PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

MODELO PREDICTIVO DE IDENTIFICACIÓN DE FALLAS EN LA INDUSTRIA MANUFACTURERA

YADIRA JAZMIN QUISPE VÁSQUEZ

Profesor Supervisor:

JORGE VERA ANDREO.

Santiago de Chile, Marzo, 2016

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

MODELO PREDICTIVO DE IDENTIFICACIÓN DE FALLAS EN LA INDUSTRIA MANUFACTURERA

YADIRA JAZMIN QUISPE VÁSQUEZ

Tesis (Proyecto) presentada(o) a la Comisión integrada por los profesores:

JORGE VERA A.

EDUARDO KIRBERG B.

SERGIO MATURANA V.

Actividad de Graduación para completar las exigencias del grado de:

Magíster en Ingeniería Industrial

Santiago de Chile, (Marzo 2016)

DEDICATORIA

Dedicada a mis padres Micaela y Efraín, quienes son la energía constante que me impulsa en cada etapa de mi vida, por su incondicional apoyo, perfectamente mantenido en el tiempo y por siempre creer en mí.

Dios y la Virgen María por la salud que me permitió lograr mis objetivos, además por su infinita bondad y amor.

AGRADECIMIENTOS

Esta actividad de graduación, fue posible gracias al esfuerzo de distintas personas, quienes participaron, de forma directa o indirecta, opinando, corrigiendo, gestionando y acompañándome durante todo el proceso y a quienes en este apartado deseo agradecer.

En primer lugar, al Ing. Jaime Caiceo, un amplio agradecimiento, por haber confiado en este trabajo, por su valiosa dirección, paciencia y apoyo para concluir este proyecto académico.

Su dominio en el área de estudio, fue mi fuente de motivación y de curiosidad durante todo este tiempo.

Al Ing. Jorge Vera, un especial agradecimiento, por manifestar siempre una gran disponibilidad para atender mis dudas, por sus consejos, su sabiduría y apoyo desde el inicio del proyecto. Resalto su calidad como persona y profesional.

Al Ing. Miguel Paz Soldán, mi jefe y gran amigo, que siempre confía en mí y me apoya de forma incondicional. Su participación, fue la pieza más importante para el desarrollo de este proyecto.

Al Ing. Maximiliano Hurtado, un agradecimiento muy especial por ser la persona que siempre me brindo actitudes y palabras motivadoras. Resalto la dedicación que muestra en cada una de las presentaciones del taller.

A mis amigas, Danny Vargas, María Fernanda Silva y Loretto González, por acompañarme desde el inicio de este gran proyecto académico y por brindarme una amistad incondicional, que traspasa fronteras.

Finalmente, y no por ello menos importante, un agradecimiento especial a aquellas encantadoras personas que forman parte del equipo de administración del MII, Carmen Harambour, Ángela Saba y Marcela Berrios, por su impecable gestión en todo el proceso y por su gran espíritu, que alegraron mis días durante los dos años de estudio.

A todos ustedes, mi mayor reconocimiento y gratitud.

INDICE GENERAL

INDICE DE TABLAS

INDICE DE FIGURAS

RESUMEN

La industria manufacturera de Perú, específicamente de línea blanca, ha presentado en los últimos años un crecimiento importante en el mercado. Al mismo tiempo, ha aumentado considerablemente la exigencia en la calidad de sus productos por parte de sus clientes. En dicho escenario se ha evidenciado que si ambos factores no se combinan de una forma adecuada, uno de los principales problemas que ocurre, es el aumento de los porcentajes de fallas en los productos, lo cual se complica cuando no existe el soporte eficiente de herramientas sofisticas para el análisis de las bases de datos de los registros de fallas y, por consiguiente, se dificulta el planteamiento de acciones tanto correctivas y como preventivas.

Adicional hoy en día estando las industrias inmersas en un mercado, global, volátil y dinámico, se torna esencial integrar en las organizaciones un proceso de toma de decisiones soportado por modelos predictivos que se actualicen de forma continua.

Es por todo ello que en este trabajo de graduación se presenta una forma de análisis a través del Big Data, con el objetivo de identificar patrones que permitan predecir el origen de las fallas en el futuro. Dicho análisis hace uso de una herramienta de Business Analytics, llamada RapidMiner, en donde se aplican dos métodos de inteligencia artificial llamados: Árbol de decisión y Clustering.

El modelo predictivo diseñado con el método "Árbol de Decisión" logró una confiabilidad del 87.86 %, es decir, que el modelo es capaz de acertar con las reglas de diseño expuestas, en un 87.86 % de los casos.
Identificando patrones en Refrigeradores que permiten predecir origen de fallas en: Partes plásticas y el sistema eléctrico. Relacionando variables como: País de producción, color, tiempo de almacenamiento y tiempo de uso.
Y en cocinas prediciendo el origen de fallas en: partes metálicas y sistema eléctrico. Relacionando variables como: País de producción, tiempo de almacenamiento y tiempo de uso.

Y por último, en relación al método de "Clustering", el modelo identificó dos cluster bien definidos, tanto para cocinas y refrigeradoras, siendo las variables más relevantes para identificar dichos Clusters: Modelo y ubicación del daño.

En conclusión el software utilizado busca de forma automática y continua, patrones para activar las distintas alertas. Estas alertas ayudan a los gestores a conocer con suficiente antelación las consecuencias de sus acciones, ayudándolos a tomar las decisiones y acciones correctivas, en función de los objetivos establecidos por la compañía.

ABSTRACT

The manufacturing industry in Peru, specifically manufacturers of home appliance components, currently, have displayed a high market growth index, and an increasing demand for higher quality products. Under this scenario, it has become evident that if both component are not combining in a proper way, one of main problems that occurs is there will be an increase in the failure rate of their products. In turn, this will become an even bigger problem if the manufacturing sector does not have proper tech support or neither counts with sophisticated tools that would allow them to analyse its registry of failures. Under such circumstances it becomes very difficult to plan both corrective and preventive actions.

Nowadays these companies are immersed in a global, volatile and dynamic market, so it is essential to integrate them in a decision-making process supported by predictive models that are permanently being upgraded.

For these reasons this Graduation Project will focus on presenting a way to use Big Data to identify and analyse patterns that will allow the user to predict the cause of malfunction. This analysis will make use of RapidMiner, a Business Analytics tool based on two AI methods: Decision Tree and Clustering.

The predictive model designed with the Decision Tree achieved a reliability rate of 87.86%. In other words, this model is able to predict the outcome of the analysed designs 87.86% of the time.

Through the patterns identification on fridges produced, we can predict the cause of malfunction in plastic parts or the electrical system. Some of the variables used are: Country of origin, colour, storage time and operation time.
In the case of kitchens we can predict the cause of malfunction in metal parts or the electrical system. The variables used are: Country of Origin, storage time and operation time.

Finally, regarding the method of "Clustering", the model identified two well-defined clusters as the most relevant ones: Model and location of the malfunction. This applies both for fridges and kitchens.

We can conclude that the software used is able to search for patterns continuously and automatically, and alert the user when an anomaly is found. These alerts allows the managers to estimate the consequences of their actions in advance, helping them to take better decisions and achieve the goals set by the company.

1. INTRODUCCIÓN

1.1. Contexto general

La investigación se realiza en una empresa que por un tema de confidencialidad de la información, llamaremos Manufacturera C & R SAC. Filial de una empresa ecuatoriana de gran éxito en la región andina dedicada a la producción y comercialización de productos de línea blanca, que nace en 1972 en Cuenca – Ecuador, con el objetivo de producir electrodomésticos que faciliten las labores del hogar y cumplan con los más altos estándares de diseño y tecnología.
En 1992 se iniciaron las gestiones de exportación al mercado andino hacia países como Perú, Colombia y Venezuela, mercados en los que creció permanentemente su participación. Desde esta fecha hasta 1996 la empresa matriz adoptó la estrategia de dar prioridad a la fabricación de productos con marca propia, con el propósito de posicionar la marca en los mercados en que participa.

En el 2004 se inicia el proceso de exportación a Centroamérica, principalmente a países como Panamá, Costa Rica, Nicaragua, Honduras, El Salvador y Guatemala.

Para el año 2005 se inician operaciones en Colombia con la apertura de una filial en la comercialización del producto y desarrollo de la marca. El 2007 se inicia la comercialización en el Caribe, teniendo presencia en Belice, Santa Lucía, Trinidad y Tobago, Guyana, entre otros países.

Inicio e implementación de Manufacturera SAC.

En el año 2008 debido al éxito de las ventas de electrodomésticos en el Perú por encima de otros países de la región, copando más del 25% de exportaciones que realizaban desde la casa matriz Ecuador y debido al crecimiento del producto bruto interno (PBI), el Gerente General de la Empresa matriz, decidió aceptar el proyecto "construir una planta en el Perú". Para ello el Directorio basó su decisión en los objetivos de posicionar la marca en la región, facilitar la exportación e incrementar las ventas.

Es así como en el 2009 nace Manufacturera C & R SAC. de capital ecuatoriano, en donde se invirtió cerca de 86 millones de soles en la construcción de la planta industrial en Perú. La construcción y los permisos legales duraron cerca de dos años y para el año 2010 se inicia operaciones a través de Manufacturera C & R SAC. La planta tiene una extensión de 60.000 metros cuadrados, con una capacidad de 390,000 productos entre cocinas y refrigeradoras.
El ámbito de negocio de "Manufacturera C & R SAC.", es la producción y comercialización de electrodomésticos, con productos de gama alta. Su participación de mercado es del 30% y distribuye sus productos a las más grandes cadenas de retail, así como a los diversos mayoristas ubicados a lo largo de todo el Perú.

Certificaciones obtenidas

Manufacturera C & R SAC, obtuvo la certificación en la norma ISO 9001:2008 en noviembre del 2012, teniendo como alcance la fabricación de cocinas de uso doméstico. En el año 2014, la alta dirección decidió incluir en el alcance al área comercial y realizar un proyecto de implementación para las normas 14001 (Medio Ambiente) y OHSAS 18001 (Seguridad y Salud ocupacional).

En octubre del 2015, participó de la auditoria de certificación por parte de la empresa colombiana ICONTEC, obteniendo como resultado la re-certificación en la ISO 9001:2008 y la certificación en las normas ISO 14001 y OHSAS 18001, con el nuevo alcance "Fabricación y comercialización de cocinas de uso doméstico", gracias a su compromiso con el consumidor y el interés en implementar sistemas de gestión que colaboren con el cumplimiento y la mejora de procesos estratégicos, comerciales, financieros y operacionales.

Actualmente el Sistema de Gestión Integrado, se encuentra en un proceso de mejora continua y próximamente se tiene planificado implementar la norma ISO 17025 en los laboratorios de la organización.

Situación del mercado de Línea Blanca

La línea blanca es para el retail en los dos primeros meses del año la categoría más importante, impulsando las ventas de productos como lavadoras y cocinas, que crecieron versus el mismo periodo del año anterior en 5% y 3% respectivamente.

La campaña de verano es la tercera más importante para la venta de los artefactos de línea blanca, después del día de la madre y de Navidad. Es así como las ventas de línea blanca en retail del país en dicha campaña (del 28 de diciembre del 2014 al 21 de febrero del 2015) crecieron en valor en 1% respecto a la del año pasado, según se dio a conocer en la presentación de "Resultados de la campaña de línea blanca, pequeños electrodomésticos 2015", realizado por GfK Consumer Choices.

Para línea blanca, las ventas en la campaña de verano representan la quinta parte de lo realizado en el año, y en el caso de refrigeración por el incremento de la temperatura representa el 25% de su venta anual"

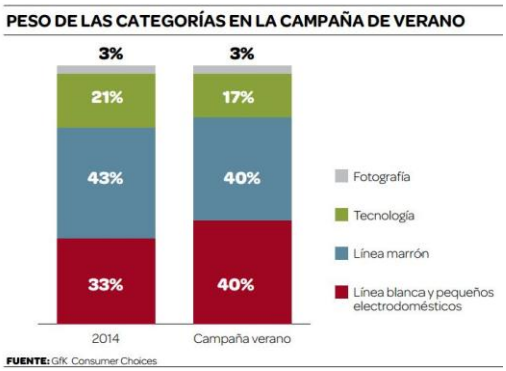


Figura 1: Peso de las categorías en la campaña de verano - 2015

Debido a este crecimiento, las consultoras de estudios de mercado vienen recomendando a las empresas participantes del mercado de artefactos eléctricos del país, a seguir invirtiendo en innovación y oferta, teniendo en cuenta que aún hay espacio para crecer.

PESO DE VENTA DE LÍNEA BLANCA

Principales ciudades

Región	Enero 2015
Lima	48.9%
La Libertad	6.7%
Arequipa	6.3%
Piura	5.8%
Lambayeque	4.9%
Ica	4.3%
Cusco	2.8%
Áncash	2.8%
Junín	2.8%
Online	1.9%
Cajamarca	1.8%
Loreto	1.7%
Puno	1.6%
San Martín	1.4%
Huánuco	1.3%
	95.1%

FUENTE: GFK CONSUMER CHOICES

PESO DE VENTA PEQUEÑOS ELECTRODOMÉSTICOS

Principales ciudades

Región	Enero 2015
Lima	59%
La Libertad	6%
Arequipa	5.4%
Piura	4.8%
Lambayeque	3.8%
Ica	3%
Junín	2.4%
Áncash	2.4%
Online	2.2%
Cusco	2.2%
Cajamarca	1.5%
Loreto	1.3%
Ucayali	1.2%
Huánuco	1.1%
Puno	0.9%
	97.3%

FUENTE: GFK CONSUMER CHOICES

Figura 2: Peso de venta de línea blanca en provincias

Según los estudios realizados por GfK Consumer Choices., la línea blanca y pequeños electrodomésticos representaron el 33% de las ventas de artefactos eléctricos en el 2014, detallando que en la campaña de verano del mismo año, se vendieron alrededor de 200,000 unidades de línea blanca en el retail.

2. EL PROBLEMA DE INVESTIGACIÓN

2.1. Planteamiento del problema

En el periodo de crecimiento y posicionamiento de la empresa en las diversas cadenas de retail a lo largo de todo el Perú, la empresa experimenta en el 2012 una baja en la calidad de sus productos, lo que se ve reflejado en un aumento gradual en los porcentajes de productos en devolución por fallas.
Dicho panorama origina una desaceleración en su participación de mercado a 30%, en relación al 35% experimentado en el año 2012 y por segundo año consecutivo no se generarían utilidades, reportando una pérdida de S/. 3 millones de soles y cuya tendencia negativa se mantuvieron para el 2013.
Tras ello el directorio decide realizar una reestructuración de las jefaturas entre las cuales se decide cambiar al Gerente General, ocupando el puesto *"un profesional con un interesante perfil, con experiencia en la industria y poseedor de perspectivas de empresas internacionales",* para que dirija los rumbos de la empresa.
Con dicha reestructuración, la empresa comenzó a estabilizarse, pero los números manifestaban que aún no era suficiente ya que la participación de mercado no creció significativamente y las tasas de productos en devolución mantenían su tendencia de incremento de forma significativa.

De acuerdo a los datos proporcionados por el área de Gestión de la Calidad, de todas las fallas comunicadas por el cliente final, el 95% se centra en productos como: cocinas, refrigeradores y campanas extractoras, de los cuales los dos primeros son fabricados por la casa matriz Ecuador y la filial de Perú (Manufacturera C & R SAC.) y el tercero es importado desde China.
La tasa de fallas acumulada en el periodo 2012 - 2014, se registró con un promedio (cocinas y refrigeradoras) del 15% y con una marcada tendencia al crecimiento. Lo cual evidencia que la forma actual de análisis e identificación de causas no venía siendo eficiente.

Las fallas son prácticamente inevitables debido al movimiento de los productos y al traslado desde la fábrica hacia los centros de distribución y de medio a medio de trasporte. Del mismo modo es vital para la industria la especificación y determinación clara de las responsabilidades por cada ocurrencia de falla.
Actualmente el área de Gestión de la calidad recurre a las consultas manuales para obtener los datos necesarios de los productos en devolución, realizando un análisis estadístico de la base de datos "Parque en Garantía", que registran los casos atendidos por el área de postventa. Dicho análisis da como resultado la tendencia de los porcentajes de tipo de fallas por producto.

El análisis evidencia varias limitantes como:

- Alto tiempo para el procesamiento de datos
- No permite establecer relaciones claras entre más de tres variables
- Poca confiabilidad para poder predecir fallas en el futuro.

En este panorama de una constante tendencia a la alza de los porcentajes de las fallas y productos en devolución, la línea de producción de refrigeradoras es la que se ve más afectada, debido a ser la que presenta los porcentajes más altos registrados durante el periodo 2012 - 2014.

