04

Fecha de presentación: julio, 2020
Fecha de aceptación: septiembre, 2020
Fecha de publicación: octubre, 2020

PERFILAMIENTO DE PROVEEDORES

PARA LA COMPRA DE MATERIA PRIMA MEDIANTE EL DESARROLLO DE UN ALMACÉN DE DATOS Y TÉCNICAS DE MINERÍAS DE DATOS CON LENGUAJE R

PROFILING OF SUPPLIERS FOR THE PURCHASE OF RAW MATERIALS THROUGH THE DEVELOPMENT OF A DATAWAREHOUSE AND DATA MINING TECHNIQUES WITH R LANGUAGE

Estalin Vladimir Arrobo Lapo[1]
E-mail: us.estalinarrobo@uniandes.edu.ec
ORCID: https://orcid.org/0000-0001-7322-3297
Silvio Amable Machuca Vivar[1]
E-mail: us.silviomachuca@uniandes.edu.ec
ORCID: https://orcid.org/0000-0002-4681-3045
Erick Fernando Méndez Garcés[1]
E-mail: us.erikmendez@uniandes.edu.ec
ORCID: https://orcid.org/0000-0003-3009-1479
[1] Universidad Regional Autónoma de Los Andes. Ecuador.

Cita sugerida (APA, séptima edición)

Arrobo Lapo, E. V., Machuca Vivar, S. A., & Méndez Garcés, E. F. (2020). Perfilamiento de proveedores para la compra de materia prima mediante el desarrollo de un almacén de datos y técnicas de minerías de datos con lenguaje R. *Revista Universidad y Sociedad*, 12(S1), 31-38.

RESUMEN

Las decisiones empresariales de hoy se centran en datos, los mismos que tienen su origen a partir de distintas fuentes, una decisión importante para las empresas es la perfilación o clasificación de proveedores, lo que permite a los tomadores de decisiones realizar diferentes acciones sobre estos como acuerdos o compromisos de servicios. El objetivo del presente trabajo de investigación fue desarrollar un almacén de datos y analizar esta información utilizando técnicas de minería de datos. El datawarehouse se lo creó utilizando la metodología Hefesto, las técnicas de minería de datos se les realizó con el lenguaje de programación R, se obtuvo un reporte de proveedores con variables compatibles entre sí que permitieron ejecutar técnicas de minería de datos.

Palabras clave: Almacén de datos, inteligencia de negocios, pentaho, lenguaje R, agrupamiento.

ABSTRACT

Today's business decisions focus on data, the same that originate from different sources, an important decision for companies is the profiling or classification of suppliers, which allows decision-makers to carry out different actions on them, such as service agreements or commitments. The objective of this research work was to develop a data warehouse and analyze this information using data mining techniques. The Datawarehouse was created using the Hefesto methodology, the data mining techniques were carried out with the R programming language, a report was obtained from suppliers with variables compatible with each other that allowed data mining techniques to be executed.

Keywords: Data Warehouse, business intelligence, pentaho, R language, clustering.

INTRODUCCIÓN

Se vive en una época conocida como la era de la información, donde todas las actividades económicas tienen un fuerte vínculo con las tecnologías de la información y la comunicación, en este contexto la información generada por estas actividades constituye una nueva clase de activo económico como lo es la moneda o el oro, por tanto, se afirma que la información es poder. El presenta trabajo "Perfilamiento de proveedores para la compra de materia prima mediante el desarrollo de un datawarehouse y técnicas de minerías de datos con lenguaje R", un sinnúmero de veces la información que generan las empresas u organizaciones como producto de sus actividades transaccionales termina en un dispositivo de almacenamiento físico, sin dar cabida a una posible extracción de información valiosa para la empresa y peor aún sin haber realizado un análisis más completo como es la minería de datos.

Por tanto, la presente investigación atienda esta necesidad evidente brindando un gran aporte científico a las empresas que no han explotado de forma adecuada la información que poseen. Con el pasar del tiempo se han presentado varias metodologías de análisis de datos como los sistemas para la toma de decisiones, sistemas para extracción del conocimiento obteniendo un resultado de reportería, recopilando la información desde las bases transaccionales y obteniendo un reporte de datos dimensional, sin llegar a la complejidad de las técnicas de minería de datos.

En el presente trabajo de investigación se modelará e implementará un almacén de datos o datawarehouse a partir de un sistema transaccional como generador de información utilizando la metodología Hefesto, de este almacén de datos se obtendrá un reporte de proveedores dimensional con variables que permiten ser ejecutadas utilizando el lenguaje de programación R, el mismo que tiene un enfoque al análisis estadístico, obteniendo de esta forma el resultado de ejecutar las técnicas de minería de datos no supervisadas conocidas como agrupamiento o clustering. La población será todos los proveedores que la empresa tiene registrados producto de sus actividades realizadas, al contar con un lenguaje estadístico se facilitó el análisis por tanto se tomará como muestra el total de los proveedores, es necesario indicar que para el presente análisis se ha obviado los nombres de los proveedores, ya que el análisis se centra en la relación de las variables que cada uno posee.

