feature Type	Numeric Feature

Visualizations

ValorFOB

Histogram

compare to None



Statistics

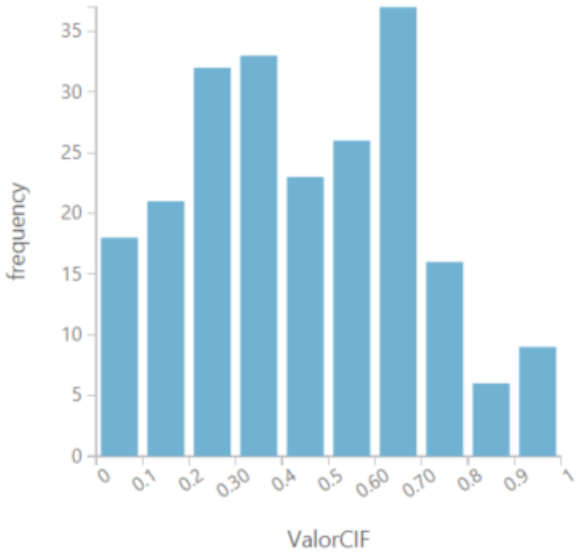
Mean	0.4484
Median	0.4303
Min	0
Max	1
Standard Deviation	0.2459
Unique Values	221
Missing Values	0
Feature Type	Numeric Feature

Visualizations

ValorCIF

Histogram

compare to None



Statistics

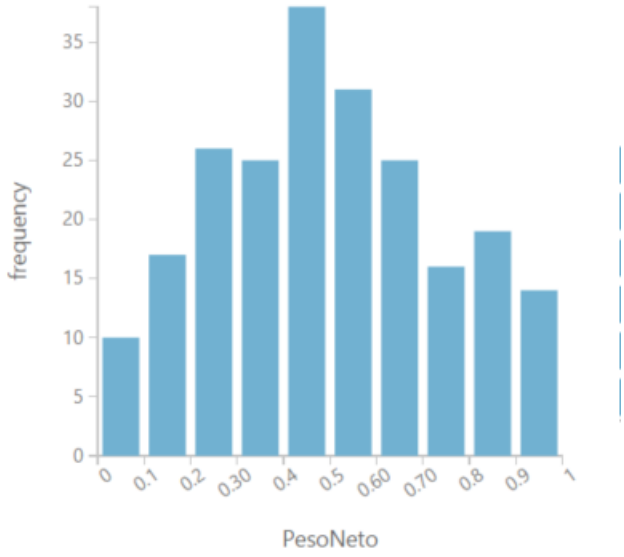
Mean	0.4963
Median	0.4785
Min	0
Max	1
Standard Deviation	0.2474
Unique Values	142
Missing Values	0
Feature Type	Numeric Feature

Visualizations

PesoNeto

Histogram

compare to None



Statistics

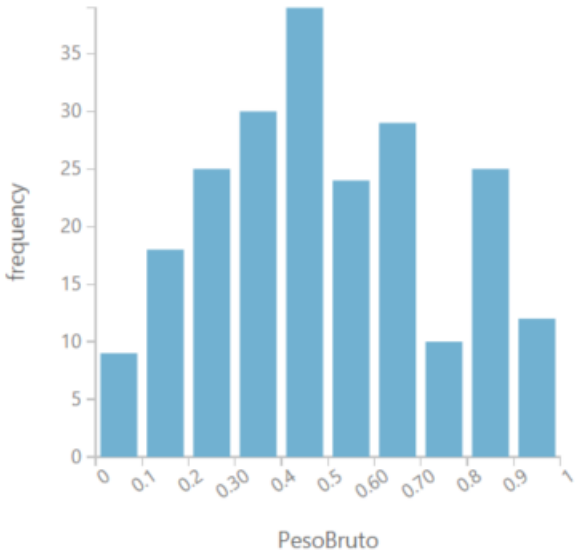
Mean	0.4915
Median	0.4734
Min	0
Max	1
Standard Deviation	0.2425
Unique Values	190
Missing Values	0
Feature Type	Numeric Feature

Visualizations

PesoBruto

Histogram

compare to None



Statistics

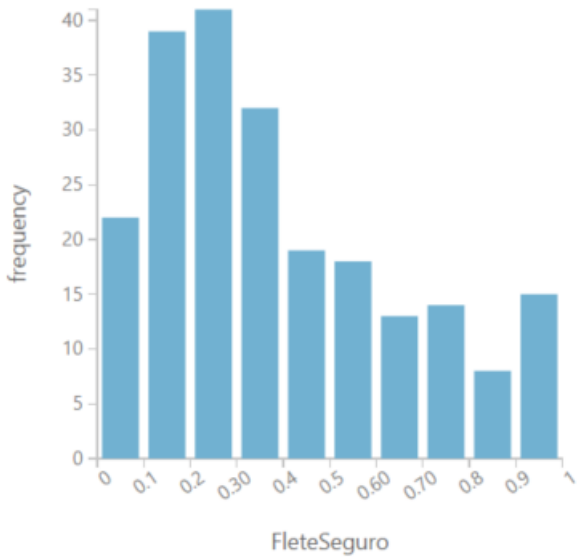
Mean	0.3919
Median	0.3246
Min	0
Max	1
Standard Deviation	0.2629
Unique Values	220
Missing Values	0
Feature Type	Numeric Feature