Esta situación lleva al directorio a cerrar por completo dicha línea de producción a inicios del 2015 y optando por solo realizar importaciones de refrigeradoras desde la casa matriz.

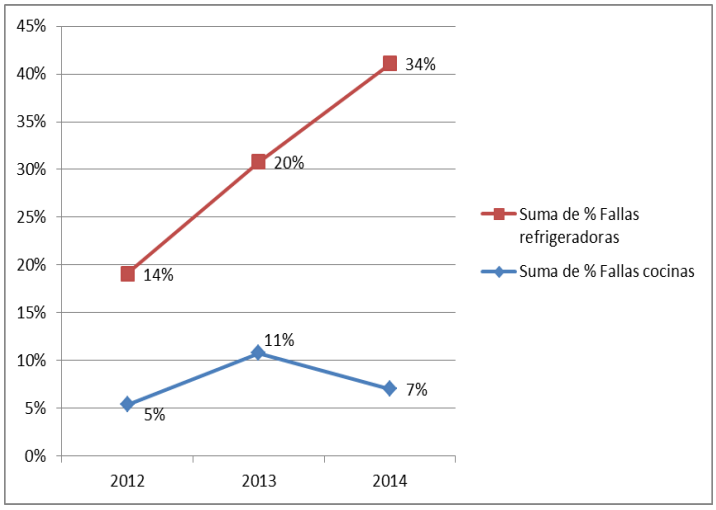


Figura 3: Porcentajes de fallas por producto

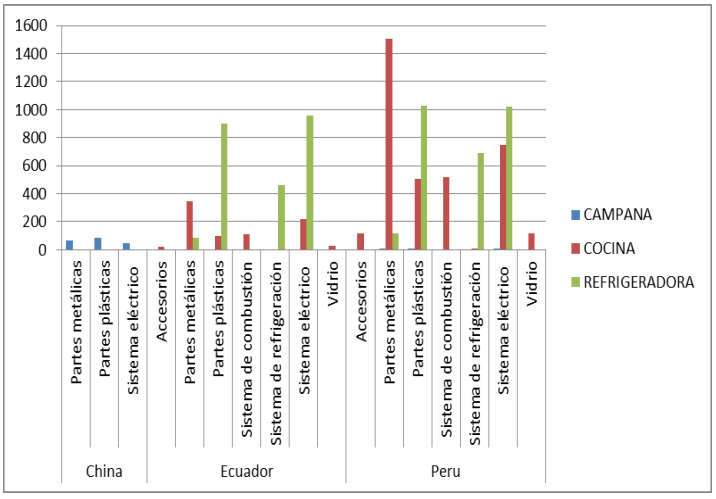


Figura 4: Tipo de fallas por producto

2.2. Formulación del problema

La minería de datos debe ser utilizada en esta actividad, debido a que no existen registros documentados de comportamientos estándares, que dan origen a fallas en la industria de fabricación de productos de línea blanca (cocinas y refrigeradoras). En la actualidad todo análisis de fallas y reparación de unidades se realiza en forma reactiva.

El problema a resolver será por lo tanto diseñar un modelo predictivo haciendo uso de la minería de datos, para la búsqueda de patrones de fallas, los cuales identifiquen variables que influyan en el origen de las fallas tanto estéticas como funcionales en el futuro.

Con el diseño del modelo predictivo y el uso de una herramienta sofisticada, permitirá a la compañía identificar patrones de falla con un alto grado de confianza, para apoyar al área de gestión de la calidad y a la coordinación del Sistema de Gestión de la calidad en el fortalecimiento del proceso de mejora continua.

Al finalizar la investigación, se espera conocer los patrones que relacionen las variables más significativas tales como: Modelo, tiempo de almacenamiento, procedencia, antigüedad, tipo de falla, etc.

2.3. Objetivos de la investigación

2.3.1. Objetivo general

Diseñar un modelo predictivo que permita aplicar la minería de datos, en el descubrimiento de patrones de fallas, que se originan a lo largo de todo el circuito tanto de producción como logístico, en productos como cocinas y refrigeradores.

2.3.2. Objetivos específicos

- Identificar variables no usadas en la actualidad y evaluar su inclusión en el modelo predictivo de fallas.

- Direccionar y plantear acciones de mejora en las diversas áreas involucradas con el caso estudiado.

- Proporcionar una herramienta sofistica de análisis de datos a la empresa, a fin de fortalecer y agilizar la toma de decisiones.

2.4. Justificación del estudio

- La investigación se encuentra justificada, debido al interés de la empresa en trabajar con herramientas sofisticas que aporten a la toma de decisiones de forma oportuna, brindando una ventaja competitiva.

- Adicional esta investigación sienta un precedente importante para futuras investigaciones en el campo de producción, logística y marketing.

2.5. Alcance del estudio

La presente investigación comprende sólo el estudio de la base de datos proporcionada por postventa "Parque en garantía", evaluando variables y herramientas de modelación predictiva. No se discute el impacto económico que pueda tener en los resultados de la empresa.

La investigación no incluye los registros de los productos devueltos a través del canal de logística y ejecutivo de ventas, por representar un porcentaje mínimo, almacenado de forma manual y sin la información suficiente para su análisis.

3. MARCO TEÓRICO

3.1. ¿Qué es la tasa de falla?

Actualmente la tasa de falla en la compañía se calcula tomando en cuenta los registros de los reclamos canalizados y atendidos por el área de postventa.

$$\text{Tasa de falla} = \frac{\text{\# de fallas registradas en un periodo determinado}}{\text{\# Total de productos fabricados en dicho periodo}} \times 100$$

En este cálculo la compañía por el momento no considera los registros de reclamos canalizados por el área de logística, gestión de calidad y de forma directa por el ejecutivo de venta. Dichos registros, se evalúan por separado y de forma diferente.

El analizarlos de forma diferente lo que origina es dilatar los tiempos de respuesta y limitar en gran parte el cálculo que evidencie la magnitud real del problema, pero a la vez significa una buena oportunidad para poder analizar y plantear propuestas de mejora en el proceso de canalización y análisis de la información de los reclamos.

3.2. Registro del área de postventa

Los datos registrados por el área de postventa, son reclamos de fallas en los productos comunicados de forma directa por el cliente a través del Call Center o comunicados a través de la página web de la compañía, los mismos que son almacenados en una base de datos.

El personal del Call Center registra y a la vez distribuye las solicitudes en: Garantía (productos que se encuentran dentro del tiempo de garantía) y solicitud de servicio (productos que se encuentran fuera del tiempo de garantía). Luego dichos registros son unificados de acuerdo al canal de comunicación y compartidos a través del software a la sección de técnicos.

Posterior el coordinador de servicio técnico, es quien se encarga de asignar las solicitudes de atención, de acuerdo a rutas ya establecidas.

Una vez que las solicitudes son asignadas, los técnicos realizan las visitas respectivas a las casas de los clientes, en donde realizan el diagnóstico y se corrobora información detallada en la garantía, registrando todo en la tarjeta de daño. Luego de la visita, el técnico traspasa dicha información al módulo de postventa en el SAP.

Luego de tener toda la información en el sistema, el coordinador de servicio técnico importa la base de datos y realiza una limpieza de cada uno de los datos ingresados, antes de compartirla con el jefe de Gestión de la Calidad.

3.3. Estadística del área de Gestión de la Calidad

El jefe de Gestión de la Calidad recibe la base de datos de los reclamos en formato Excel, en donde realiza una revisión y limpieza de los datos según sea el caso, para luego elaborar las estadísticas respectivas que evidencien el estado actual de los reclamos.

Dentro de la estadística trabajada, se encuentran los porcentajes de fallas por tipo de producto de acuerdo al parque en garantía (Todos los productos aún en fecha de garantía).

Dicha estadística de fallas son calculadas en un periodo de tiempo trimestral, las mismas que son expuestas en una reunión coordinada también por el área de gestión de la calidad y en donde se discuten y direccionan acciones en base al análisis realizado.

A la reunión se convoca a un representante del área de Logística, Línea de producción de cocinas, Línea de producción de refrigeradoras y postventa.

Todas las acciones son registradas por la coordinadora del sistema de gestión, quién es la encargada de realizar el seguimiento de acuerdo a las fechas de cumplimiento propuestas por los responsables de las áreas antes mencionadas.

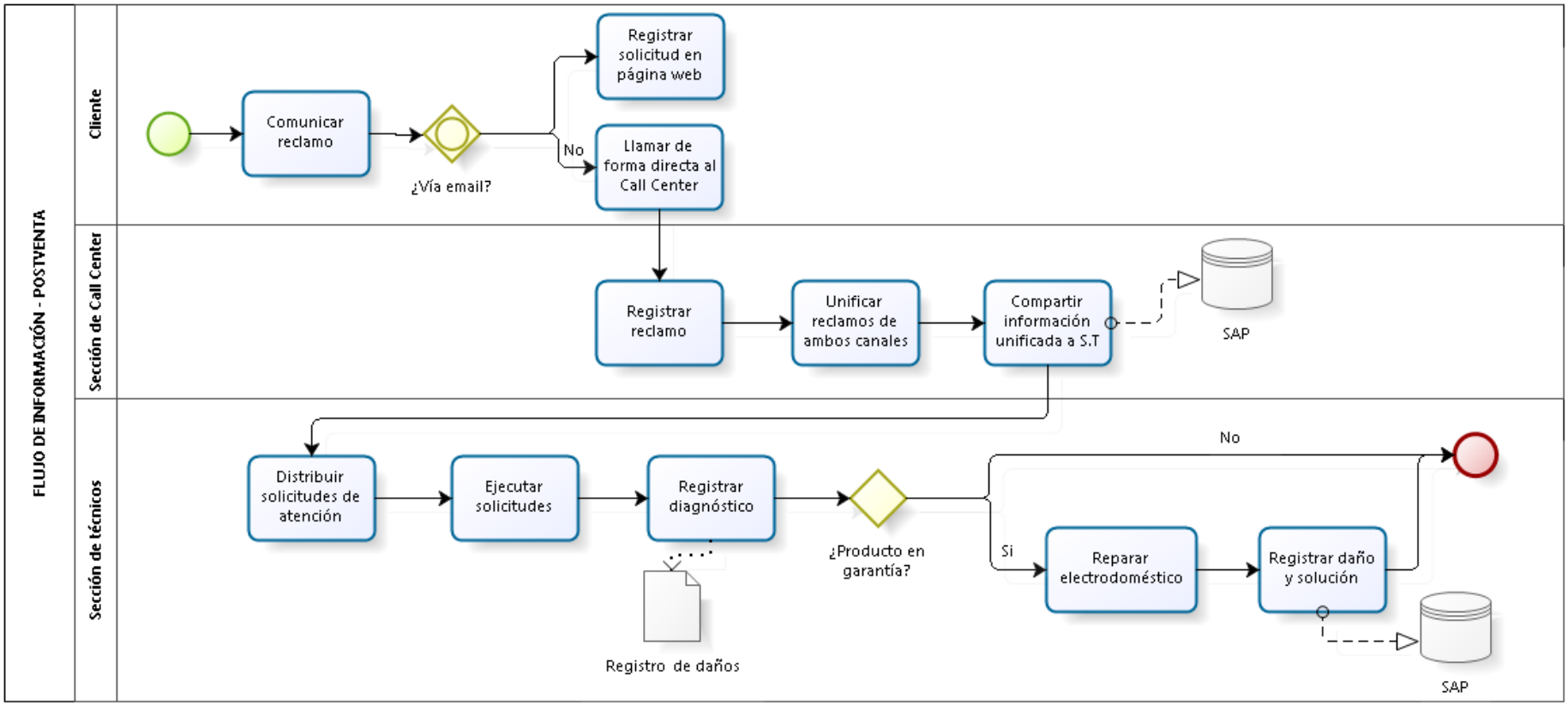


Figura 5: Modelo BPMN Proceso de Gestión de Reclamos

4. METODOLOGÍA

4.1. Descripción del trabajo

En este capítulo se da a conocer el método y herramienta propuestos para aplicar Minería de datos al proceso de devolución de cocinas y refrigeradores, con el objetivo de diseñar un modelo predictivo para la búsqueda de patrones de fallas, identificando las variables más relevantes que dan origen a algún tipo de falla tanto funcional como estética, a partir de la base de datos de los registros de fallas comunicados por los clientes a través del Call Center.

A partir de los resultados, se definirán las recomendaciones inmediatas para las áreas de la empresa que se ven directamente relacionadas con los patrones de fallas.

4.2. Minería de datos

Conceptualmente la Minería de Datos, se puede definir como un conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos con motivo de predecir de forma automatizada tendencias, comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos.
Desde un punto de vista empresarial la minería de datos puede ser definida como un conjunto de áreas que tiene como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisiones.
Lo que realmente hace el data mining es reunir las ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, el Datawarehouse y el Procesamiento Masivo, principalmente usando como materia prima bases de datos.
Para tener una aproximación cercana a las diversas definiciones encontradas se puede concluir que la minería de datos es un proceso con el cual se pueden descubrir y cuantificar relaciones predictivas en los datos, y del resultado de este proceso es posible obtener conocimiento útil para el negocio.

Hoy en día realizando las consultas (simplemente navegando en los datos) convencionales a grandes bases de datos no es suficiente para resolver problemas de negocios, sino que se hace necesario seguir una metodología ordenada para aplicar herramientas tecnológicas y técnicas disponibles en informáticas para así obtener conocimiento y resultados confiables que permitan a las compañías obtener un beneficio.

4.3. Metodología CRISP-DM

Para la explotación de datos aplicando minería de datos existen diferentes técnicas las cuales pueden ser desarrolladas según diferentes metodologías. Para el desarrollo de este trabajo de investigación se utiliza la metodología CRISP-DM (Cross Industry Standard Process for Data), definido en el año 1997, por una agrupación de empresas europeas del sector TIC, como una metodología y guía de buenas prácticas para la gestión de los datos de cara a su procesamiento y análisis.

CRISP-DM es un modelo de proceso de minería de datos mas utilizada, tal como demuestran las encuestas realizadas a lo largo del tiempo por diferentes organismos. Por ejemplo, la encuesta "What main methodology are you using for your analytics, data mining, or data science projects? Realizada por KDnuggets, refleja la amplia aceptación que tiene el modelo en el Mercado.

En el siguiente gráfico se muestran los resultados obtenidos tras la realización de 200 encuestas en 2014 frente a los resultados de 2007.

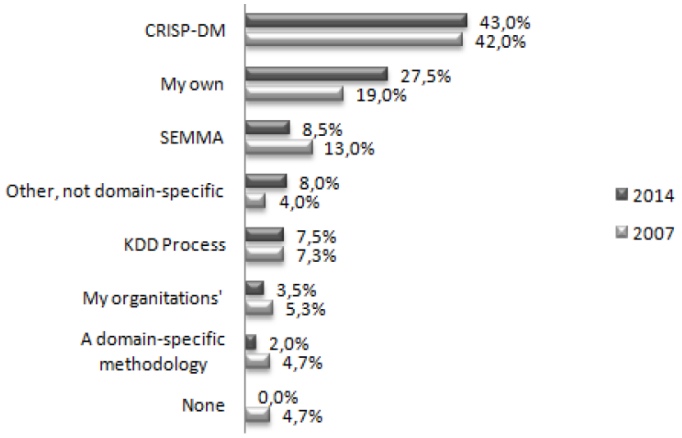


Figura 6: Estadística de modelos de Data Mining

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Data Mining en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto (Figura 7). Las flechas indican relaciones más habituales entre las fases, aunque se pueden establecer relaciones entre cualquier fase. El círculo exterior en el diagrama simboliza la naturaleza cíclica de la minería de datos en sí, ya que un proceso de minería de datos continúa después del despliegue de una solución porque las lecciones aprendidas durante el proceso pueden provocar nuevas preguntas de negocio, a menudo más centradas.

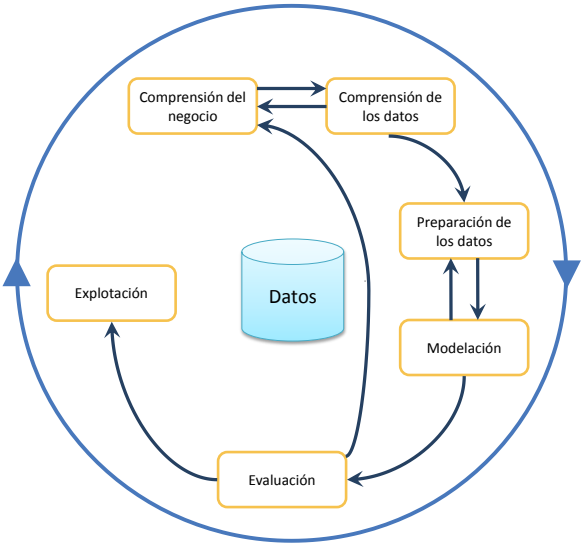


Figura 7: Metodología CRISP - DM

A continuación se describe brevemente cada una de las fases:

Comprensión del negocio (*Business Understanding*): Esta fase inicial se enfoca en la comprensión de los objetivos del proyecto y exigencias desde una perspectiva de negocio, para convertir el conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Comprensión de los datos (*Data Understanding*): La fase de comprensión de datos comienza con la recolección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar problemas de calidad de datos (a través del análisis de comportamiento histórico, tendencias, etc.) y detectando variables o comportamientos interesantes que permitan formular hipótesis.