MATERIALES Y MÉTODOS

Como se describió anteriormente, en primera instancia se desarrolló un almacén de datos, para esto se utilizó la metodología "Hefesto v3", ya que esta metodología el proceso eliminando cierta complejidad innecesaria de otras metodologías, una metodología base para este tipo de proyectos es la metodología Kimball & Ross (2013), que contiene los conceptos bases para el desarrollo de un almacén de datos de esta forma se logró entregar lo más pronto posible una primera implementación para mostrar las ventajas de la misma, las etapas fueron (Tabla 1):

Tabla 1. Fases de la metodología Hefesto v3.

1. Análisis de requerimientos	• Preguntas del negocio • Indicadores y perspectivas • Modelo Conceptual
2. Análisis de data sources	• Hechos e indicadores • Mapeo • Granularidad • Modelo conceptual ampliado
3. Modelo lógico del dw	• Tipología • Tablas de Dimensiones • Tablas de Hechos
4. Integración de datos	• Carga Inicial • Actualización

Fuente: Bernabeu & García Mattío (2017).

Luego de culminar con la creación del almacén de datos, se obtuvo un reporte de todos los proveedores registrados en la empresa, de igual forma de las diferentes variables que se pretenda analizar utilizando el lenguaje de programación estadística R, se utiliza la técnica de minería de datos no supervisada conocida como agrupamiento o clustering, con el Objetivo de analizar la relación entre dos variables del reporte mencionado, que a criterio del lector se deben tomar en cuenta para el clustering.

RESULTADOS Y DISCUSIÓN

Para el proyecto se utilizó la base de datos de un sistema transaccional, se obtuvo los requerimientos de los usuarios mediante preguntas de negocio. Luego se analizó estas preguntas para identificar los indicadores y las perspectivas o dimensiones, las mismas que se utilizaron para la construcción del modelo conceptual del almacén de datos.

Las siguientes fueron las preguntas de negocio (Tabla 2 y 3) (Figura 1):

Tabla 2. Prefunds del negocio.

Tema:	**Almacén de datos para el análisis multidimensional de la gestión de compras.**
Preguntas:	¿Qué cantidad se ha logrado comprar en los últimos años en kilogramos? ¿Qué porcentaje ha logrado captar la empresa de la disponibilidad total del grano en el país? ¿Cuáles son los proveedores que demuestran mayor fidelidad a los convenios de entrega con la empresa? ¿Qué proveedores entregar la materia prima de mejor calidad? ¿Cómo se relacionan los parámetros de calidad respecto al precio y volumen de compra que obtienen cada proveedor?
Fuentes de datos:	• Base de datos del sistema transaccional
Fuentes de información:	• Analistas de información. • Asistente agrícola • Asistente administrativo
Mecanismos de entrega de información:	• Cubo de análisis OLAP disponible en página web con estructura cliente – servidor.

Tabla 3. Identificar Indicadores y Perspectivas.

Perspectivas | Dimensiones:	• Tiempo • Centros de compras • Pesaje • Pagos • Proveedores • Artículos • Geografía
Indicadores:	• Peso bruto • Peso tara • Kilos recibidos • Cantidad en quintales • Cantidad kg subió a kardex • Costo neto qq • Costo compra qq • Humedad grano • Impureza grano • Líquido por pagar • Costro tratamiento por humedad • Descuento a proveedor • Retención en la fuente • Valor del flete

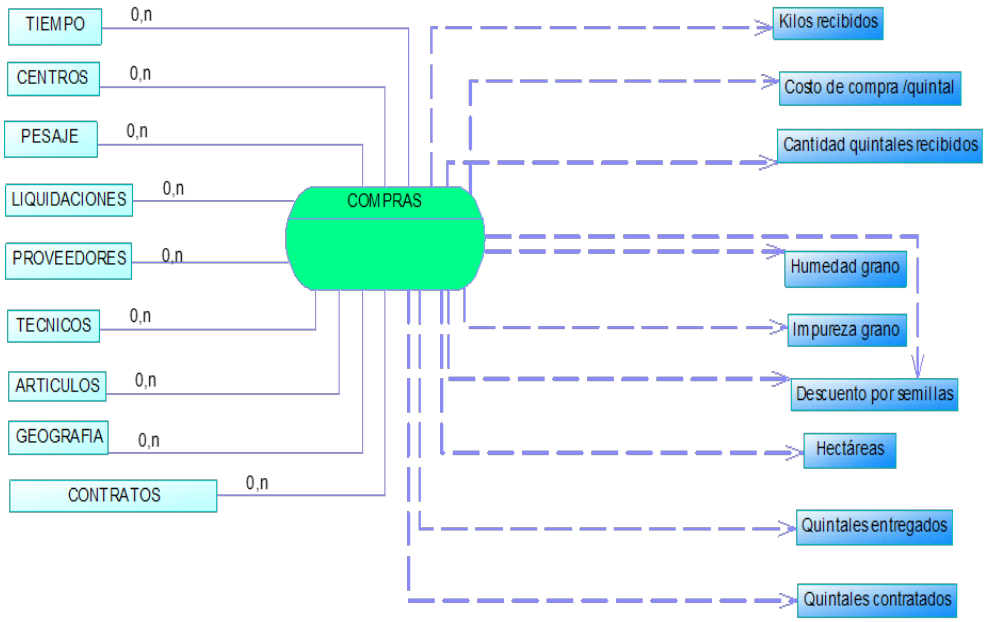


Figura 1. Modelo conceptual de almacén de datos.

