1. **TÍTULO: Desarrollo de modelos de características mediante técnicas de minería de datos**

2. **PROGRAMA: Maestría en Ingeniería, Énfasis: Ingeniería del software.**

3. **NOMBRE DEL ESTUDIANTE Y DEL GRUPO DE INVESTIGACIÓN RESPONSABLE: José Miguel García (josepplloo@gmail.com), Grupo de investigación Ingeniería y Tecnologías de las Organizaciones y de la Sociedad - ITOS**

4. **NOMBRE DEL DIRECTOR: Germán Urrego (gaurrego015@gmail.com)**

5. **NOMBRE DEL ASESOR: Germán Urrego (gaurrego015@gmail.com)**

6. **RESUMEN DEL PROYECTO**

La minería de datos se define como un conjunto de técnicas aplicables a diversas representaciones de datos que tienen como objetivo la asociación, predicción, clasificación, transformación, carga y extracción de información a partir de dichos datos. El reto en la minería de datos, más que el tamaño, formato o el almacenamiento de la información, es el análisis de la misma, es decir, la implementación de la técnica que se ajuste más a la situación que se desea evaluar [1]. Escenarios tan simples como el conocimiento de las tendencias en los consumidores —para lo cual es preciso ajustar los datos a una curva mediante técnicas estadísticas y descriptivas— son muy llamativos para las industrias innovadoras que quieren ajustar sus productos a las necesidades de sus clientes, mejorando la capacidad productiva de su empresa, abaratando costos productivos y generando mayores ingresos.

Mediante el uso de la minería de datos y de diferentes técnicas de agrupamiento es posible recolectar la información necesaria para crear una gran gama de productos del mismo tipo, en donde varíen el color, el tamaño, la capacidad, la velocidad, etc. Como resultado se descubren las relaciones jerárquicas y las restricciones que se deben tener en cuenta a la hora de generar masivamente los productos que las compañías desean producir. Este proceso se encuentra enmarcado dentro de la etapa de análisis de la ingeniería de líneas de productos y, a partir del mismo, se conciben los modelos de características. Las líneas de producto tienen como paradigma compartir y administrar una gran cantidad de características con el objetivo de construir productos que satisfagan las necesidades específicas de un segmento o misión particular del mercado y que se desarrollan a partir de un conjunto común de activos básicos de una manera prescrita [2]. De esta forma, compañías del sector bancario, tecnológico y manufacturero han hecho uso de las técnicas de minería de datos para el desarrollo de sus productos mediante la ingeniería de líneas de producto[3], [4], [5].

En la actualidad, los modelos de características se generan de forma manual. Este trabajo se fundamenta en los conceptos de minería de datos e ingeniería de líneas de producto con el propósito de crear de forma autónoma los modelos de características; teniendo en cuenta la información arrojada por las técnicas de

minería de datos y buscando que estos modelos sean de utilidad en el desarrollo de las líneas de productos.

Palabras clave:
Minería de datos, Líneas de producto, Modelos de características.

## 6. ABSTRACT

Data mining is defined as a set of techniques applicable to various data representations with aim association, prediction, classification, transformation, loading and extraction of information from such data. The challenge in data mining, rather than the size, format or storage of information, is the analysis of the same, that is, the implementation of the technique that best fits the situation to be evaluated [1] . Scenarios as simple as knowledge of consumer trends —which data must be adjusted to a curve by statistical and descriptive techniques— are striking for innovative industries that want to adjust their products to the needs of their customers, improving the productive capacity of the company, lowering production costs and generating higher revenues.

Using data mining and different clustering techniques it is possible to collect the information needed to create a wide range of products of the same type, where color, size, capacity, speed, etc. are varying. Thus, the hierarchical relationships and the restrictions that must be considered when massively generating products that companies wish to produce are discovered. This process is framed within the stage of analysis of the engineering of product lines and, from the same, the models of characteristics are conceived. The product lines have as paradigm to share and manage great number of characteristics with the aim of constructing products that satisfy the specific needs of a market segment or mission, and that are developed from a common set of basic assets in a prescribed way [2]. In this way, companies in the banking, technology and manufacturing sectors have made use of data mining techniques for the development of their products through product lines engineering [3], [4], [5].