Preparación de los datos (*Data Preparation*): Engloba todas las actividades necesarias para construir el conjunto de datos final, que será usado en la fase de modelado a partir de los datos iniciales. Estas tareas de preparación de datos van a ser ejecutadas repetidas veces y no pueden realizarse en cualquier orden. En general incluyen la selección y transformación de tablas, registros y atributos así como la transformación y la limpieza de datos para las herramientas que modelan. En esta etapa es muy importante considerar la opinión de los expertos.

Modelado (*Modeling*): En esta etapa se seleccionan varias técnicas de modelado para aplicarlas a los datos disponibles y sus parámetros son calibrados para obtener resultados óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos que pueden ser aplicadas.
Algunas técnicas tienen requerimientos específicos sobre la forma de los datos, por lo que es necesario volver a la fase de preparación, de ahí el carácter cíclico del proceso CRISP-DM.

Evaluación (Evaluation): En esta etapa en el proyecto, se ha construido un modelo (o modelos) que parece tener una alta calidad de la perspectiva de análisis de datos. Antes de proceder al despliegue final del modelo es importante evaluar a fondo el modelo y la revisión de los pasos ejecutados para crearlo para comparar el modelo correctamente obtenido con los objetivos del negocio. Un objetivo clave es determinar si hay algún aspecto importante del negocio que no ha sido suficientemente considerado para modificar los modelos creados de minería de datos.

Explotación: La creación del modelo no es el final del proyecto ya que el conocimiento obtenido tendrá que ser organizado y presentado de forma que los usuarios internos de la empresa puedan utilizar el conocimiento adquirido. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos.

4.4. Método estadístico a emplear

En minería de datos existen múltiples paradigmas, de los cuales emplearemos Árbol de decisiones y Clúster, que se detallan a continuación:

4.4.1. Árbol de decisión

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, son diagramas lógicos que representan una serie de condiciones o reglas de decisión sucesivas. Un árbol de decisión particiona el espacio formado por las variables predictoras en un conjunto de hiper-rectángulos, en cada hiper-rectángulo se ajusta un modelo sencillo, generalmente una constante.

Pueden utilizarse valores discretos o continuos:

- Valores discretos = Clasificación
- Valores continuos = Regresión

El modelo de árbol decisión en este trabajo se utiliza para clasificar los productos, de acuerdo a sus tipos de fallas, tomando en cuenta las variables independientes registradas en la base de datos por el área de postventa de la compañía.

Componentes de un Árbol

Nodo:

- Representa un atributo de entrada
- Un nodo interno contiene un test sobre algún valor de una de las variables.

Arco:

- Representan los diferentes valores que pueden tomar los atributos.

Hoja:

- Son los valores de salida de la función.

Rama:

- Entregan los posibles caminos que se tienen de acuerdo a la regla establecida.

✓ **Ventajas de los árboles de decisión**

- Los resultados son fáciles de entender e interpretar
- No tiene problema en trabajar con datos nulos
- Realiza automáticamente la selección de variables
- Es robusto ante la presencia de "outliers"
- Es un clasificador no-paramétrico, es decir, no requiere suposiciones
- Toma en cuenta las interpretaciones que pueden existir entre las variables predictoras.

✓ **Desventajas de los árboles de decisión**

- El proceso de selección de variables es sesgado hacia las variables con más valores diferentes.
- Dificultad para elegir el árbol óptimo
- Requiere un gran número de datos para asegurar que la cantidad de observaciones en los nodos terminales sea significativo.

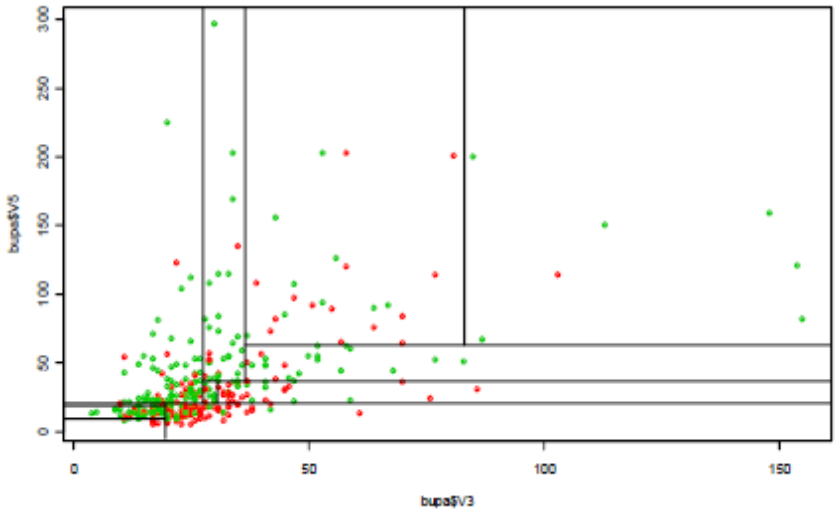


Figura 8: Partición del espacio muestral

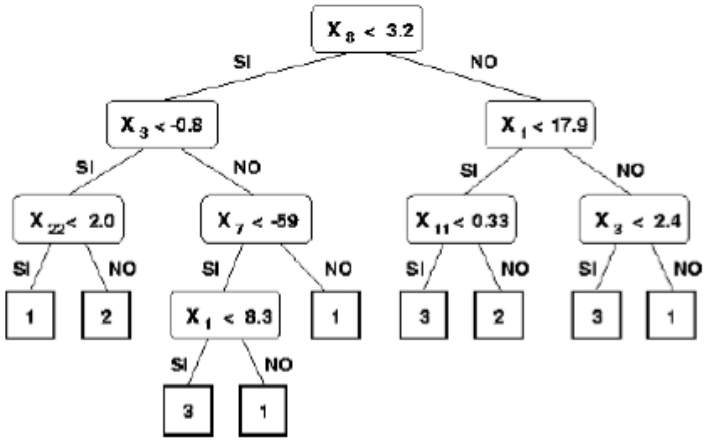


Figura 9: Ejemplo de un árbol de decisión

4.4.2. Clustering

Técnica que permite encontrar grupos en los cuales los objetos de un grupo sean similares entre sí y diferentes de los objetos de los otros grupos.

La formación de grupos es muy utilizada como paso analítico en la minería de datos, ya que permite analizar las variables en cada uno de los grupos. Una vez obtenidos los grupos, si se desea predecir, lo que se realiza es determinar a qué grupo pertenece con mayor probabilidad la característica que buscamos, y en función del grupo, se realiza una determinada acción sobre el mismo.

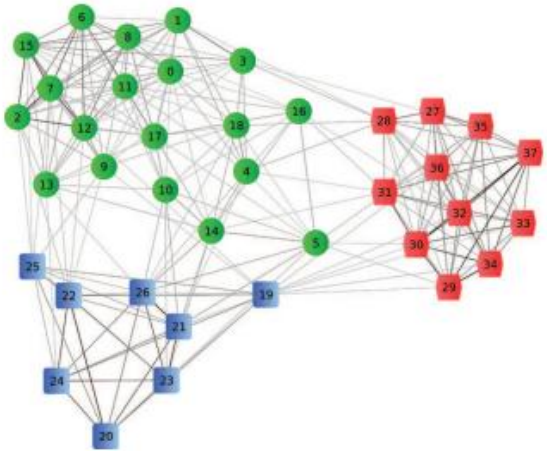


Figura 10: Formación de grupos - Cluster

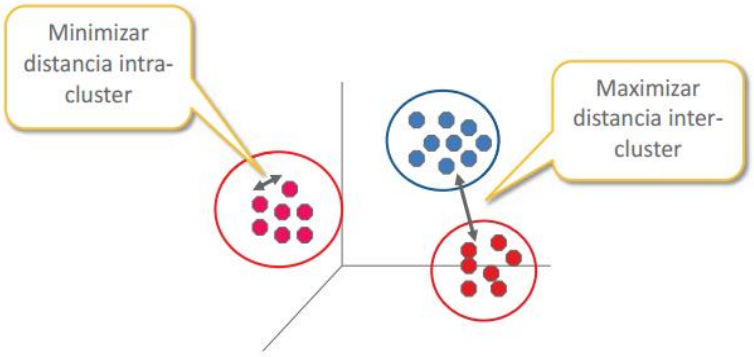


Figura 11: Distancias entre Cluster

Un buen Clustering, debe ser: Escalable, tratar distintos tipos de variables, descubrimiento de clusters con formas arbitrarias, capaz de tratar datos con ruido y outliers, insensible al orden de los registros y generar resultados interpretables.

El modelo de agrupamiento desarrollado en esta actividad de graduación, se basa en la utilización del algoritmo de K-means, el cual, para su funcionamiento requiere parámetros previamente especificados. Se define un orden en cuanto a la selección de los parámetros, aunque este no interfiere en el buen funcionamiento del algoritmo.

Para aplicar el algoritmo se requiere el número de clusters (k) como dato de entrada, cada cluster tiene asociado un centroide (centro geométrico del cluster). Los puntos se asignan al cluster cuyo centroide esté más cerca (utilizando alguna métrica de distancia), luego iterativamente, se van actualizando los centroides en función de las asignaciones de puntos a clusters, hasta que los centroides dejen de cambiar.

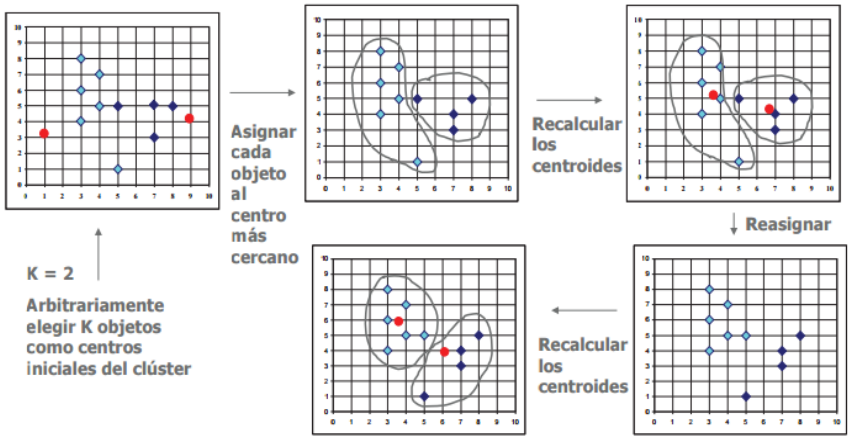


Figura 12: Desplazamiento de centroides – Cluster

4.4.3. Análisis y descripción de la muestra

Se realizó el análisis de los datos correspondientes a 28 064 registros de fallas identificadas por el cliente final y comunicadas de forma directa al Call Center del área de postventa de la empresa. El tamaño de la muestra después de la etapa de limpieza se reduce a 9.797 registros de fallas entre cocinas y refrigeradoras.

4.5. Herramienta utilizada

Para el desarrollo de este trabajo de investigación, se aplica una herramienta de uso libre que permite la generación de flujos de trabajo de forma ágil.
La herramienta utilizada, es llamada Rapid Miner, que es una de las principales herramientas utilizadas para el diseño de modelos predictivos, en proyectos reales, tal como se evidencia en una de las encuestas realizadas por KDnuggets, sitio líder de Business Analytics, Data mining y Data Science, sobre el uso de herramientas en proyectos reales, haciendo una comparación entre los años 2013, 2014 y 2015, "What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real Project?", en el que se identifican las principales herramientas usadas en el sector.

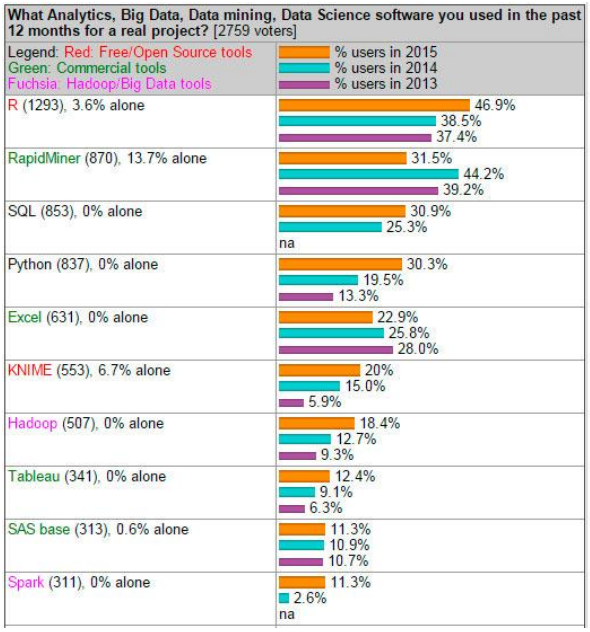


Figura 13: Herramientas utilizadas en los años 2013, 2014 y 2015

En la ilustración anterior aparecen las herramientas más usadas en los últimos 3 años y se muestra la comparación de uso entre ellas. Las herramientas de analítica que son consideradas Open-Source están señaladas en rojo, las comerciales están en verde y las genéricas de color negro.

4.5.1. Descripción de la Herramienta

La herramienta a utilizar para el desarrollo del trabajo de investigación es RapidMiner, Utilizaremos la versión "RapidMiner Studio Basic" la cual es gratuita.

RapidMiner (anteriormente YALE, Yet Another Learning Environment) era la plataforma de minería de datos de código abierto más usada (con más de 3 millones de descargas), antes de convertirse en un producto comercial.

Ofrece un entorno integrado para el aprendizaje automático, minería de datos, minería de texto, análisis predictivos y análisis de negocio, incorpora extracción, transformación, carga de datos, y presentación de informes de predicción. La interfaz de usuario y las herramientas de visualización gráficas son excelentes, con una inteligencia considerable integrada en el proceso de construcción de flujos de trabajo. Esto proporciona el reconocimiento de errores y sugerencias de soluciones rápidas. Su capacidad de transformación de datos es única entre las herramientas de esta naturaleza y permite que los resultados sean inspeccionados al momento.

Se utiliza para aplicaciones industriales y de negocio, así como para la investigación, educación, creación rápida de prototipos y el desarrollo de aplicaciones. Es compatible con todos los pasos del proceso de minería de datos, incluyendo la visualización, la validación y optimización de los resultados.

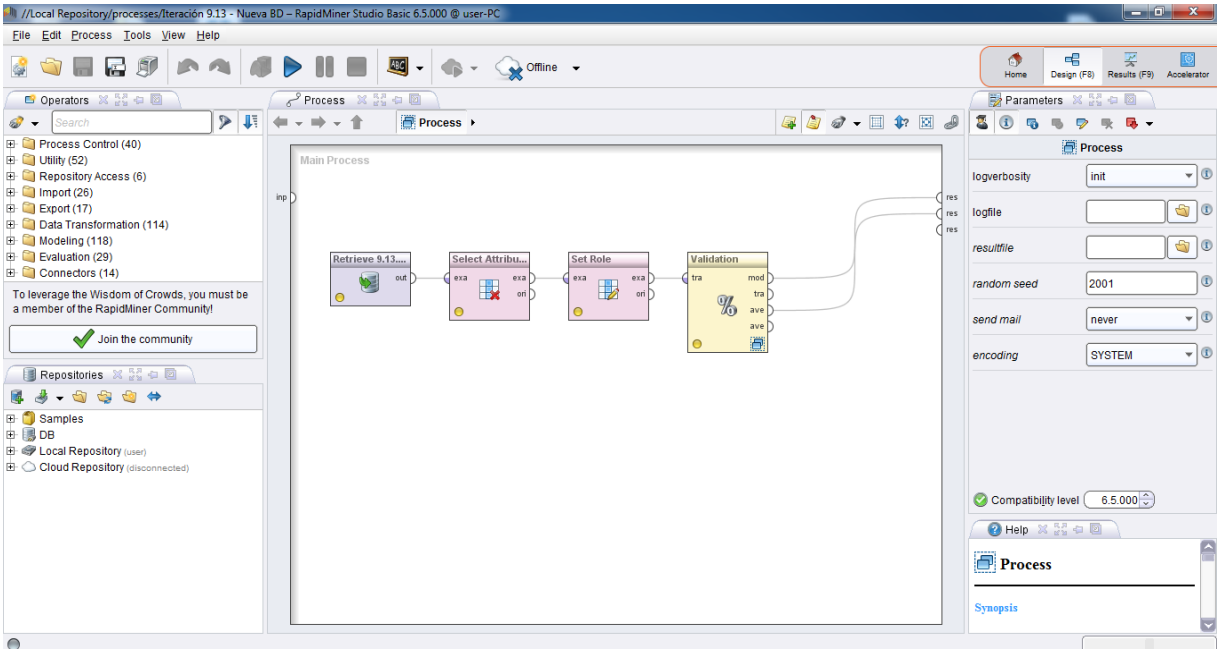


Figura 14: Área de modelamiento en RapidMiner

5. DISEÑO Y RESULTADOS

En esta etapa se describe la propuesta del modelo predictivo de análisis de datos, los resultados que se obtienen y se evalúa el grado en que el modelo encuentra los objetivos del negocio haciendo uso de la minería de datos.