En este punto se analiza las fuentes de información, en este caso la base de datos del sistema transaccional, con el objetivo de definir la fórmula de cálculo, en este caso los indicadores se los extraerá de la base transaccional y se utilizará en el modelo la función de sumarización para la suma (González Ortega, et al., 2019) (Tabla 4).

Tabla 4. Muestra de Indicadores.

Indicadores	Fórmula	Detalle
Kilos recibidos	Hechos: Kilos recibidos Función de sumarización: SUM	Sumatoria de los kilos recibidos por la compra del grano.
Cantidad quintales recibidos	Hechos: cantidad quintales recibido Función de sumarización: SUM	Sumatoria en quintales recibidos en los centros.
Quintales entregados	Hecho: quintales entregados Función de sumarización: SUM	Sumatoria de los quintales entregados por parte de los proveedores
Quintales contratados	Hecho: quintales contratados Función de sumarización: SUM	Sumatoria de los quintales contratados con los proveedores.

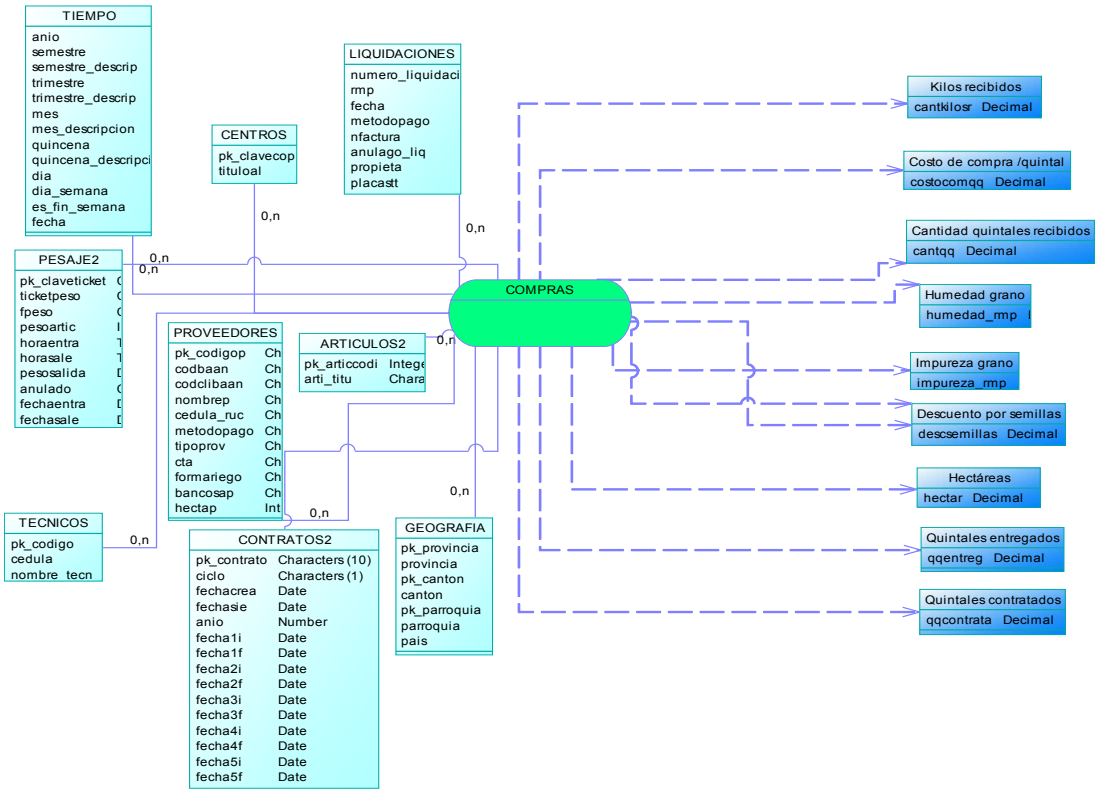


Figura 2. Modelo conceptual ampliado del almacén de datos.