Visualizations

FleteSeguro

Histogram

compare to None



Statistics

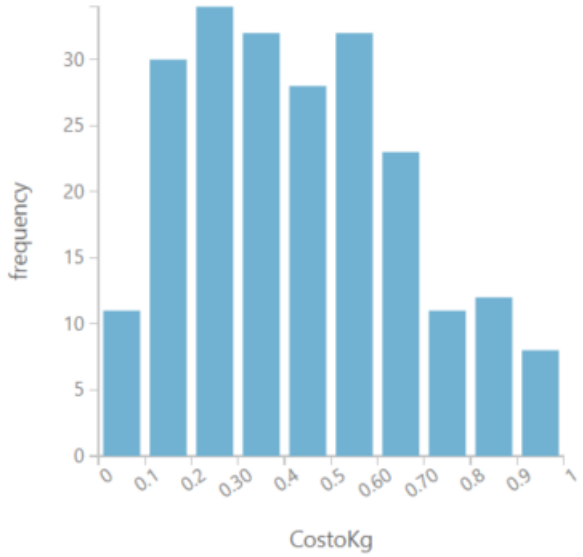
Mean	0.4337
Median	0.4264
Min	0
Max	1
Standard Deviation	0.2366
Unique Values	221
Missing Values	0
Feature Type	Numeric Feature