5.1. Diseño del Modelo Predictivo

El objetivo que se persigue, es identificar patrones de fallas en productos terminados de la empresa Manufacturera C&R SAC.

Para llevar a cabo el objetivo planteado, lo primero que se necesita es tener una idea clara sobre la base de datos que se está utilizando y los objetivos que persigue la empresa.

Por lo tanto en esta actividad de graduación se propone un modelo de análisis de datos que busca ejecutar de forma consecutiva dos técnicas de análisis predictivos, llamados: Decision Tree y Clustering. La primera técnica diseñada con elementos que permiten identificar patrones de fallas relacionando variables tanto de producción como de logística.

Una vez identificados aquellos patrones más potentes, se integra la segunda técnica de análisis llamada Clustering, en el que se va a utilizar la información aportada por el árbol de decisión, con el objetivo de observar cómo se agrupan de forma individual las variables presentes en los patrones de fallas. Esta técnica es utilizada para complementar la información proporcionada a través de los patrones de fallas.

El modelo propuesto, busca que la empresa cuente con información específica de forma gráfica acerca de las variables que podrían causar fallas en los productos terminados en el futuro. Dicha información obtenida en un tiempo mínimo, con el objetivo de agilizar la toma de decisiones.

Por otro lado el modelo predictivo, se encuentran alineados con los objetivos estratégicos, comerciales y operacionales de la empresa Manufacturera C&R SAC.

Para llevar a cabo el diseño de cada uno de los modelos, fue necesario realizar un análisis minucioso de la base de datos, en donde se toma en cuenta eliminar aquellos registros y variables que agregan no agregan valor al objetivo planteado.

5.1.1. Selección de los datos

Se va a trabajar con la base de datos proporcionados por el área de postventa de la compañía. Esta base de datos almacena información sobre todos aquellas fallas en los productos terminados, comunicados por el cliente final a través del Call center, entre los años 2013 - 2015.

Los datos se encuentran almacenados en el módulo del postventa del sistema SAP.

A continuación, se muestra un fragmento de la base de datos importada en formato Excel:

ID	TIPO DE PRODUCTO	MODELO	COLOR	MARCA	CRITICIDAD DE FALLA	DESCRIPCIÓN DE FALLA	UBICACIÓN DEL DAÑO	TIPO DE FALLA	CLIENTE	VENTA EN RETAIL	N° SERIE	MES DE PRODUCCIÓN	AÑO DE PRODUCCIÓN	FECHA DE COMPRA
1	COCINA	Florencia Qua	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	BENITES GOIC	NO	P254614110143	4	2013	04-07-2013
2	COCINA	Florencia Qua	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	IMPORTACION	SI	P254608530153	5	2013	19-06-2013
3	COCINA	Florencia Qua	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	GONZALES SA	NO	P254616740143	4	2013	08-07-2013
4	COCINA	Florencia Qua	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	ELECTROTIENE	NO	P254606200133	3	2013	19-06-2013
5	COCINA	Florencia Qua	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	ROSAS CARM	NO	P254601580133	3	2013	30-06-2013
6	COCINA	Galicia Quarzo	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	CERRON ALVA	NO	P265900390143	4	2013	21-06-2013
7	COCINA	Galicia Quarzo	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	ROSAS DE PEÑ	NO	P265902330163	6	2013	20-07-2013
8	COCINA	Granada Quar	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	SANGUINETTI	NO	P265413840143	4	2013	07-07-2013
9	COCINA	Granada Quar	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	LUPACA QUISI	NO	P265401460143	4	2013	09-05-2013
10	COCINA	Milan Quarzo	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	VILCHEZ GONZ	NO	P265504680163	6	2013	27-06-2013
11	COCINA	Montecarlo Q	Croma	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	HERRERA CESI	NO	P250502700143	4	2013	24-06-2013
12	COCINA	Niza Quarzo	Negro	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	MARTINEZ ALV	NO	278000090113	1	2013	19-07-2013
13	COCINA	Palermo Quar	Negro	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	ECHEGARAY H	NO	268101200123	2	2013	29-05-2013
14	REFRIGERADO	RI-530 Avant	Blanco	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	PIZANGO YAH	NO	122601400113	1	2013	10-03-2013
15	REFRIGERADO	RI-530 Avant	Blanco	INDURAMA	Leve	Accesorios da	Partes plástica	Estético	ESPINOZA GUI	NO	122600830133	3	2013	10-06-2013

En esta etapa, se evidenció lo siguiente: Filas con más de 2 celdas vacías, números de series de los productos incompletos, fechas de compra menor a fechas de producción, fechas de identificación de falla menor a fechas de compra, nombre de tipo de producto en singular y en plural. Una de las categorías de daños, se describe como “Actividades Generales” (Categoría que fue eliminada de la base de datos, debido a que no proporciona información específica de daños en los productos)

Luego de la etapa de limpieza de datos, se consideró eliminar algunas variables que no contenían información relevante para el objetivo del modelo, dichas variables son:

- Nombre del técnico
- Nombre del cliente
- Estado del producto (Garantía – Sin garantía)

Así mismo para aumentar la precisión del modelo fue necesario diseñar variables con ayuda de las ya registradas en la base de datos, dichas variables se detallan en el cuadro N°2.

Tabla 1: Variables de la Base de datos – Postventa

N°	Nombre Variable	Descripción
1	Fecha de compra	-
2	Fecha de identificación de falla	-
3	Año de producción	-
4	Mes de producción	-
5	Descripción de falla	-
6	Modelo del producto	-
7	**N° de serie del producto**	-
8	**Nombre del cliente**	-
9	**Nombre del técnico**	
10	Estado del producto	Garantía o sin Garantía
11	Color del producto	Blanco, croma, gris, negro, steel
12	Origen del producto	Perú, Ecuador, China
13	Marca del producto	Indurama o Global
14	Criticidad de falla	Grave o Leve
15	Tipo de falla	Funcional o estética
16	Tipo de producto	Cocina o Refrigeradora

Tabla 2: Diseño de nuevas variables

N°	Nombre de Variable	Descripción
19	Años de antigüedad del producto	Fecha de Falla – Fecha Producción
20	Tiempo de almacenamiento	Fecha de compra – Fecha de producción
21	Tiempo de uso	Fecha de falla – Fecha de compra
22	Ubicación del daño	Se agruparon las fallas en siete categorías: Accesorios, partes plásticas, partes metálicas, sistema de combustión, sistema de refrigeración, sistema eléctrico y daños en vidrio.

Nota: La variables "Tiempo de almacenamiento" y "Tiempo de uso", se estructuraron en tres niveles: Bajo (0-3 meses), Medio (4-8 meses) y Alto (Mayor a 9 meses)

5.1.2. Descripción de resultados – Árbol de Decisión

Luego del diseño y ejecución del modelo predictivo de fallas, los resultados que el software nos muestra se detallan en la matriz de confusión y el árbol de decisión propiamente dicho, los cuales se describen a continuación:

Matriz de confusión

Tabla 3: Matriz de confusión

accuracy: 87.41%

	true COCINA	true REFRIGERADORA	true CAMPANA	class precision
pred. COCINA	1044	109	0	90.55%
pred. REFRIGERADORA	260	1464	1	84.87%
pred. CAMPANA	0	0	61	100.00%
class recall	80.06%	93.07%	98.39%	

En los valores detallados en la matriz de confusión, podemos observar el criterio Accuracy, que define el grado de precisión del modelo, que en este caso se obtuvo un valor de 87,41 %, es decir que el modelo es capaz de acertar con las reglas de diseño expuestas anteriormente, en un 87.41 % de los casos. En dicha matriz se observan valores horizontales que son los que el modelo diseñado predice y valores verticales, que es lo que realmente ocurrió.

Nota: De acuerdo a los resultados y al número reducido de registros de fallas en campanas, se omite dicho producto, concentrando el análisis sólo en cocinas y refrigeradores.

Los valores de la matriz se leen de la siguiente forma:

Lo que el modelo predice (Valores Horizontales)
Fallas en cocinas:

- Que 1044 fallas que se predicen en cocinas, efectivamente se presentaron en cocinas.
- Que 109 fallas de las que se predicen en cocinas, en realidad se presentaron en refrigeradoras.
- Que 0 fallas de las que se deberían presentar en cocinas, 0 se presentaron en campanas.

Con lo cual se concluye que del total de las fallas que el modelo predice en cocinas, un 90.55% de ellas realmente se presentan en cocinas.

Fallas en refrigeradores:

- Que 260 fallas que se predicen en refrigeradores, en realidad se presentan en cocinas.
- Que 1464 fallas que se predicen en refrigeradores efectivamente se presentan en refrigeradoras.
- Que 1 de los registros de fallas que se predicen en refrigeradores, en realidad se presentan en campanas.

Con lo que se concluye que del total de las fallas que el modelo predice en refrigeradores, un 84,87 % de ellos realmente se presentan en refrigeradores.

Lo que realmente ocurrió (Valores Verticales)

Fallas en cocinas:

- Del total de registros de fallas que se presentan en cocinas, el 80.06% quedó correctamente clasificado.

Fallas en refrigeradores:

- Del total de registros de fallas que se presentan en refrigeradores, el 93.07% quedó correctamente clasificado.

Árbol de decisión

El árbol de decisión que el modelo presenta, se muestra a continuación:

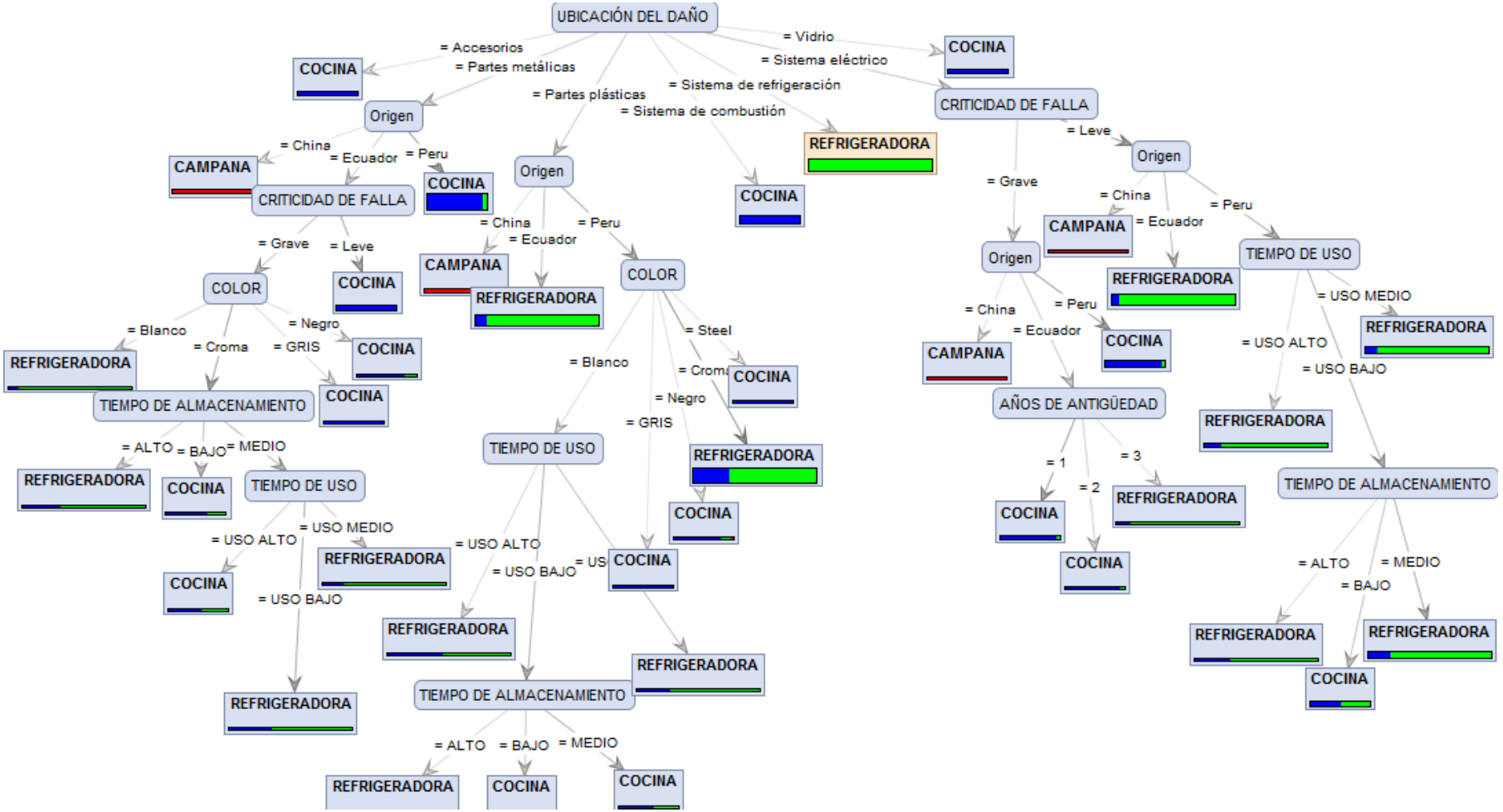
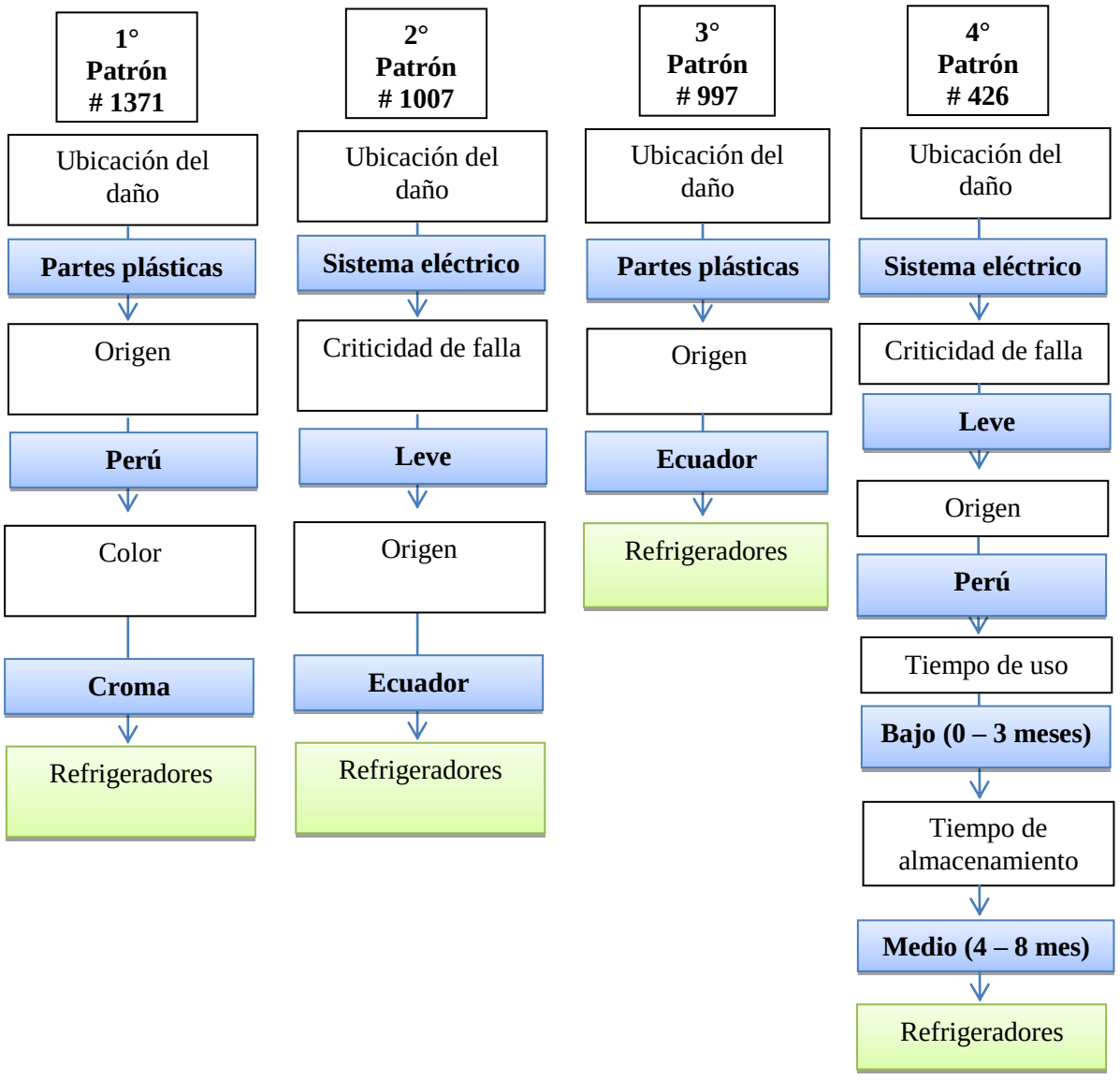