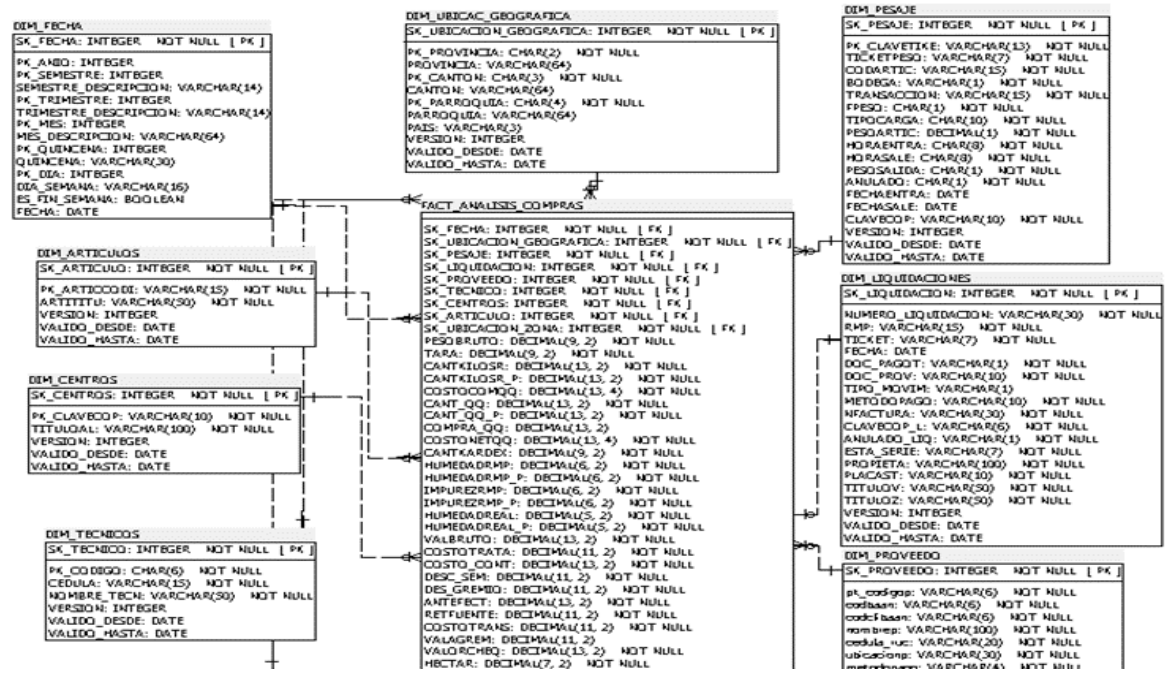


Figura 3. Modelo Lógico del Almacén de datos.

En esta etapa se realiza la ingesta de datos (Figura 2 y 3), realizando una selección desde las bases de datos transaccionales hacia el modelo del almacén de datos, utilizando una herramienta para la extracción, transformación y carga de datos, conocida como ETL (Extraction, Transformation, Load) (Figura 4).

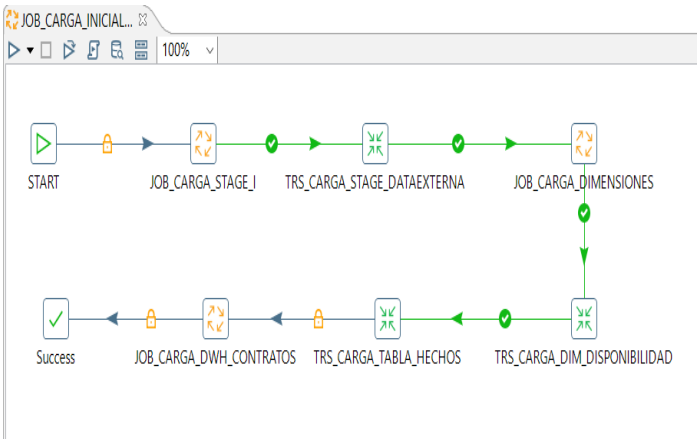
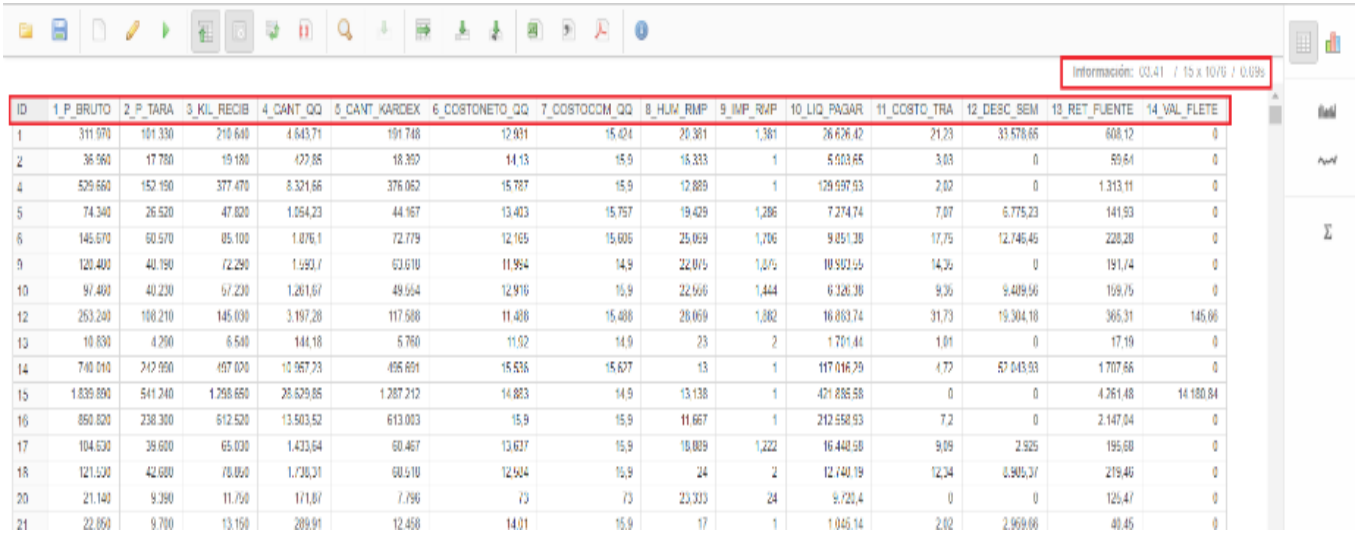