Currently, models of characteristic are generated manually. This work is based on the concepts of data mining and product line engineering with the purpose of creating autonomous models of characteristics; considering the information provided by the techniques of data mining and looking for these models to be useful in the development of the product lines.

Keywords:
Data mining, Product line engineering, Feature models.

## 7. PLANTEAMIENTO DEL PROBLEMA

En la actualidad la sociedad se enfrenta a un cambio de paradigma en los sistemas de comunicación e información. Debido a la masificación de la tecnología a nivel

mundial, el mejoramiento de las tecnologías de la información y la comunicación (TIC) se ha convertido en un factor clave en el desempeño productivo y el crecimiento económico e industrial. En Colombia, las empresas han aumentado significativamente el uso de las computadoras y el Internet. Se estima que para el año 2014 de 8.659 empresas el 99% poseía computador y estaba conectada a Internet [6]. El Ministerio de Tecnologías de la información y las comunicaciones ha invertido hasta $373.993 millones de pesos Colombianos hasta marzo del 2014 solo en el proyecto de conectividad de alta velocidad, el cual busca que el 100% de los municipios del país tengan acceso a Internet de alta velocidad [7].

Los avances mencionados anteriormente han generado que las industrias modernas puedan almacenar grandes cantidades de datos en diferentes sistemas de información. Estos datos crecen rápidamente al ser recolectados por todo tipo de dispositivos, y son coleccionados por las industrias porque son una fuente valiosa de conocimiento, la cual puede ser usada para mejorar las decisiones relacionadas con la productividad. Sin embargo, actualmente el uso de estos datos históricos es limitado, ya que una gran cantidad de productos y datos quedan aislados y dispersos en los diferentes sistemas [8] generando que las industrias sean ricas en datos, pero pobres en información [9]. De esta manera, la organización de estos datos y la búsqueda de conocimiento se convierte en un desafío para la minería de datos [10].

La variabilidad de aplicaciones que generan el tráfico de datos en Internet es uno de los temas de investigación de las líneas de producto de software dinámicas, la *autonomic computing* y los *web services* [11], ya que debido a dicha variabilidad los modelos de características industriales incluyen cientos de características derivadas de las preferencias de los clientes, lo cual los hace muy complejos y difíciles de configurar [12]; esto a su vez genera problemas y dificultades en la configuración y caracterización de los productos personalizados. Por estas razones surge la siguiente pregunta de investigación: ¿Cómo las técnicas de minería de datos pueden transformar los datos históricos de las compañías en modelos de características?

# 8. MARCO TEÓRICO Y ESTADO DEL ARTE

## 8.1. MARCO TEORICO

### Minería de Datos

La minería de datos es el proceso de encontrar patrones y relaciones en los datos con el fin de realizar actividades descriptivas y predictivas. La minería de datos descriptiva busca descubrir en grandes volúmenes de datos las estructuras, relaciones, tendencias, grupos y valores atípicos que están contenidos en los datos. Por su parte, la minería de datos predictiva construye modelos y procedimientos de regresión, clasificación, reconocimiento de patrones y tareas de aprendizaje de

máquinas que evalúan la capacidad predictiva de estos modelos en datos frescos o nuevos [13].
El modelamiento de los datos mediante las técnicas de minería de datos puede ser usado para predecir el comportamiento de un individuo, segmentar una población, determinar las relaciones entre una población, determinar las características que más afectan a un resultado en particular; y en las empresas, estas técnicas tienen el objetivo de desarrollar estrategias competitivas en. Los datos de las compañías que se analizan pueden presentarse en todo tipo de formatos y estructuras y pueden estar almacenados en todo tipo de infraestructura [14]. Por esta razón la minería de datos ha optado por clasificarse en varios tipos de funciones y técnicas que no se apartan del proceso tradicional para minar o extraer datos, como se ilustra en la Figura 1.