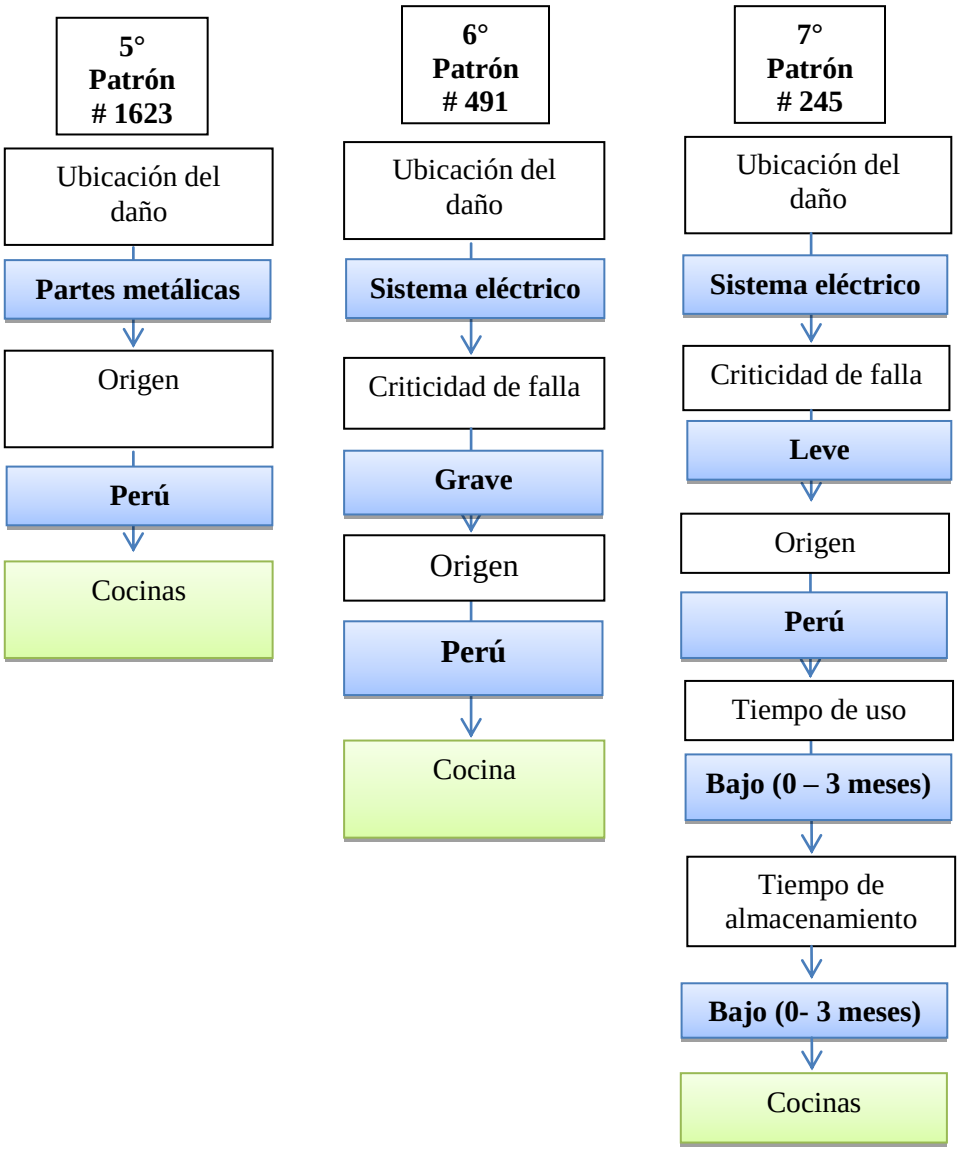
Figura 15: Esquema del árbol de decisión

A continuación se muestran los patrones identificados en el árbol de decisión y que de acuerdo al número de ocurrencias han sido considerados como los más potentes.

Patrones identificados en Refrigerados:



Patrones identificados en Cocinas:



Lectura de patrones:

1° Patrón: Lectura de abajo hacia arriba
Existe una alta probabilidad que refrigeradores, fabricadas en Perú, presenten fallas ubicadas en los accesorios plásticos.

2° Patrón:
Existe una alta probabilidad que refrigeradores, fabricadas en Ecuador, presenten leves, ubicadas en el sistema eléctrico.

3° Patrón:
Existe una alta probabilidad que refrigeradores, fabricadas en Ecuador, presenten fallas ubicadas en los accesorios plásticos.

4° Patrón:
Existe una alta probabilidad que refrigeradores, con un tiempo de almacenamiento entre 4 y 8 meses, fabricadas en Perú, presenten fallas leves, ubicadas en el sistema eléctrico, las mismas que el cliente identifica en un tiempo de entre 0 y 3 meses de uso.

5° Patrón:
Existe una alta probabilidad que cocinas fabricadas en Perú, presenten ubicadas en las partes metálicas.

6° Patrón:
Existe una alta probabilidad que cocinas, fabricadas en Perú, presenten fallas graves, ubicadas en el sistema eléctrico.

7° Patrón:
Existe una alta probabilidad que cocinas, con un tiempo de almacenamiento entre 0 y 3 meses, fabricadas en Perú, presenten fallas leves, ubicadas en el sistema eléctrico, las mismas que el cliente identifica en un tiempo de entre 0 y 3 meses de uso.

5.2. Diseño del proceso “CLUSTERING”

El modelamiento haciendo uso de la técnica de “Clustering”, en este trabajo de actividad de graduación, se considera como un complemento al árbol de decisión, con el objetivo de identificar información adicional que permita ver como los datos se comportan en conjunto y en particular como se agrupan, permitiendo entregar información específica a la empresa para la toma de decisiones.
La modelación se realizó por separado tanto cocinas como refrigeradores.

5.2.1. Resultados del modelo “Clustering” - Refrigeradores

Se identificaron dos tipos de Cluster: Cluster 0 con 3557 items y Cluster 1 con 1700 items.

Cluster Model

```
Cluster 0: 3557 items
Cluster 1: 1700 items
Total number of items: 5257
```

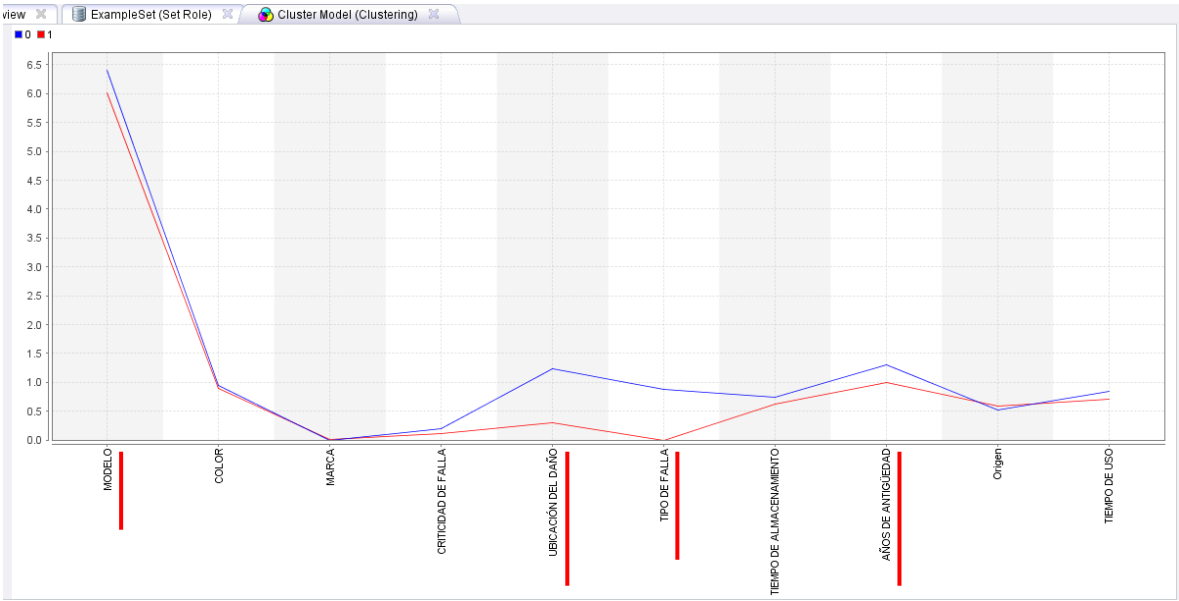
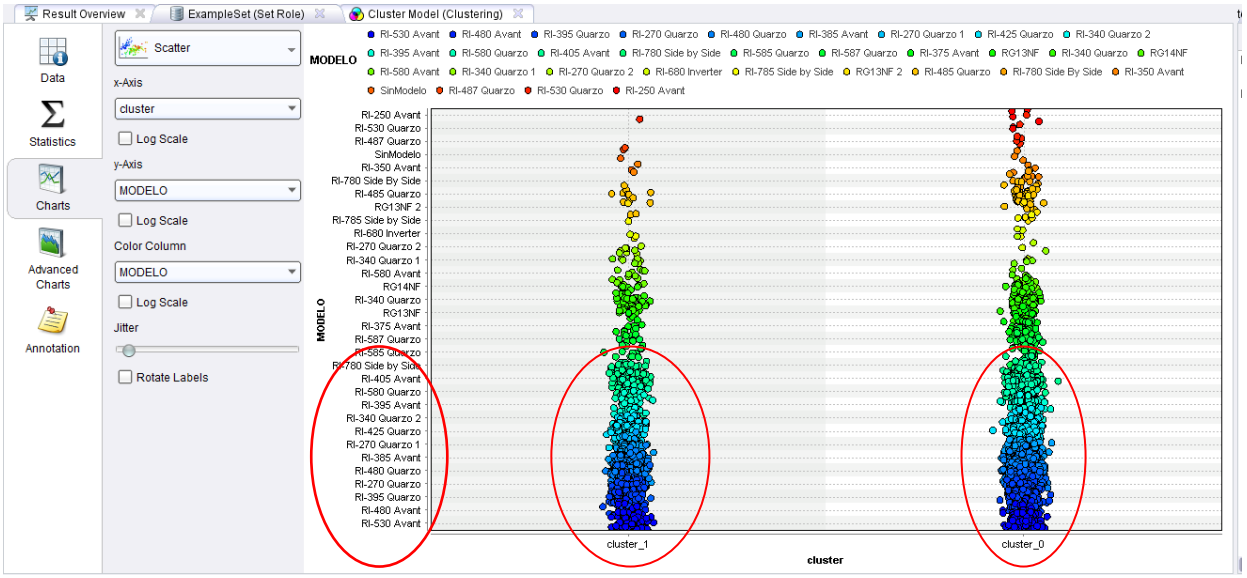


Figura 16: Valores de los centroides en cada variable - Refrigeradores

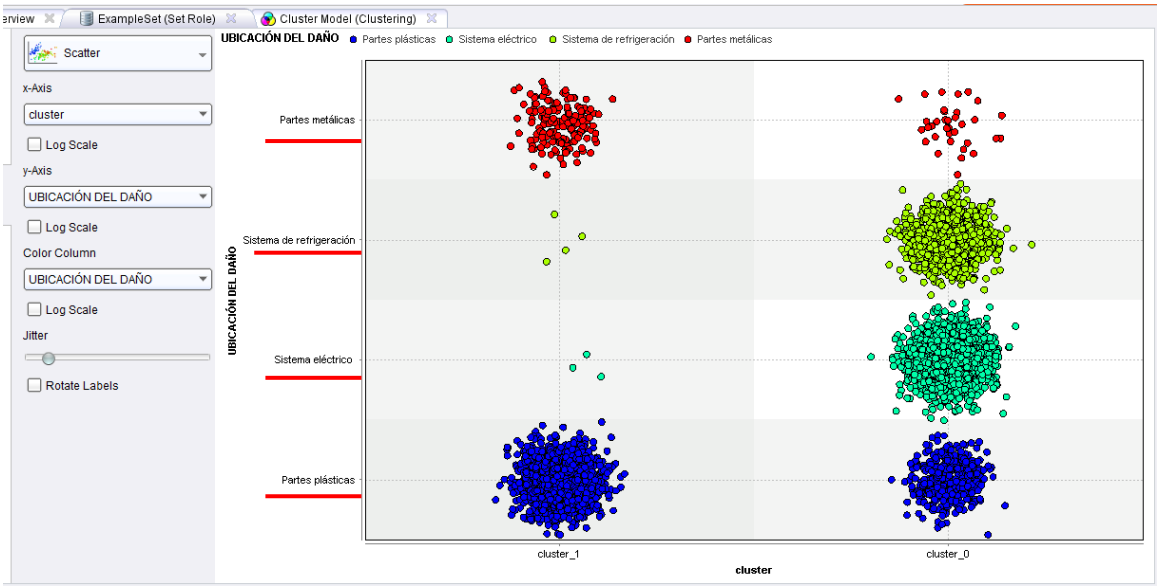
En la figura 14, se observan los valores que tienen los centroides de cada variable y en cada cluster. Se aprecia que los centroides de las variables: “Ubicación del año”, “Tipo de falla” y “Años de antigüedad”, están más distantes entre ellos que el resto. Lo cual quiere decir que dichas variables aportan a la clasificación de los cluster.

A continuación se verá de forma gráfica los valores de dichos atributos y su clasificación:

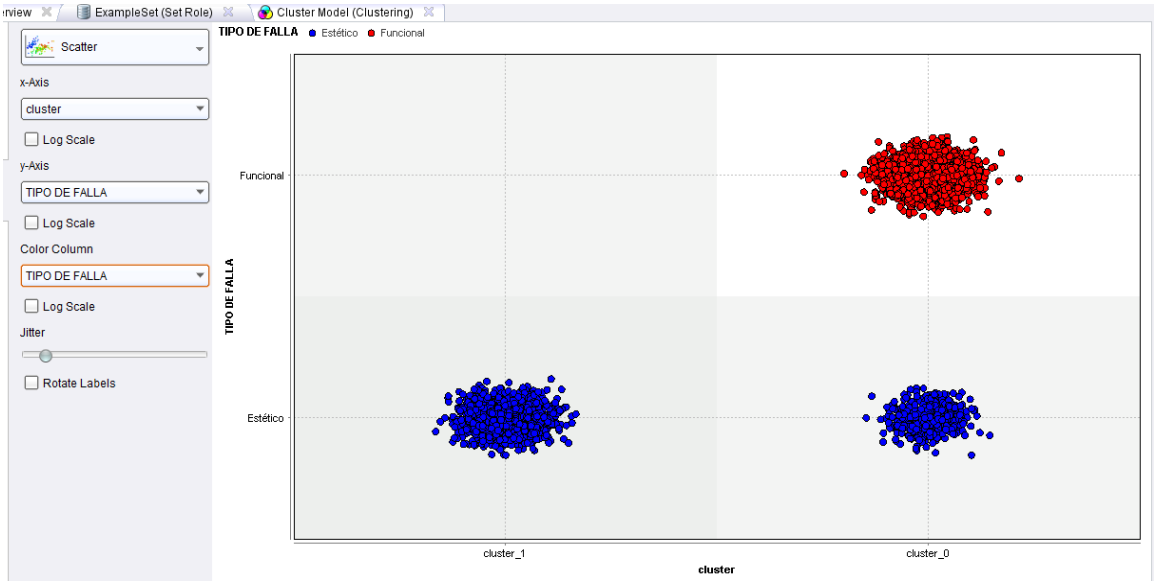
- **Modelo:** Seleccionando cluster por modelo, se observa que la concentración se encuentra en los siguientes modelos: RI-530 Avant, RI-480 Avant, RI-395 Quarzo, RI-270 Quarzo, RI-480, Quarzo, RI-385 Avant, RI-270 Quarzo, RI-425 Quarzo, RI-340 Quarzo, RI-395 Avant, RI-580 Quarzo y RI-405 Avant.