Figura 4. Job de carga inicial.

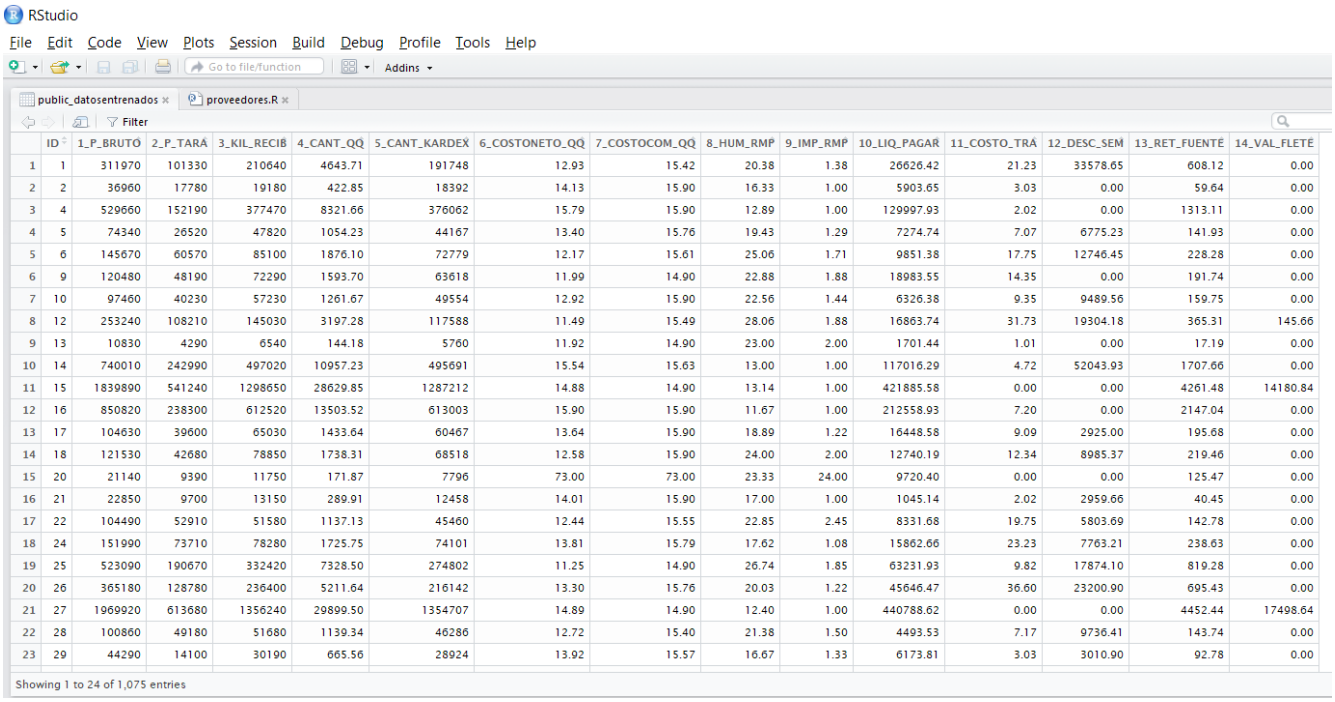
Con el almacén de datos implementado fue posible obtener el siguiente reporte (Figura 5):