Visualizations

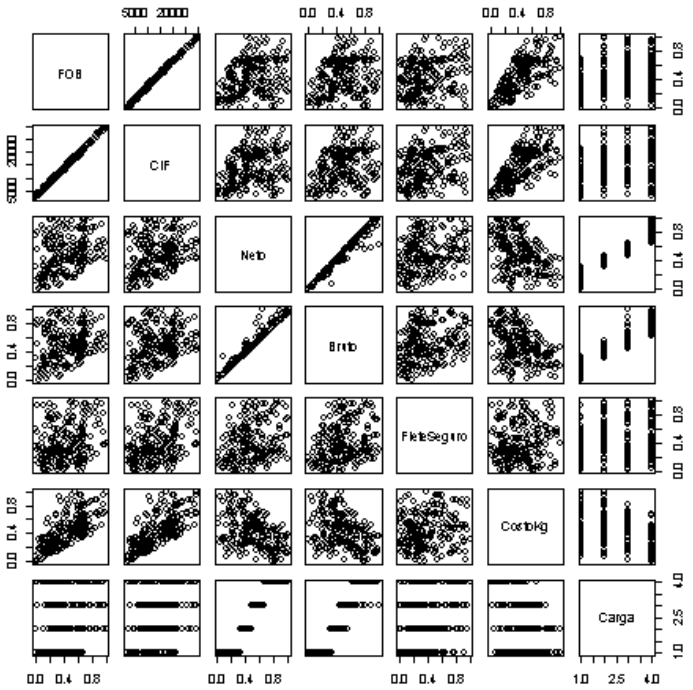
CostoKg

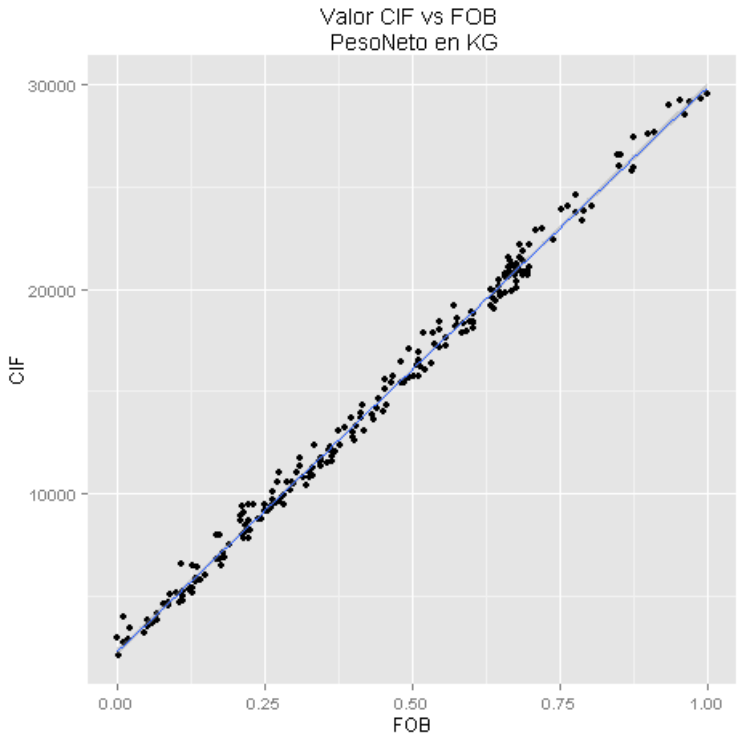
Histogram

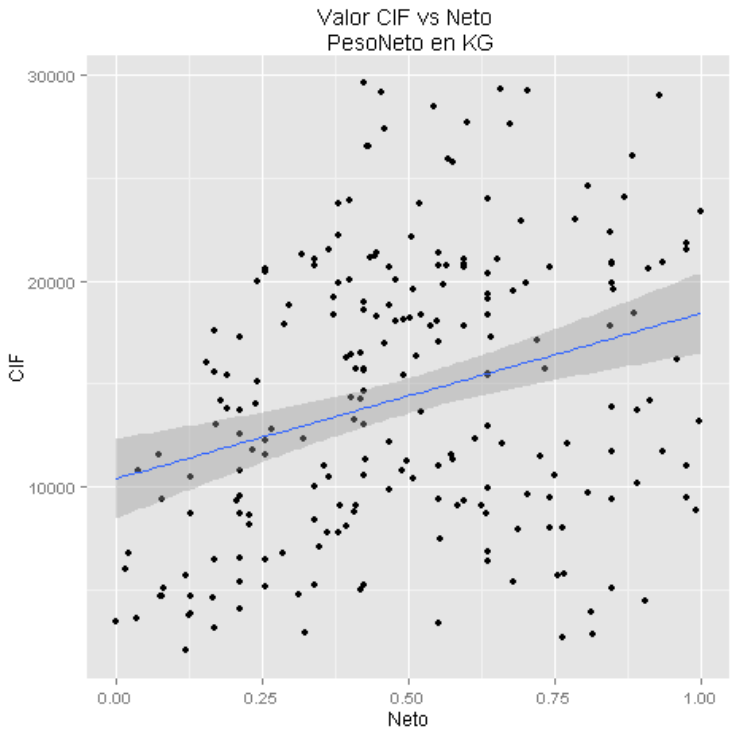