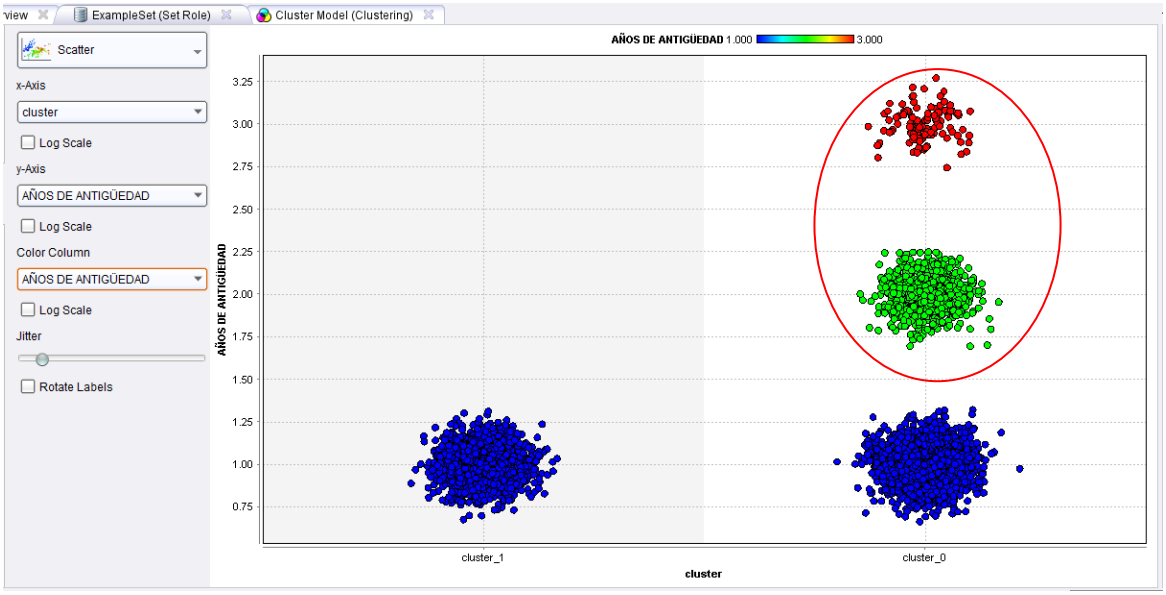
- **Ubicación del daño:** Seleccionando cluster por ubicación del daño, se observa que los grupos se concentración en: Sistema eléctrico, partes plásticas, sistema de refrigeración y partes metálicas.



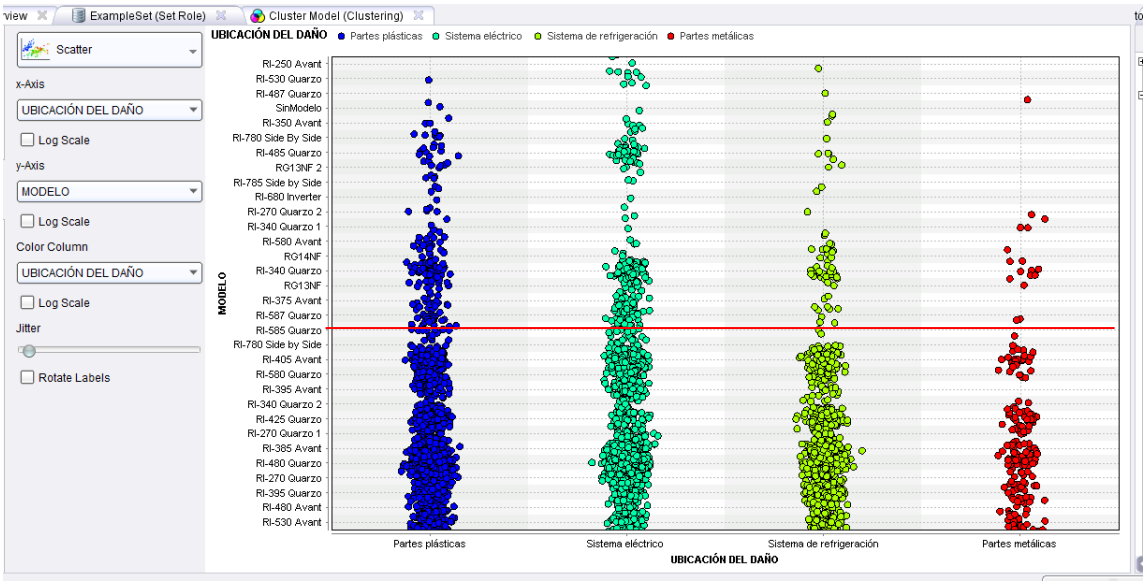
- **Tipo de falla:** Seleccionando cluster por tipo de falla, se observa cluster muy bien definidos tanto para tipo de falla estético como funcional.



- **Años de antigüedad:** Seleccionando cluster por años de antigüedad, se observa cluster muy bien definidos para 2 y 3 años de antigüedad.

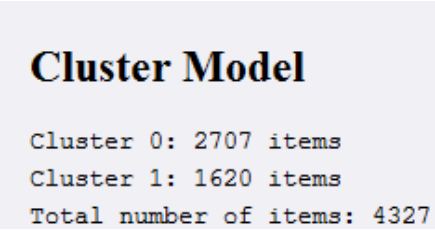


- En el siguiente gráfico se muestra la distribución por “Ubicación del daño” con respecto al modelo. Observándose que la mayor concentración de fallas, se encuentra en “Partes plásticas” y “Sistema eléctrico”, específicamente en los modelos: RI-530 Avant, RI-480 Avant, RI-395 Quarzo, RI-270 Quarzo, RI-480, Quarzo, RI-385 Avant, RI-270 Quarzo, RI-425 Quarzo, RI-340 Quarzo, RI-395 Avant, RI-580 Quarzo y RI-405 Avant.



5.2.2. “Clustering Cocinas”

Se identificaron dos tipos de Cluster: Cluster 0 con 2707 items y Cluster 1 con 1620 items.



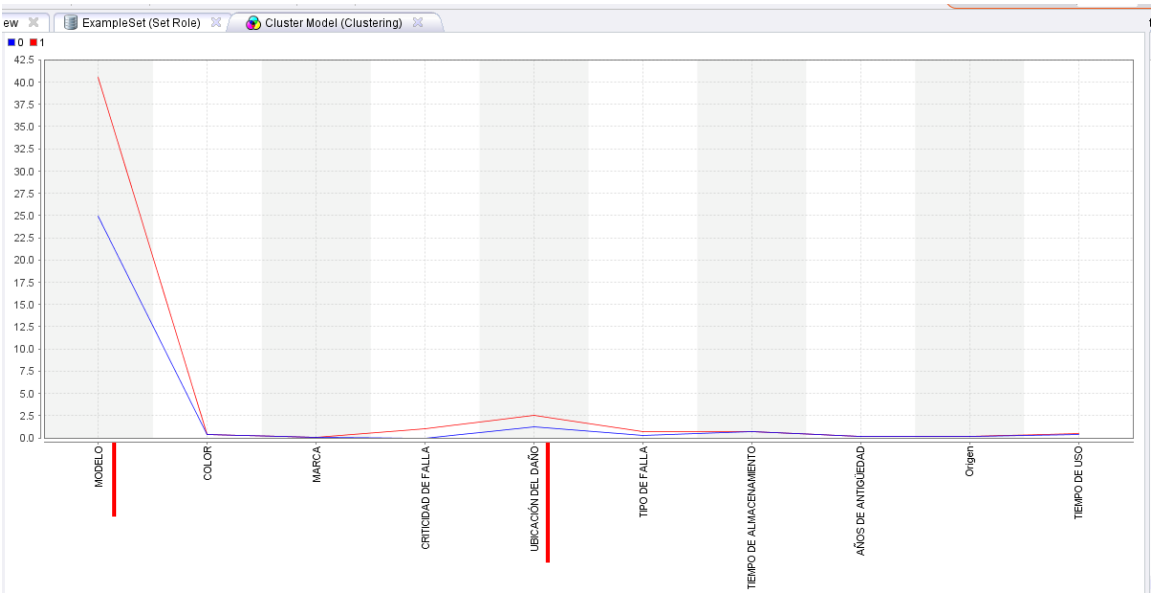
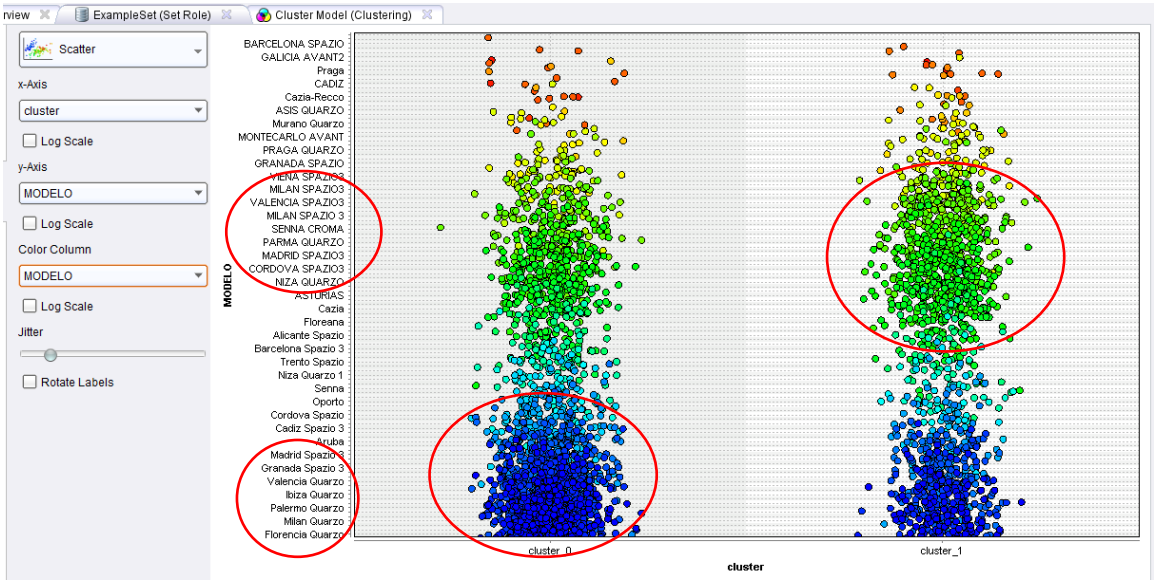


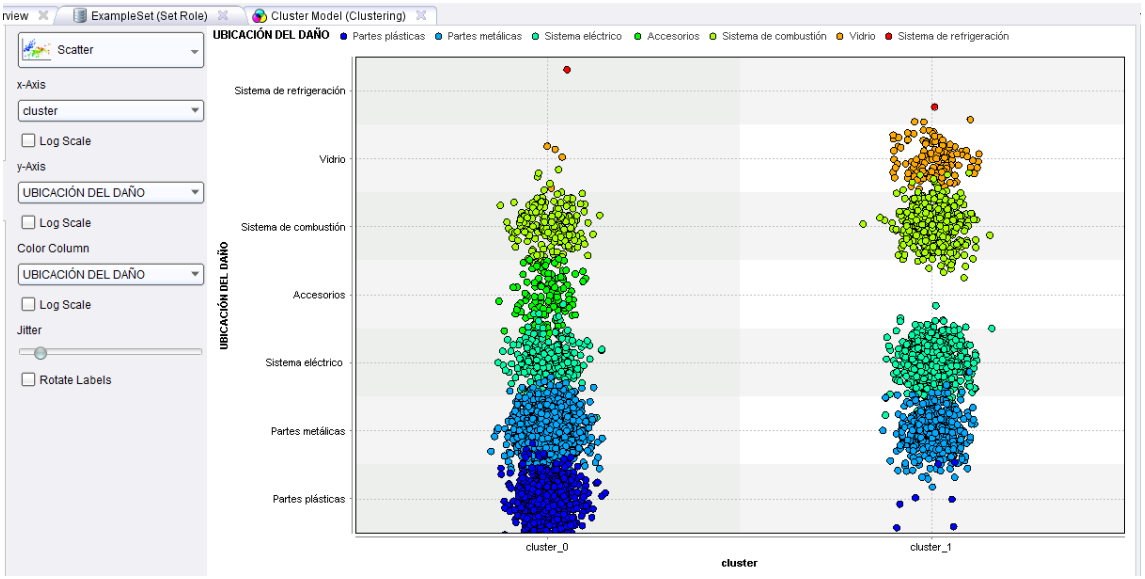
Figura 17: Valores de los centroides en cada variable - Cocinas

En la figura 15, se observan los valores que tienen los centroides de cada variable y en cada cluster. En este caso para cocinas se ve que los centroides de las variables: **"Modelo"** y **"Ubicación del daño",** están más distantes entre ellos que el resto. Lo cual quiere decir que dichas variables, aportan a la clasificación de los cluster.
A continuación se verá de forma gráfica los valores de dichas variables y su clasificación:

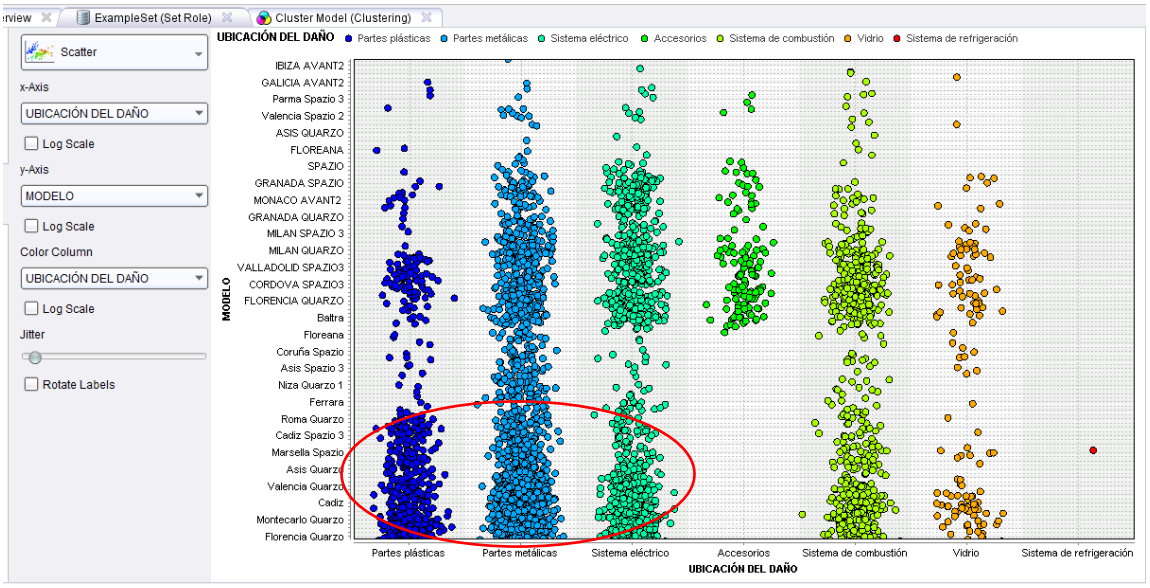
- **Modelo:** Seleccionando cluster por modelo, se observa que la concentración se encuentra en los siguientes modelos: Florencia Quarzo, Milán Quarzo, Palermo Quarzo, Ibiza Quarzo, Valencia Quarzo, Granada Spazio3, Madrid Spazio, Asturias, Niza Quarzo, Cordova Spazio, Parma Quarzo, Senna Quarzo y Milán Spazio.



- **Ubicación del daño:** Seleccionando cluster por ubicación del daño, se observa que la concentración se encuentra en partes plásticas, partes metálicas y sistema eléctrico.



- En el siguiente gráfico a lo igual que para refrigeradores, la concentración de fallas, se concentra en: “Partes plásticas” y “Sistema eléctrico”, con la única diferencia que se suman a la concentración, fallas en “Partes metálicas”.
  En relación al modelo, se observa que la distribución es bastante homogénea.



6. CONCLUSIONES

Durante el desarrollo del trabajo, desde el inicio hasta la evaluación de los resultados, se logró detallar diversas conclusiones, siendo las principales, las que se muestran a continuación:

a) Árbol de decisión

- El modelo predictivo diseñado en este trabajo, tiene una alta capacidad para definir patrones, representado con una confianza de 87.41%, es decir que el modelo es capaz de acertar con las reglas de diseño expuestas, en un 87.41 % de los casos.

- Se concluye que el origen de fallas de los productos de la empresa Manufacturera C&R SAC. se divide en sistemáticas y aleatorias.
  Sistemáticas: Reflejadas por patrones en el Sistema eléctrico, partes plásticas y partes metálicas.
  Aleatorias: Reflejadas por concentraciones que no relacionan más de 1 variables, como fallas en: Sistema de refrigeración y sistema de combustión.
  Lo que quiere decir que las causas no siguen un patrón, sino por el contrario se encuentran dispersas, lo cual se convierte en algo crítico ya que evidencia un panorama de causas aleatorias, en los procesos de la empresa.