ID	1_P_BRUTO	2_P_TARA	3_KIL_RECIB	4_CANT_QQ	5_CANT_KARDEX	6_COSTONETO_QQ	7_COSTOCOM_QQ	8_HUM_RMP	9_IMP_RMP	10_LIQ_PAGAR	11_COSTO_TRA	12_DESC_SEM	13_RET_FUENTE	14_VAL_FLETE
1	311.970	101.330	210.640	4.643,71	191.748	12,931	15,424	20,381	1,381	26.626,42	21,23	33.578,65	608,12	0
2	36.960	17.780	19.180	422,85	18.352	14,13	15,9	16,333	1	5.003,65	3,03	0	59,64	0
4	529.660	152.190	377.470	8.321,66	376.062	15,787	15,9	12,889	1	129.597,63	2,02	0	1.313,11	0
5	74.340	26.520	47.820	1.054,23	44.167	13,403	15,757	19,429	1,286	7.274,74	7,07	6.775,23	141,93	0
6	146.670	60.570	86.100	1.876,1	72.779	12,165	15,606	25,059	1,706	9.851,38	17,75	12.746,45	228,28	0
9	120.400	40.190	72.290	1.593,7	63.610	11,994	14,9	22,075	1,675	10.503,55	14,35	0	191,74	0
10	97.480	40.230	67.230	1.261,67	49.564	12,916	15,9	22,966	1,444	6.326,38	9,35	9.409,66	109,75	0
12	253.240	108.210	145.030	3.197,28	117.588	11,428	15,488	28,069	1,882	16.863,74	31,73	19.304,18	365,31	145,66
13	10.830	4.290	6.540	144,18	5.760	11,52	14,9	23	2	1.701,44	1,01	0	17,19	0
14	740.010	242.990	497.020	10.967,23	495.691	15,536	15,627	13	1	117.016,29	4,77	57.043,93	1.707,66	0
15	1.839.890	541.240	1.298.650	28.629,85	1.287.212	14,883	14,9	13,138	1	421.886,58	0	0	4.261,48	14.180,84
16	860.820	238.300	612.520	13.503,52	613.003	15,9	15,9	11,667	1	212.558,93	7,2	0	2.147,04	0
17	104.630	39.600	65.030	1.433,64	60.467	13,637	15,9	18,889	1,222	16.448,58	9,09	2.925	195,68	0
18	121.530	42.680	78.850	1.738,31	80.510	12,504	15,9	24	2	12.740,19	12,34	8.905,37	219,46	0
20	21.140	9.390	11.750	171,87	7.796	75	75	23,333	24	9.720,4	0	0	125,47	0
21	22.850	9.700	13.150	289,91	12.458	14,01	15,9	17	1	1.045,14	2,02	2.969,66	40,45	0

Figura 5. Reporte obtenido del almacén de datos

Se exportó el reporte (Figura 6) al formato csv, ya que este formato es universal para las herramientas de análisis estadístico. Se procedió a cargar el dataset csv al programa R Studio, en el cual se ejecutará el algoritmo para el análisis de minería de datos utilizando la técnica de Clustering.