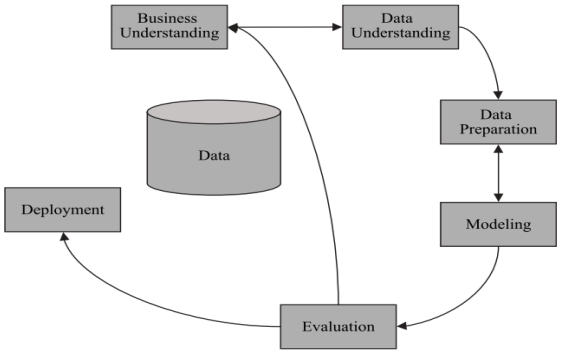


**Figura 1**: Proceso Industrial estándar para minar datos

La Figura 1 ilustra el proceso empleado tradicionalmente en la minería de datos, *Cross Industry Standard Process for Data Mining (CRISP-DM)* [15]. Es importante diferenciar las técnicas o funciones que se aplican para cada una de las etapas de este proceso:

1. **Entendimiento del negocio**. El primer paso es el más importante, se llama entendimiento del negocio y se fundamenta en la identificación del problema, en esta fase se analiza qué recursos son pertinentes para solucionar el problema y en algunos casos el retorno de la inversión [15].
2. **Entendimiento de los datos**. En este paso se deben identificar los datos a través de métodos estadísticos, por ejemplo, en qué rango de valores se encuentran los datos y qué fuentes de datos son necesarias para la solución del problema [15].
3. **Preparación de los datos**. En esta etapa los datos ya se encuentran identificados y es necesario limpiarlos y formatearlos para presentarlos correctamente y no crear ambigüedad o manejar escalas de medición diferentes, por ejemplo, el género de los sujetos que aparece como masculino y femenino se muestra como 0 y 1 respectivamente [15].
4. **Modelado**. En esta etapa es en donde se aplican las diferentes técnicas de regresión y clasificación para minar los datos dependiendo del análisis que el usuario haya realizado en la etapa de entendimiento del negocio [15].

5. **Evaluación**. Esta etapa es necesaria para probar que el modelo alcanzó los objetivos y mostró los resultados esperados. Durante la etapa de evaluación se puede direccionar el esfuerzo a mirar aspectos que no se hayan cubierto en la etapa de modelado [15].
6. **Despliegue**. La etapa de despliegue termina con un reporte. Una solución de minería de datos que se puede repetir y está integrada con el proceso de negocio de la compañía [15].

Los datos históricos que se encuentran almacenados en los diferentes silos de datos en las compañías contienen una gran cantidad de registros que deben ser explorados para extraer el conocimiento. El descubrimiento de estos datos se conoce como conocimiento descubierto en bases de datos o KDD de sus siglas en inglés Knowledge discovery in databases [13], [15]. Los procedimientos que se realizan para analizar estos datos se pueden clasificar en las siguientes técnicas y funciones.

1. **Clasificación**: Se usa para hacer predicciones en ejercicios como las encuestas, la segmentación de usuarios, el análisis de créditos bancarios, procesos estadísticos, detección de patrones etc.[13] La clasificación de datos se realiza dependiendo de sus valores, por ejemplo, si se tiene a una persona que es de sexo femenino, y su estatura es mayor de 180 cm, se clasifica su talla como L.
2. **Regresión**: Es usada para hacer predicciones en escenarios de datos continuos; es valiosa para el *forecast*, predicciones de series de tiempo y modelos médicos y ambientales [15]. La regresión a diferencia de la clasificación usa los datos de entrada para crear una función de entrenamiento y poder realizar pronósticos sobre estos datos de entrada [13], por ejemplo, el valor de venta para una casa estará dado por la cantidad de baños multiplicado por su área.
3. **Atributo importancia**: En el análisis de negocio es muy importante identificar las características en los datos que influencian el comportamiento de los beneficios económicos [15]. Para tener una idea rápida de un dato como este, basta pensar en el estrato socioeconómico y su importancia para determinar el cobro de la cuenta de los servicios públicos.
4. **Asociación**: Es una técnica fuertemente usada en el análisis transaccional, se especializa en encontrar implicaciones en los datos o dependencias entre elementos repetidos en diferentes transacciones [15]. En la Figura 2 se puede ver el descubrimiento de una asociación.