compare to None

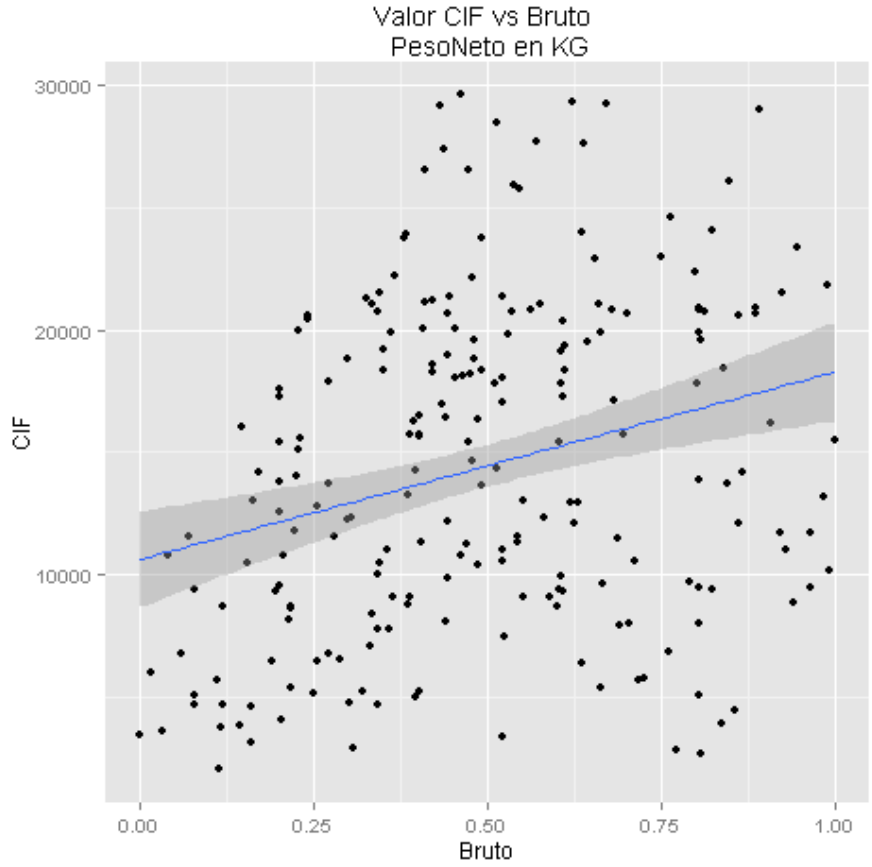


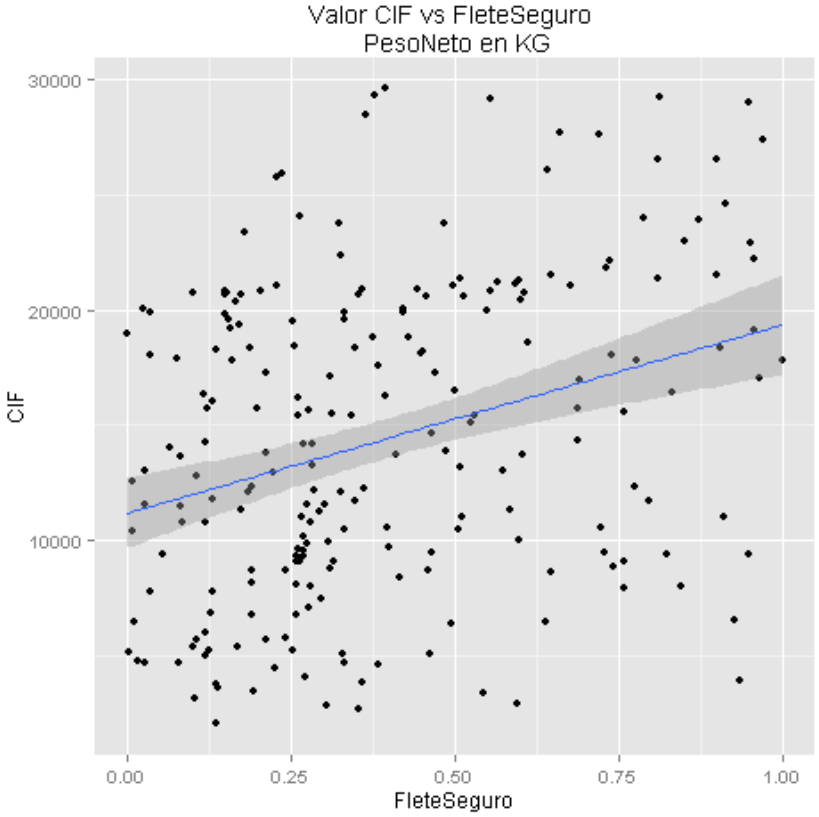
Análisis de relación entre las variables de la familia de aceites