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

public_datosentrenados × proveedores.R ×

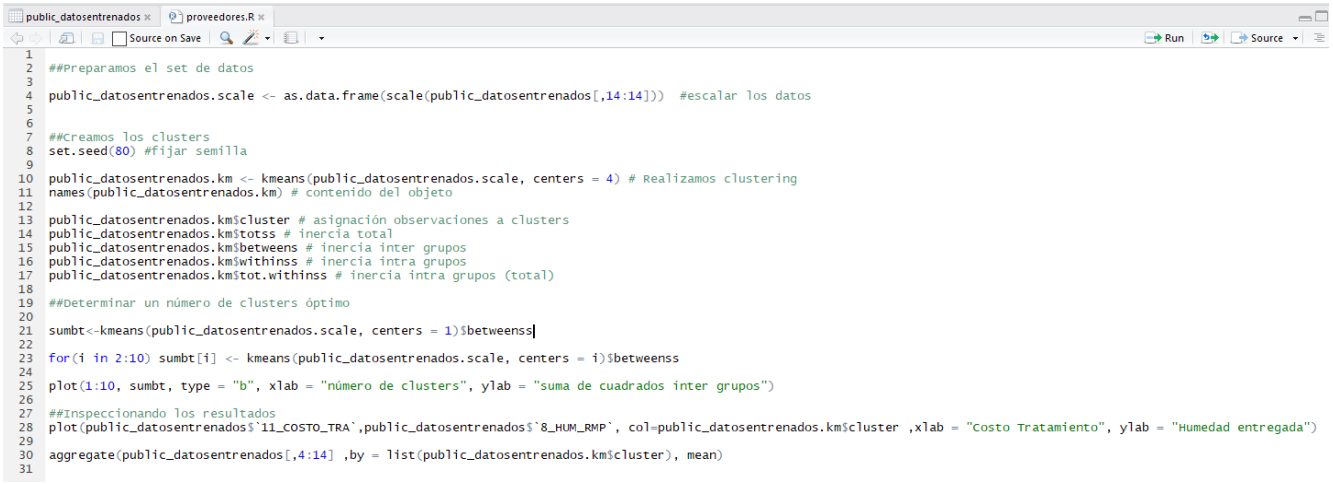
Filter

	ID	1_P_BRUTO	2_P_TARA	3_KIL_RECIB	4_CANT_QQ	5_CANT_KARDEX	6_COSTONETO_QQ	7_COSTOCOM_QQ	8_HUM_RMP	9_IMP_RMP	10_LIQ_PAGAR	11_COSTO_TRA	12_DESC_SEM	13_RET_FUENTE	14_VAL_FLETE
1	1	311970	101330	210640	4643.71	191748	12.93	15.42	20.38	1.38	26626.42	21.23	33578.65	608.12	0.00
2	2	36960	17780	19180	422.85	18392	14.13	15.90	16.33	1.00	5903.65	3.03	0.00	59.64	0.00
3	4	529660	152190	377470	8321.66	376062	15.79	15.90	12.89	1.00	129997.93	2.02	0.00	1313.11	0.00
4	5	74340	26520	47820	1054.23	44167	13.40	15.76	19.43	1.29	7274.74	7.07	6775.23	141.93	0.00
5	6	145670	60570	85100	1876.10	72779	12.17	15.61	25.06	1.71	9851.38	17.75	12746.45	228.28	0.00
6	9	120480	48190	72290	1593.70	63618	11.99	14.90	22.88	1.88	18983.55	14.35	0.00	191.74	0.00
7	10	97460	40230	57230	1261.67	49554	12.92	15.90	22.56	1.44	6326.38	9.35	9489.56	159.75	0.00
8	12	253240	108210	145030	3197.28	117588	11.49	15.49	28.06	1.88	16863.74	31.73	19304.18	365.31	145.66
9	13	10830	4290	6540	144.18	5760	11.92	14.90	23.00	2.00	1701.44	1.01	0.00	17.19	0.00
10	14	740010	242990	497020	10957.23	495691	15.54	15.63	13.00	1.00	117016.29	4.72	52043.93	1707.66	0.00
11	15	1839890	541240	1298650	28629.85	1287212	14.88	14.90	13.14	1.00	421885.58	0.00	0.00	4261.48	14180.84
12	16	850820	238300	612520	13503.52	613003	15.90	15.90	11.67	1.00	212558.93	7.20	0.00	2147.04	0.00
13	17	104630	39600	65030	1433.64	60467	13.64	15.90	18.89	1.22	16448.58	9.09	2925.00	195.68	0.00
14	18	121530	42680	78850	1738.31	68518	12.58	15.90	24.00	2.00	12740.19	12.34	8985.37	219.46	0.00
15	20	21140	9390	11750	171.87	7796	73.00	73.00	23.33	24.00	9720.40	0.00	0.00	125.47	0.00
16	21	22850	9700	13150	289.91	12458	14.01	15.90	17.00	1.00	1045.14	2.02	2959.66	40.45	0.00
17	22	104490	52910	51580	1137.13	45460	12.44	15.55	22.85	2.45	8331.68	19.75	5803.69	142.78	0.00
18	24	151990	73710	78280	1725.75	74101	13.81	15.79	17.62	1.08	15862.66	23.23	7763.21	238.63	0.00
19	25	523090	190670	332420	7328.50	274802	11.25	14.90	26.74	1.85	63231.93	9.82	17874.10	819.28	0.00
20	26	365180	128780	236400	5211.64	216142	13.30	15.76	20.03	1.22	45646.47	36.60	23200.90	695.43	0.00
21	27	1969920	613680	1356240	29899.50	1354707	14.89	14.90	12.40	1.00	440788.62	0.00	0.00	4452.44	17498.64
22	28	100860	49180	51680	1139.34	46286	12.72	15.40	21.38	1.50	4493.53	7.17	9736.41	143.74	0.00
23	29	44290	14100	30190	665.56	28924	13.92	15.57	16.67	1.33	6173.81	3.03	3010.90	92.78	0.00

Showing 1 to 24 of 1,075 entries

Figura 6. Carga del data set de proveedores a R Studio.

Se presenta el Código en lenguaje R (Figura 7), el mismo que se ejecutó para obtener el análisis de Clustering.



public_datosentrenados × proveedores.R ×

Source on Save | Run | Source

```
##Preparamos el set de datos

public_datosentrenados.scale <- as.data.frame(scale(public_datosentrenados[,14:14]))  #escalar los datos


##Creamos los clusters
set.seed(80) #fijar semilla

public_datosentrenados.km <- kmeans(public_datosentrenados.scale, centers = 4) # Realizamos clustering
names(public_datosentrenados.km) # contenido del objeto

public_datosentrenados.km$cluster # asignación observaciones a clusters
public_datosentrenados.km$totss # inercia total
public_datosentrenados.km$betweens # inercia inter grupos
public_datosentrenados.km$withinss # inercia intra grupos
public_datosentrenados.km$tot.withinss # inercia intra grupos (total)

##Determinar un número de clusters óptimo

sumbt<-kmeans(public_datosentrenados.scale, centers = 1)$betweenss

for(i in 2:10) sumbt[i] <- kmeans(public_datosentrenados.scale, centers = i)$betweenss

plot(1:10, sumbt, type = "b", xlab = "número de clusters", ylab = "suma de cuadrados inter grupos")

##Inspeccionando los resultados
plot(public_datosentrenados$`11_COSTO_TRA`,public_datosentrenados$`8_HUM_RMP`, col=public_datosentrenados.km$cluster ,xlab = "Costo Tratamiento", ylab = "Humedad entregada")

aggregate(public_datosentrenados[,4:14] ,by = list(public_datosentrenados.km$cluster), mean)

```

Figura 7. Código en lenguaje R.