- La recolección de los datos por el área de postventa y modelados a través del Rapid Miner, provee a los gestores de producción, logística y calidad, de una información adicional y enriquecida sobre las variables que podrían causar, fallas tanto estéticas como funcionales en los próximos lotes de producción.

- De todas las variables estudiadas, las que tienen una mayor relación entre ellas son:
  Tiempo de almacenamiento, criticidad de falla, origen, color y tiempo de uso.

Patrones de Refrigeradoras

- Se logró identificar patrones para predecir el origen de fallas ubicadas en el sistema eléctrico y accesorios plásticos.

- Las fallas ubicadas en los accesorios plásticos, se pueden originar tanto en productos fabricados en Perú como en Ecuador, específicamente con mayor porcentaje en productos de color Croma fabricados en Perú.

- Las fallas ubicadas en el sistema eléctrico, se pueden originar tanto en productos fabricados en líneas de producción de Perú como de Ecuador.
- Se concluye que el proceso de fabricación e inspección de refrigeradoras existen problemas críticos, debido a que las fallas en el sistema eléctrico se presentan en un tiempo mínimo de uso entre 0 y 3 meses.

Patrones de Cocinas

- Se logró identificar patrones para predecir el origen de fallas sólo en el sistema eléctrico y partes metálicas.

- Las fallas ubicadas en las partes metálicas, se pueden originar en mayor porcentaje en productos fabricados en la planta de producción de Perú.

- Las fallas ubicadas en el sistema eléctrico, se pueden originar en mayor porcentaje en productos fabricados en Perú.

- Se concluye que el proceso de fabricación e inspección de refrigeradoras existen problemas críticos, debido a que las fallas en el sistema eléctrico se presentan en un tiempo mínimo de uso (entre 0 y 3 meses).

- Se concluye que el tiempo de almacenamiento no es determinante para originar fallas en los productos, debido a que el producto permanece almacenado un tiempo mínimo entre 0 y 3 meses.

b) Clustering

- Tanto en cocinas como en refrigeradoras las variables comunes que permiten identificar "Cluster" son, modelo y ubicación del daño. Con lo cual se pudo distinguir cuales son los modelos específicos en donde existe mayor probabilidad de ocurrencia de fallas.

- Se observa una mejor distribución de grupos que relacionan la ubicación de fallas con respecto a los diferentes modelos, en cocinas.

- Los grupos observados a través de la técnica de "Clustering", tiene una alta relación con los patrones identificados a través de la técnica de "Decisión Tree".

- El modelo identificó dos cluster, tanto para cocinas como para refrigeradoras.

7. RECOMENDACIONES Y TRABAJOS FUTUROS

7.1. RECOMENDACIONES

Conociendo ahora los patrones de fallas que en el futuro podrían dar origen a fallas tanto estéticas como funcionales, se procede a detallar algunas recomendaciones de mejora en los siguientes áreas de la empresa Manufacturera C & R SAC.

Procesos:

- Se recomienda realizar una inspección mucho más minuciosa al final de las líneas de producción, específicamente en aquellas partes que de acuerdo a los patrones, son más susceptibles de presentar fallas en el futuro. Tales como: Partes plásticas, metálicas y sistema eléctrico.
  Una de las herramientas a utilizar, podría ser un Check-list dirigido, incluyendo dentro de los ítem a inspeccionar todas aquellas variables detallas en los patrones del modelo predictivo, con el objetivo de optimizar el tiempo de inspección de los productos terminados.

- Revisar los controles de proceso en las secciones que ensamblan y fabrican: Partes plásticas, metálicas y sistema eléctrico.

- Evaluar la robustez de las pruebas funcionales de los materiales eléctricos, antes de su salida de la línea de producción.

- Implementar en las líneas de producción, un Check-list dirigido para la inspección de los materiales al momento de la recepción.

- Implementar cuadros informativos en las líneas de proceso, detallando las tendencias de las fallas en los productos y de los patrones de fallas identificados. Con el objetivo de hacer parte del proceso de mejora a cada uno de los operarios de la las líneas de producción.

Proveedores:

- Realizar una clasificación de proveedores de todos los materiales utilizados en la fabricación de productos de color Croma. Con el objetivo de direccionar acciones específicas tales como:
  - Disminuir el tiempo de las evaluaciones a proveedores críticos.
  - Auditorías in – situ en los procesos productivos.

Postventa:

- Estandarizar la forma de llenado de la base de datos de los daños registrados a través del Call Center y los técnicos. Con el objetivo de evitar diferentes formatos y descripciones para una misma falla en los productos.
- Realizar capacitaciones para concientizar la importancia del registro completo de las fallas comunicadas por el cliente.

7.2. TRABAJOS FUTUROS

- A partir del diseño del modelo predictivo y del clustering, realizado en este trabajo de investigación, se recomienda ampliar el estudio incluyendo la base de datos de las fallas identificadas en productos que se canalizan de forma directa a través del ejecutivo de venta y del área de logística, con el objetivo de poder evidenciar la magnitud real de los patrones que dan origen a las fallas en los productos terminados de la empresa.
  Adicional permitir descubrir nuevas variables que podrían ser la causa del origen de las fallas en los productos terminados.

- Realizar un análisis Lean en la ruta de los procesos que integran las tareas desde el Call center hasta el área de gestión de la calidad, ello con el objetivo de eliminar todas las tareas que no agregan valor y por consiguiente vienen consumiendo tiempo, retrasando el análisis y la toma de decisiones.

8. BIBLIOGRAFIA

Jiménez, S. P., Puldón, J. J., & Andrade, R. A. E. (2012). Modelo clustering para el análisis en la ejecución de procesos de negocio. *Investigación Operacional*, *33*(3), 210-222

Flores, H. (2009). *Detección de Patrones de Daños y Averías en la Industria Automotriz* (Doctoral dissertation, Tesis de Maestría en Ingeniería en Sistemas de Información. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional).

Aguilera Arranz, V. (2013). Estudio sobre las causas del abandono de los estudiantes de economía de la Universidad Oberta de Catalunya entre los años 1998 y 2008.

Núñez, d. L. Modelos predictivos del Churn–abandono de clientes–para operadores de telecomunicaciones.

.

Chamorro, A. C. (2013). *Método para aplicar Minería de Procesos a la Distribución de Bebestibles no Alcohólicos*, Tesis para optar al grado de Magíster en Ciencias de la Ingeniería. Escuela de Ingeniería. Pontificia Universidad Católica de Chile.

KDnuggets. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Recuperado del sitio web http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-datascience-projects.html.

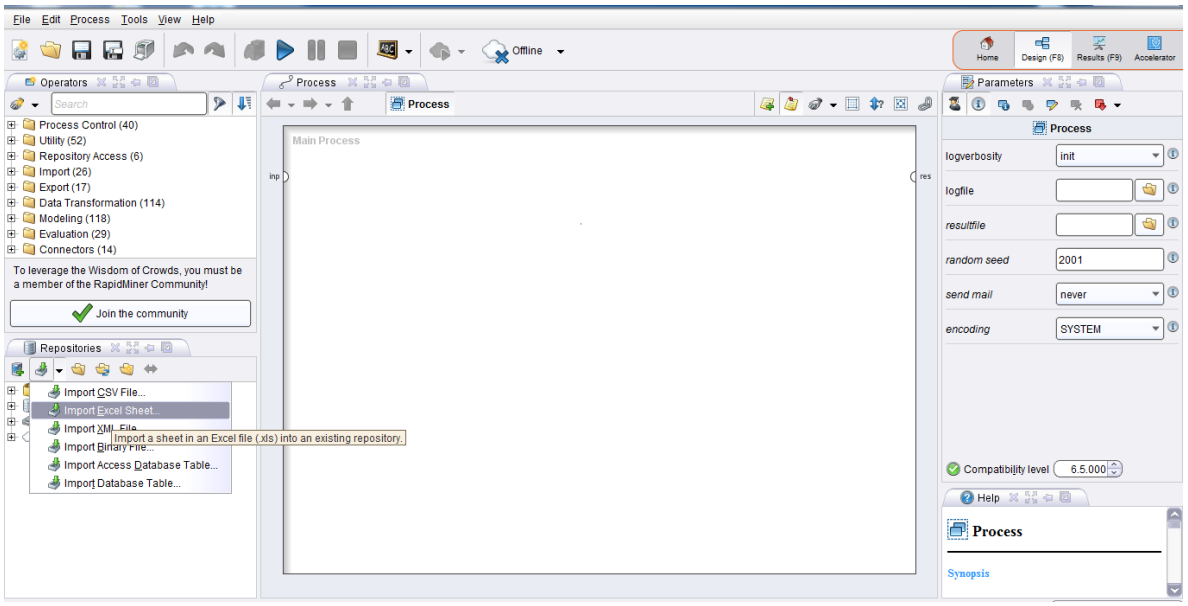
KDnuggets. "What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real Project?. Recuperado del sitio web http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html

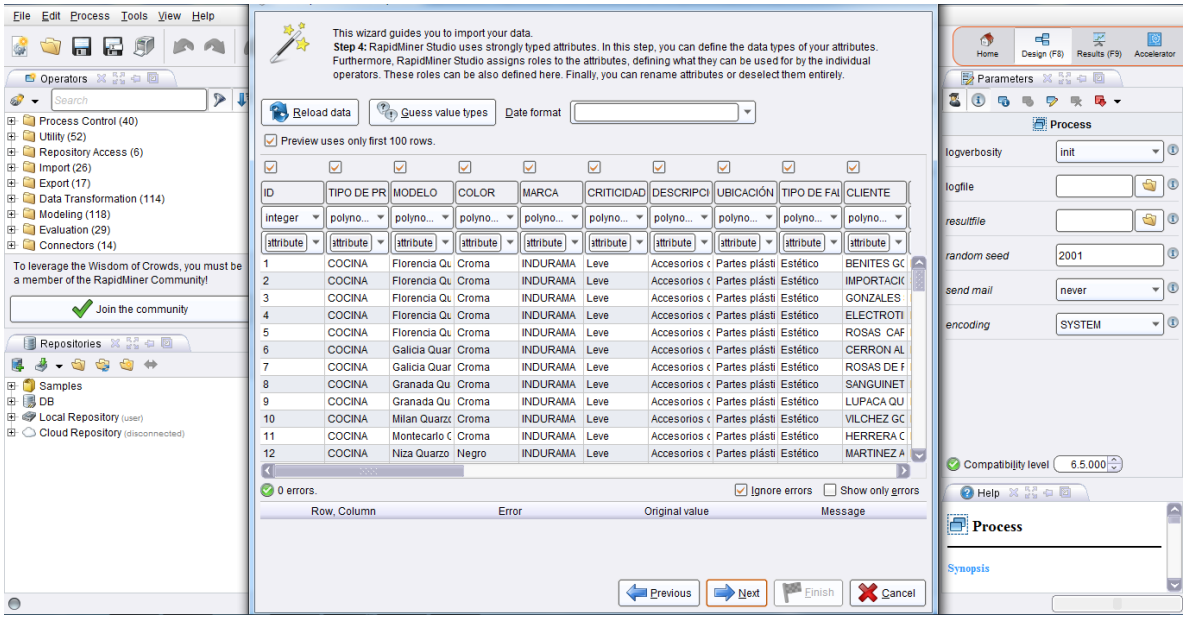
9. ANEXOS

9.1. ANEXO A: DESCRIPCIÓN Y ESQUEMA DE DISEÑO DEL MODELO PREDICTIVO – ARBOL DE DECISIÓN, EN RAPIDMINER

1. Importar planilla Excel

En esta etapa se carga la base de datos de los productos en devolución, indicando el tipo de variable de cada una de las columnas del Excel, ya sea binomial, polinomial, real, etc.

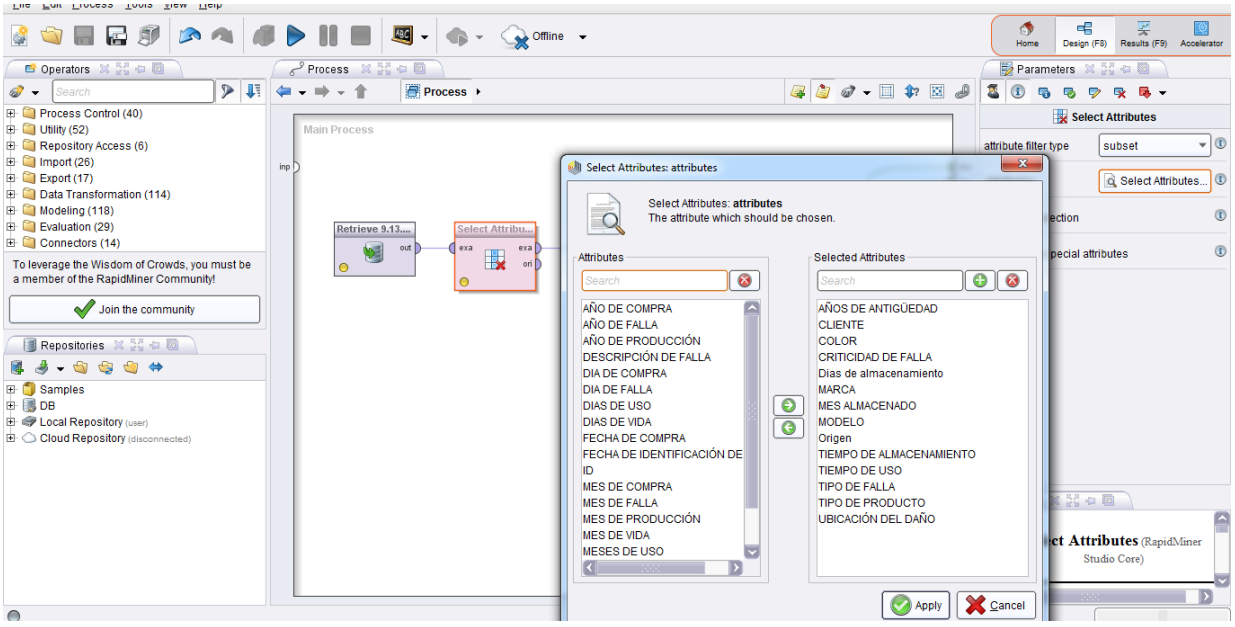




2. Diseño del área de trabajo

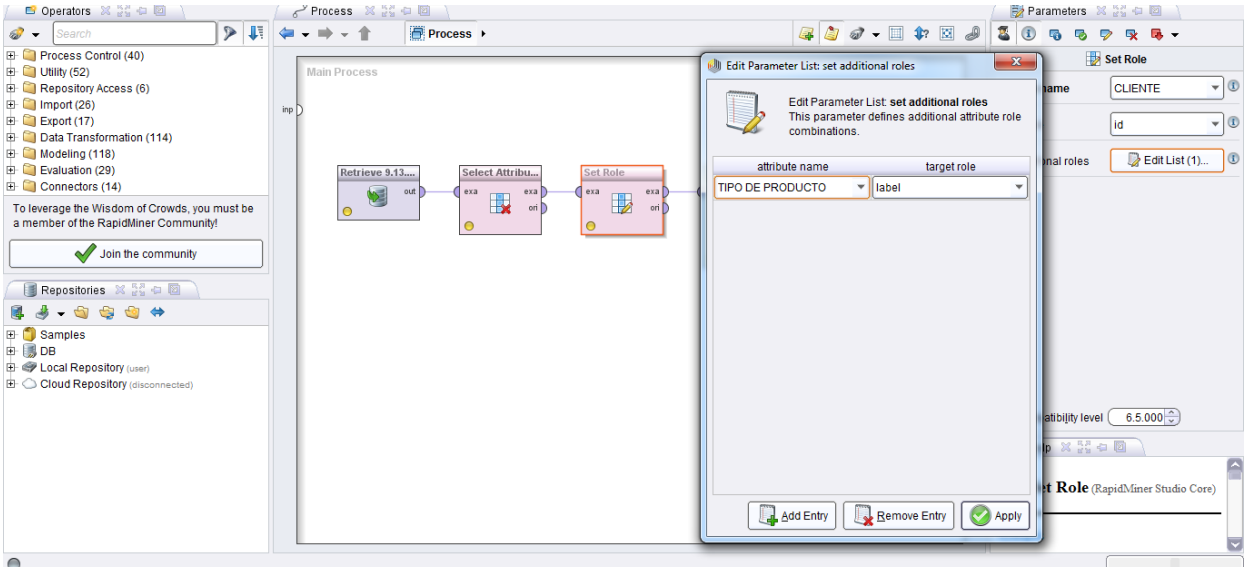
Selección del componente “Select Attributes”

Se inicia con el traslado del recuadro de la Base de datos hacia el área de trabajo y se selecciona el componente “Select Attributes”, con el uso del filtro “Subset”, el cual permitirá seleccionar los atributos a modelar.