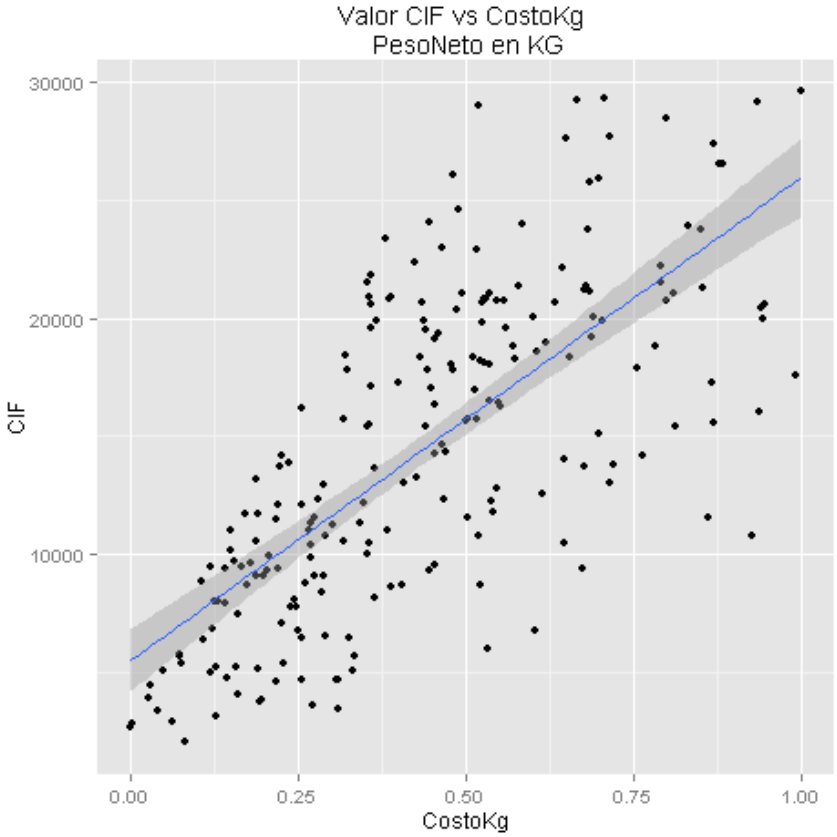

FOB
CIF
Neto
Bruto
FleteSeguro
CostoKg
Carga

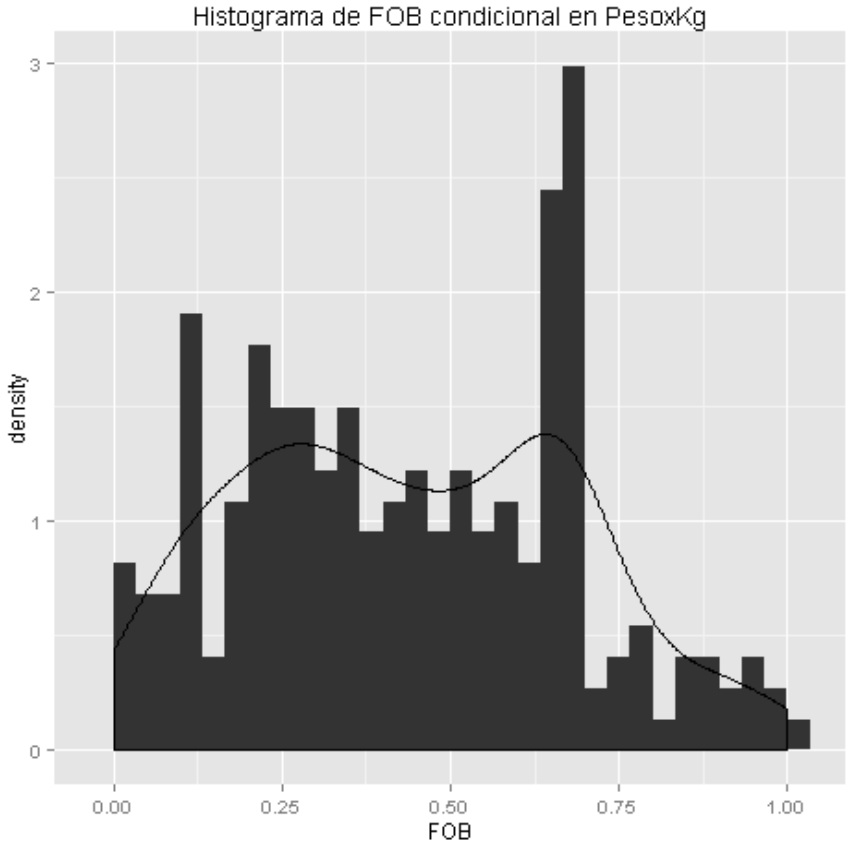

Valor CIF vs FOB
PesoNeto en KG
CIF
FOB
30000
20000
10000
0.00
0.25
0.50
0.75
1.00

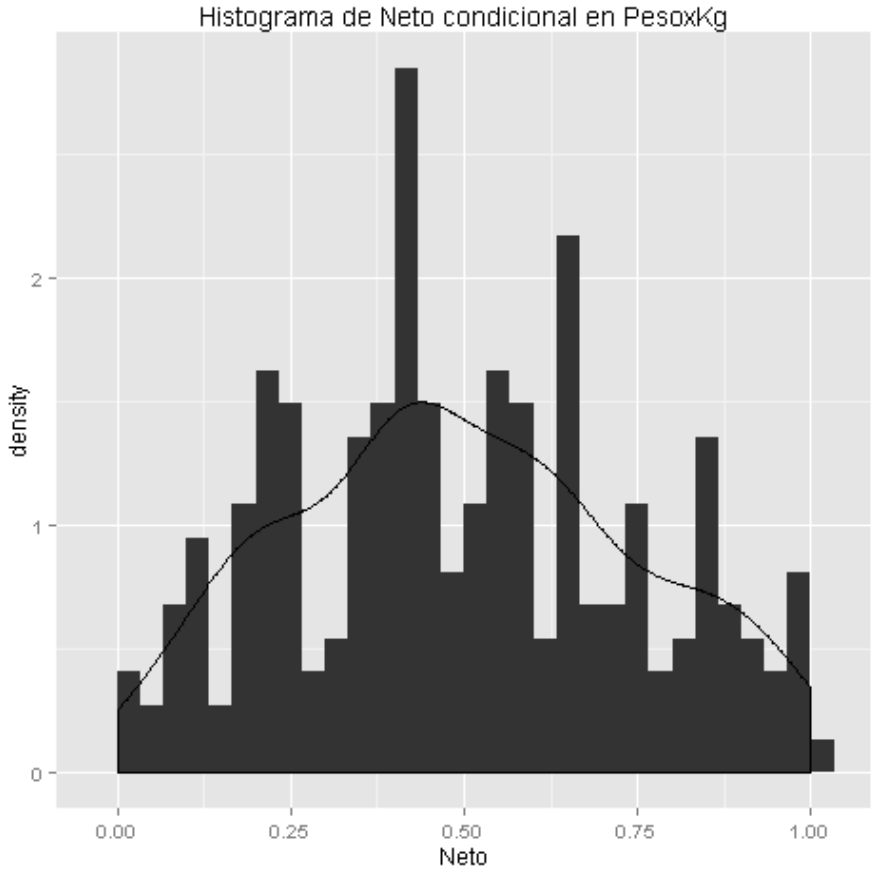

Valor CIF vs Neto
PesoNeto en KG
30000
20000
10000
CIF
0.00
0.25
0.50
0.75
1.00
Neto