La técnica de Clustering utilizada fue K-medias, la misma que busca que los elementos que pertenecen a un grupo sean los más homogéneos posible entre sí y al mismo tiempo obtener la máxima heterogeneidad entre los distintos grupos, luego de la ejecución se obtuvo 4 clústeres (Figura 8).

```
> aggregate(public_datosentrenados[,4:14] ,by = list(public_datosentrenados.km$cluster), mean)
  Group.1 3_KIL_RECIB  4_CANT_QQ 5_CANT_KARDEX 6_COSTONETO_QQ 7_COSTOCOM_QQ 8_HUM_RMP 9_IMP_RMP 10_LIQ_PAGAR 11_COSTO_TRA 12_DESC_SEM 13_RET_FUENTE
1       1 10764918.6 237321.836    10576873.0       14.80714      15.29857  14.27857  1.074286   3289290.54    112.68571  217141.799     35418.536
2       2   182685.4   4004.355      173905.6       15.50408      17.70894  20.17115  2.130385     50965.37      9.59360    7668.265       319.647
3       3  4358166.8  96079.521     4320230.1       15.02421      15.30947  13.47053  1.062105   1334870.42     16.56684  107381.854     14568.204
4       4  1700348.7  36999.591     1652460.0       15.90587      16.56587  15.12365  1.301429    519261.61     21.78524   39192.906      5666.593
>
```

Figura 8. Cuatro clústeres generados con R Studio.

Se tomó como referencia de análisis (Figura 9), las variables “Costo de tratamiento” frente a la “Humedad entregada”, obteniendo el siguiente gráfico.

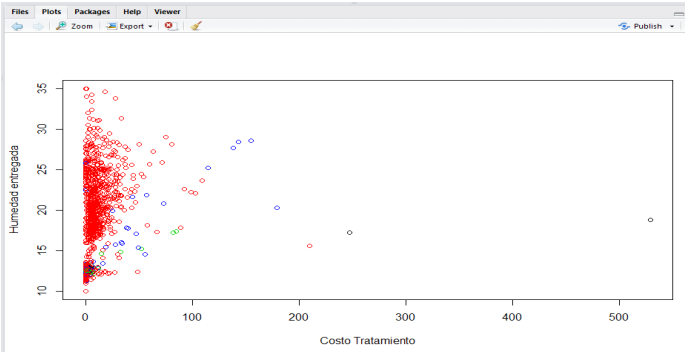


Figura 9. Plot de la técnica no supervisada clustering.

Se observa en el análisis de clustering que hay una mayor concentración el primero clúster graficado de color rojo, la misma que indica se ha generado un costo de tratamiento bajo o cero para los proveedores que han entregado materia prima con una humedad de hasta 27 puntos.

Este resultado se presenta debida a que las entregas que realizan los proveedores tienen un costo bajo, ya que la humedad no alcanza picos altos, hay que tomar en cuenta que también existen valores dispersos con valores altos de tratamiento, esto es justificable porque a más de la humedad existen otros parámetros que en caso de tener valores altos requerirían también tratamiento, como las impurezas, mala calidad de grano, etc.

CONCLUSIONES

El presente análisis de clustering respecto a la información de entrega de materia prima de los proveedores es muy convincente, debido a que se complementa en primera instancia con la creación de un almacén de datos o conocido como datawarehouse, para disponer del dataset apropiado para proceder al análisis con R.

Este proyecto deja un referente para los consumidores de datos que requieran de una guía completa desde la extracción de datos transaccionales, pasando por normalización de datos con el datawarehouse hasta el análisis de minería de datos con R, convirtiéndose en un proyecto íntegro respecto a la analítica de datos.

REFERENCIAS BIBLIOGRÁFICAS

Bernabeu, D., & García Mattío, M. (2017). Hefesto Data Warehousing V3. https://raw.githubusercontent.com/magm3333/materialClases/master/Hefesto_v3.pdf

González Ortega, R., Iviedo Rodríguez, M. D., Leyva Vázquez, M.., Estupiñán Ricardo, J., Sganderla Figueiredo, J. A., & Smarandache, F. (2019). Pestel analysis based on neutrosophic cognitive maps and neutrosophic numbers for the sinos river basin management. Infinite Study.

Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit. Wiley.