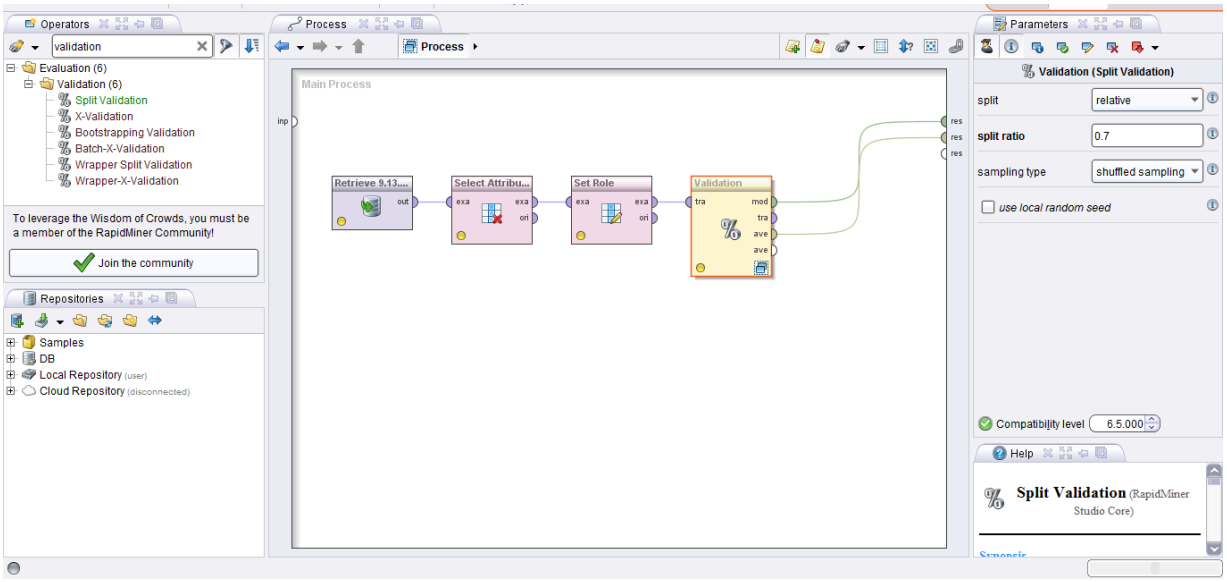
Selección del componente “Set Role”

Luego se selecciona el componente “Set Role”, indicando que el Cliente tiene el rol de ID y en la lista se indica que la variable “Tipo de producto” es la variable dependiente del set de datos, lo cual se indica a través del rol “Label”.



Selección del componente "Validation"

Una vez concluida la modelación, se continúa con el diseño del entrenamiento y validación. Para lo cual se utiliza el componente "Validation" específicamente "Split Validation", en donde se indica el tipo de validación relativa, con un 70% de los registro seleccionados para entrenamiento y con muestreo al azar.



Configuración de "Training" y "Testing"

Por último se ingresa a través del componente "Validation", a la configuración del "Training" y "Testing".

- En "Training" se procede a entrenar el modelo, que en este caso es el modelo de Árbol de decisión, seleccionando "Decision Tree", conectando la salida del modelo hacia el área de "Testing". Para configurar el componente "Decision Tree":

  - Lo primero que se indica es el criterio de particionamiento, que en este caso por estar utilizando datos tanto continuos como binomiales, se utilizará el "Gini_index".

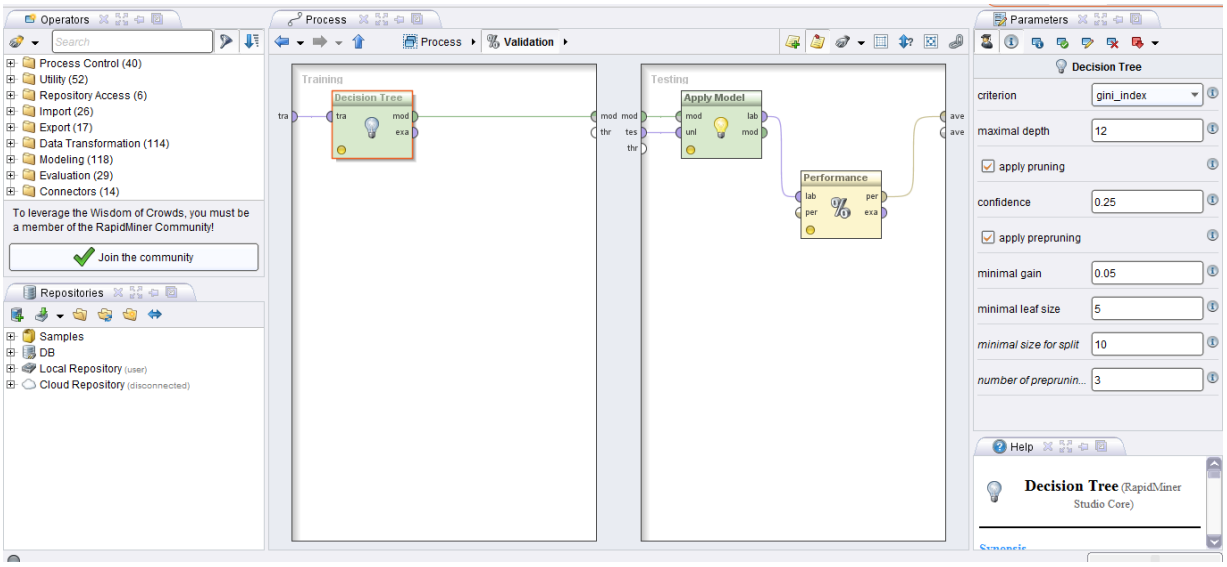
  - Posteriormente se indica la máxima profundidad que se le permite al árbol, que para el caso en estudio será 12.

  - A continuación se configura la poda, para lo cual se indica la confianza, que para este caso será 0.25.

  - En los aspectos de pre poda, se indica la ganancia mínima a exigir a los particionamientos "Minimal gain" igual a 0.05.

    Luego el tamaño mínimo de la hoja "Minimal leaf size" igual a 5 y el tamaño mínimo para el corte "Minimal size for Split" igual a 10.

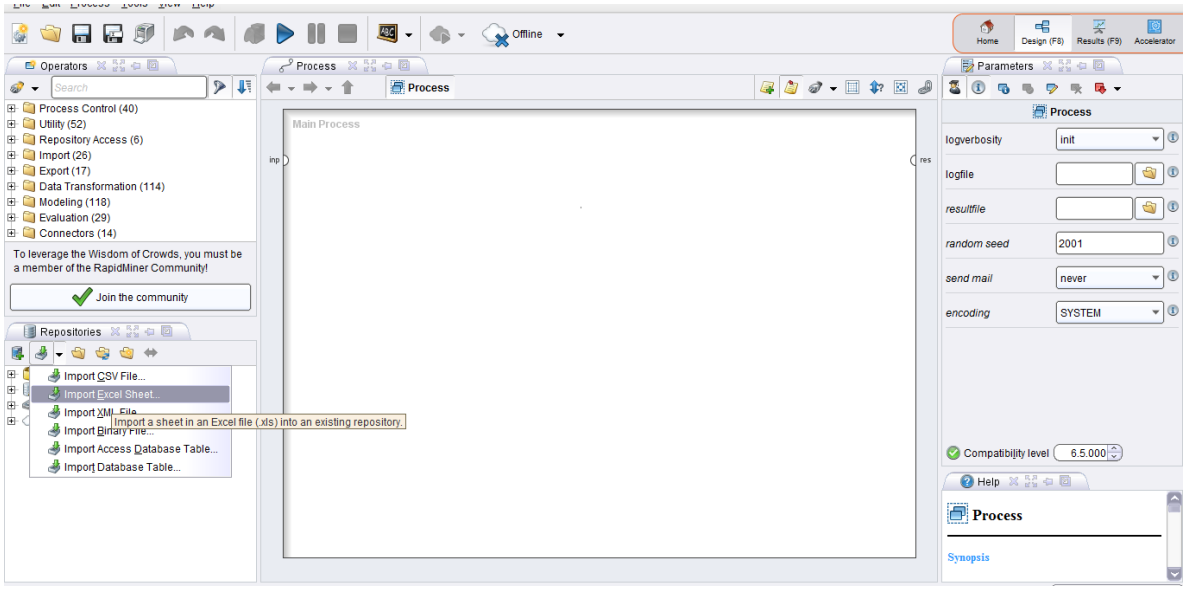
- Por último se configura el área de “Testing”, seleccionando el componente “Apply Model” y para concluir se selecciona “Performance” el cual nos permitirá medir el rendimiento del modelo predictivo diseñado.

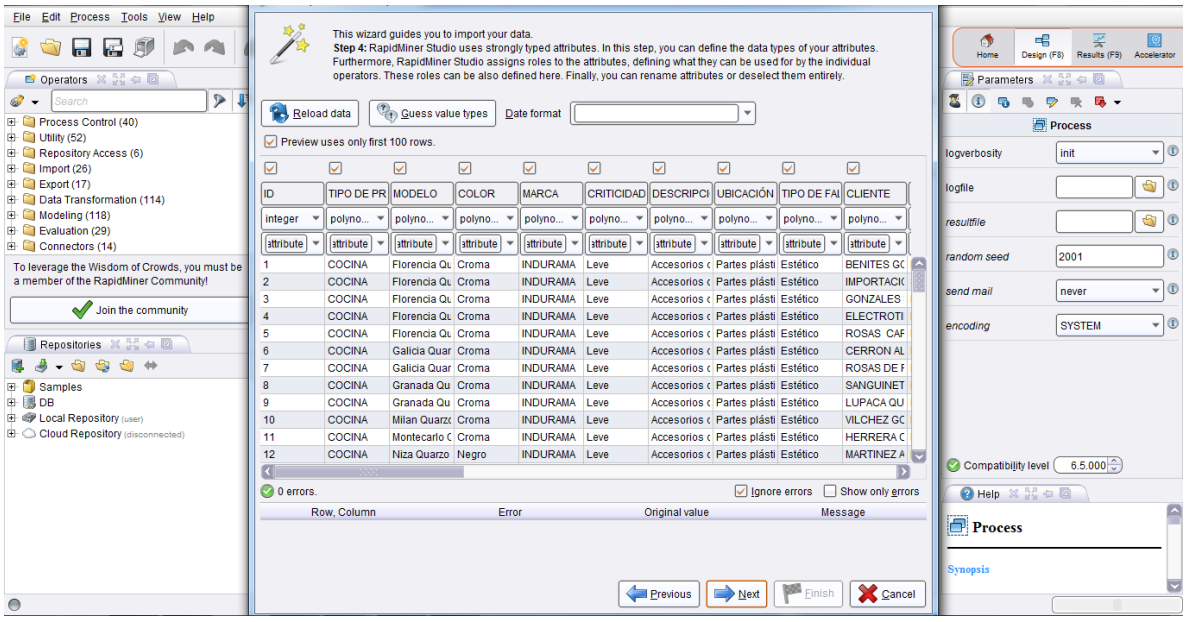


9.2. ANEXO B: ESQUEMA DE DISEÑO DEL MODELO “CLUSTERING”, EN RAPIDMINER

1. Importar planilla Excel

En esta etapa se carga la base de datos de los productos en devolución, indicando el tipo de variable de cada una de las columnas del Excel, ya sea binomial, polinomial, real, etc.

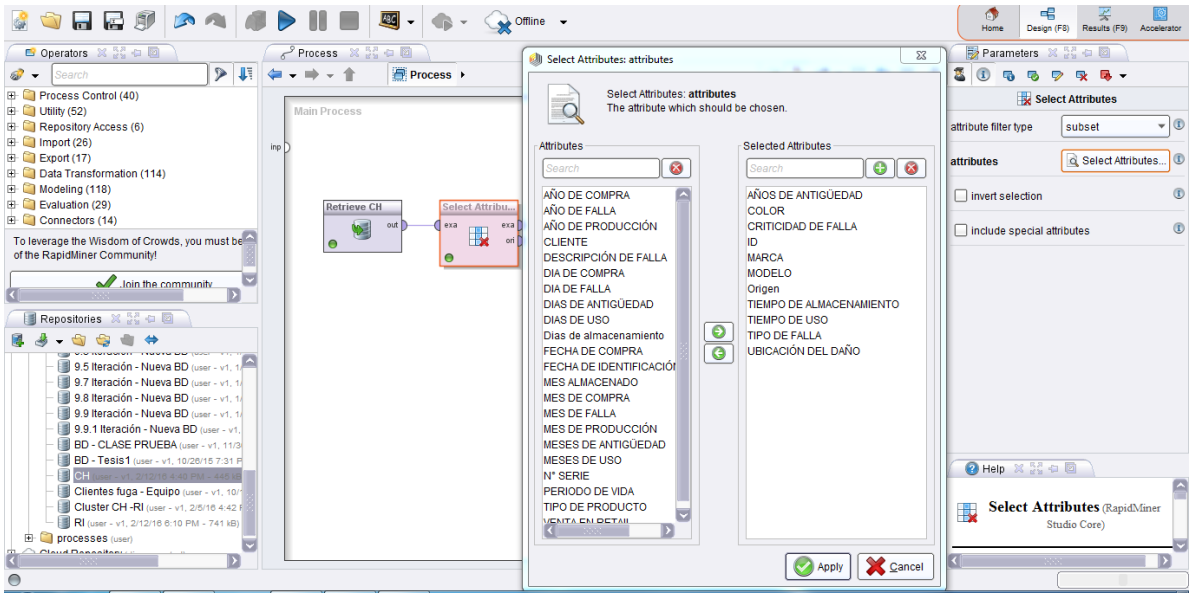




2. Diseño del área de trabajo

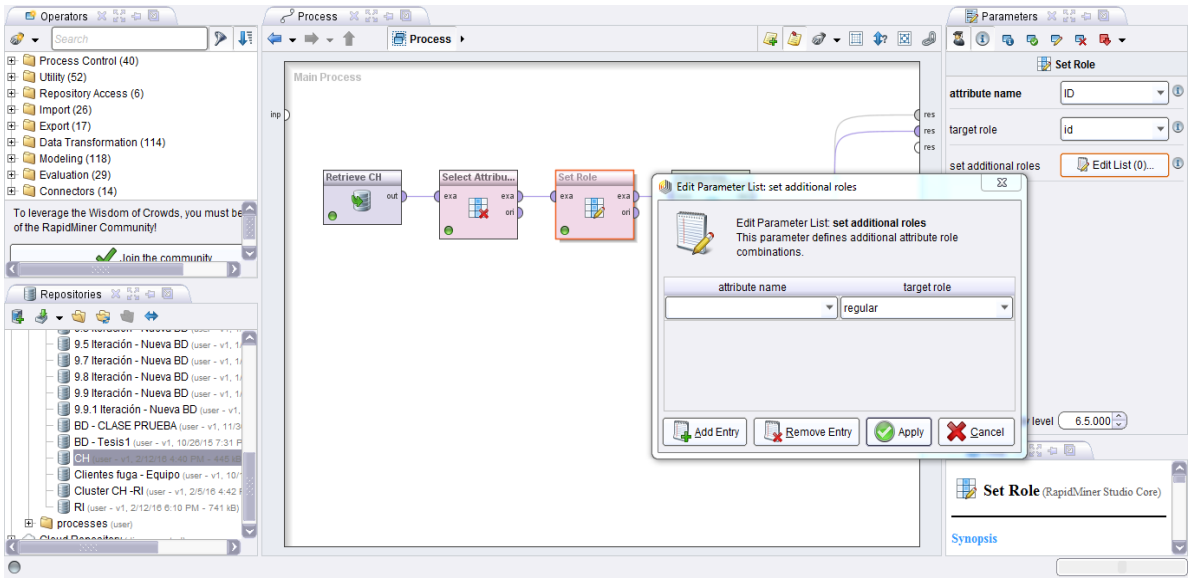
Selección del componente "Select Attributes"

La modelación inicia con el traslado de la Base de datos hacia el área de trabajo y seleccionando el componente "Select Attributes", con el uso del filtro "Subset", el cual permite seleccionar los atributos a modelar.



Selección del componente "Set Role"

Luego se selecciona el componente "Set Role", indicando en "Target role" como ID, ya no siendo necesario indicar en la lista la variable dependiente, ya que en este caso el algoritmo "K-Means" a utilizar es un algoritmo no supervisado.



Selección del componente “Clustering”

A continuación se selecciona el algoritmo “K-Means”, que pertenece al grupo de algoritmos de Clustering. En su configuración, se indica Kmin = 2, Kmax = 10 y en “Mixed types” = MixedMeasures, que permite trabajar con variables tanto continuas como nominales.

Y por último para completar el flujo, se conecta las dos salidas del modelo de Cluster a las dos salidas del flujo completo.