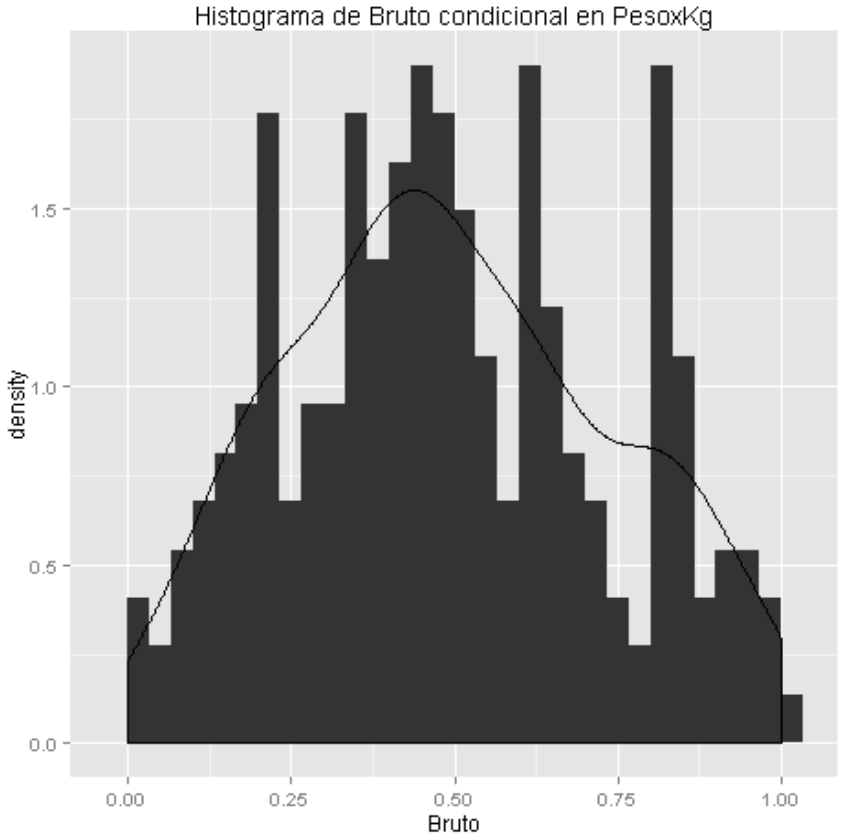

Valor CIF vs Bruto
PesoNeto en KG
30000
20000
10000
CIF
0.00
0.25
0.50
0.75
1.00
Bruto

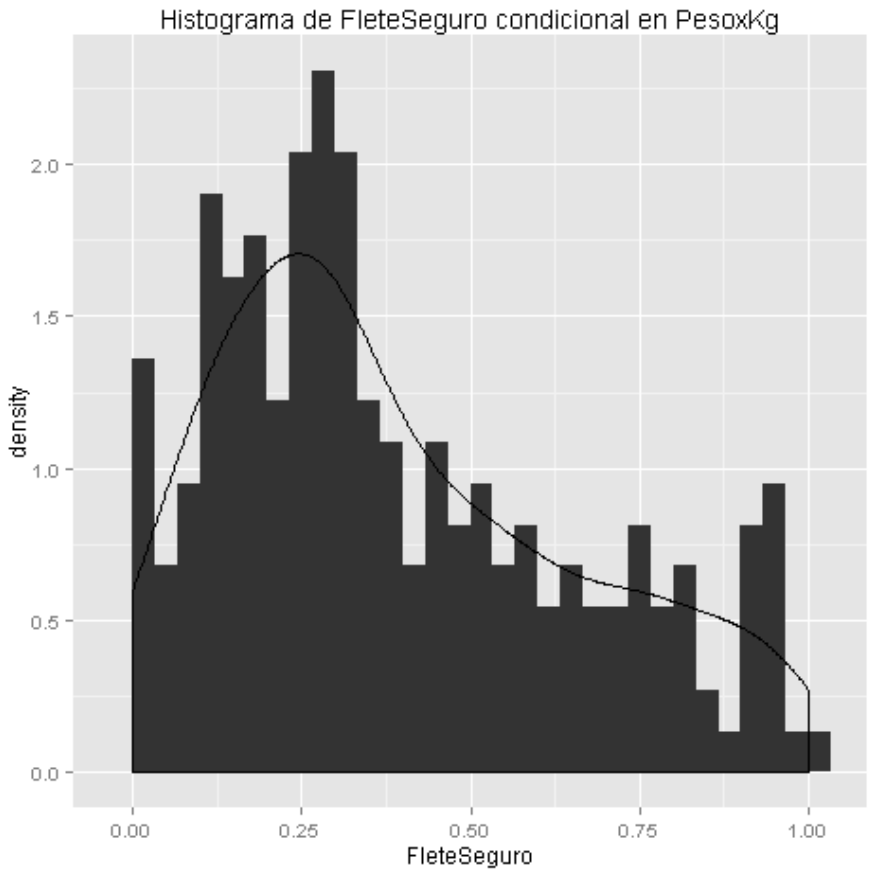

Valor CIF vs FleteSeguro
PesoNeto en KG
CIF
30000
20000
10000
0.00
0.25
0.50
0.75
1.00
FleteSeguro

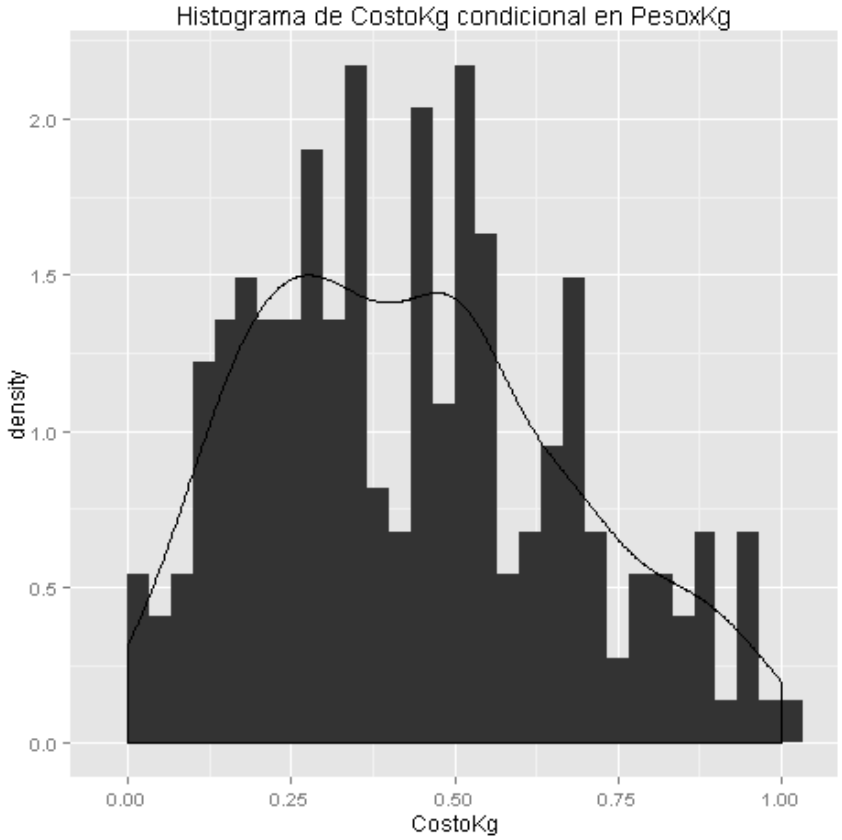

Valor CIF vs CostoKg
PesoNeto en KG
CIF
30000
20000
10000
0.00
0.25
0.50
0.75
1.00
CostoKg

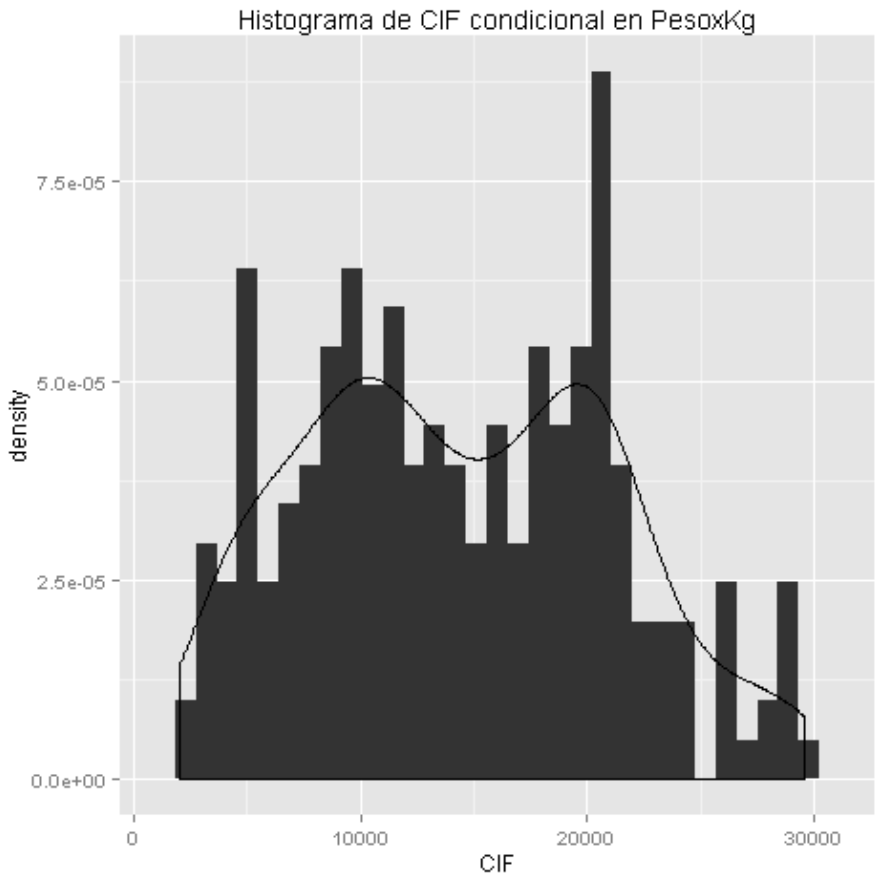

Histograma de FOB condicional en PesoxKg
density
FOB
0
1
2
3
0.00
0.25
0.50
0.75
1.00

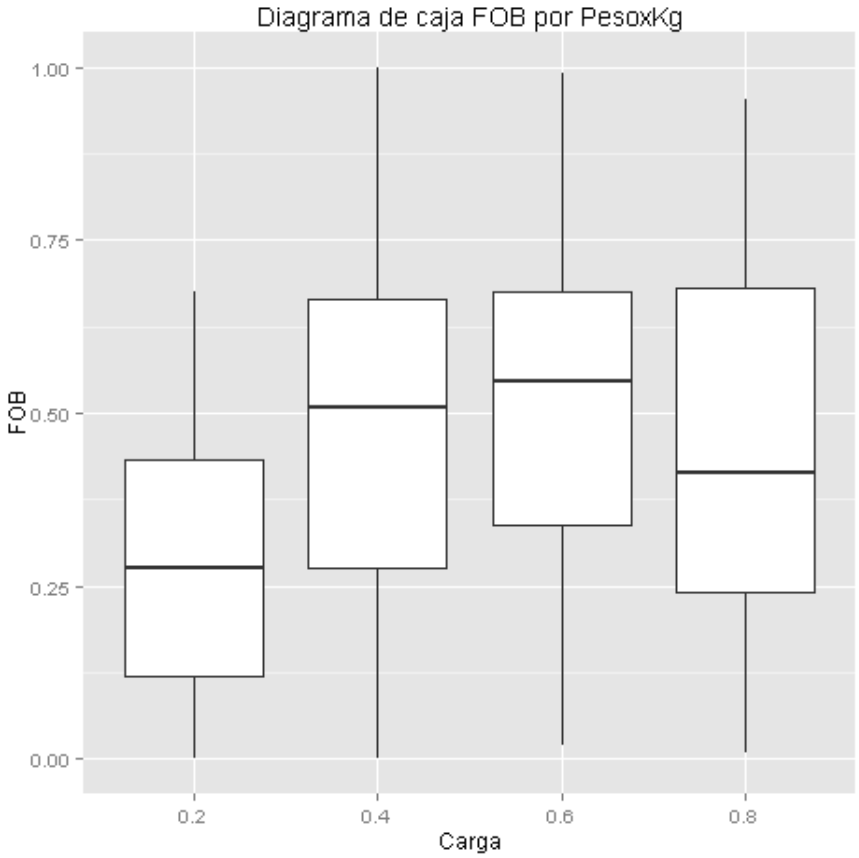

Histograma de Neto condicional en PesoxKg
density
Neto
0
1
2
0.00
0.25
0.50
0.75
1.00


Histograma de Bruto condicional en PesoxKg
density
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
Bruto


Histograma de FleteSeguro condicional en PesoxKg
density
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
FleteSeguro


Histograma de CostoKg condicional en PesoxKg
density
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
CostoKg


Histograma de CIF condicional en PesoxKg
density
7.5e-05
5.0e-05
2.5e-05
0.0e+00
0
10000
20000
30000
CIF



8.8 APÉNDICE 8

Código en R para la creación de variables polinomiales.

Execute R Script

R Script

```r
eeframe <- maml.mapInputPort(1)
library(dplyr)
eeframe = mutate(eeframe,
ValorFOB2 = ValorFOB^2,
PesoNeto2 = PesoNeto^2,
PesoBruto2 = PesoBruto^2 ,
Fleteseguro2 = FleteSeguro^2,
Costokg2 = CostoKg^2,
ValorFOB3 = ValorFOB^3,
PesoNeto3 = PesoNeto^3,
PesoBruto3 = PesoBruto^3 ,
Fleteseguro3 = FleteSeguro^3,
Costokg3 = CostoKg^3)
maml.mapOutputPort('eeframe')
```