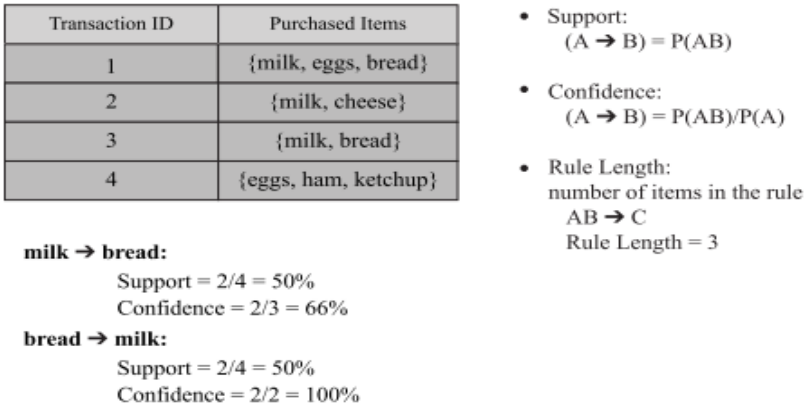


**Figura 2**: Soporte de una regla de asociación.

La Figura 2 ilustra que hay una asociación entre el pan y la leche, que se basa en las veces que aparecen los dos productos en una transacción.

5. ***Clustering* o Agrupamiento**: Esta es una función muy importante en la logística, en las cadenas de producción, en el análisis genético y en la minería de texto [15]. En el *clustering* se compara un objeto de un conjunto de datos contra muchos conjuntos de datos, el resultado de este algoritmo se llama dendrograma. Los algoritmos de *clustering* son iterativos y pueden retornar la distancia entre cada resultado o incluso crear jerarquías de los datos encontrados en los diferentes conjuntos o clúster. Los algoritmos más usados son: *K-means clustering, K-medioids clustering, Hierarchical clustering, Kernel K-means, soft K-means*, etc. La diferencia de los algoritmos radica en la función de distancia usada [13].
6. **Predicción**: Es una función que ofrece una salida dependiendo de un conjunto histórico de datos; usa árboles de decisión y redes neuronales para argumentar su salida y procede dependiendo del conjunto de datos de entrada, por ejemplo puede ser usada para detectar defectos o calcular el mantenimiento de una máquina [9].
7. **Estimación de densidad no paramétrica**: Es una técnica alternativa para el estudio de los datos multivariados, en donde estos datos no pertenecen a una distribución de probabilidad conocida. Como resultado al aplicar esta técnica se obtiene la estimación no paramétrica de una función de distribución de densidad para los datos [15].
8. **Inferencia**: El objetivo de esta función consiste en estimar las relaciones existentes entre dos variables, especialmente como la variable dependiente cambia en función de las independientes, $Y=f(X)$. Se puede hablar de tres tipos de funciones en esta categoría, las funciones paramétricas, las funciones no paramétricas y las funciones semi paramétricas [13].
9. **Remuestreo**: El objetivo de estas técnicas es generar nuevos datos a partir de un modelo teniendo en cuenta la flexibilidad y el error. Las técnicas de

remuestreo son muy costosas a nivel computacional y entre las funciones más usadas se encuentran Bootstrap y Cross-Validation [10].

10. **Subset selection**: Es una función de selección que identifica un grupo de predictores o asume unas variables *X* que tiene mucha influencia en la variable respuesta *Y*, con esta información, la función crea un modelo que se ajusta a los predictores X mediante la suma de sus cuadrados [1],[13].
11. **Shrinkage or Regularization**: Es una función de selección de características que da muy buenos resultados por su ajuste a la varianza de los datos, estos métodos proponen seleccionar las variables que más aportan en la suma de cuadrados y penalizar las que no aportan [13]. Los algoritmos más conocidos que aplican esta función son *Ridge Regression* y *Least Absolute Shrinkage and Selection Operation (LASSO)* propuesto en 1996 por Tibshirani [1]. Las generalizaciones del algoritmo *LASSO* se han convertido en un tema de gran interés, entre las más importantes están: *Generalized Linear Models (GLM), Elastic Net, Dantzig selector, SVN (Support Vector Machine), high dimensional matrix estimation* y *multivariate methods*.

Existe una relación entre las técnicas de la minería de datos y las funciones usadas en la minería de datos cada técnica y función tiene su dominio y su contexto de uso, el cual varía según los tipos de datos y la finalidad en la implementación [16]. En la Figura 3 se presentan las relaciones más comunes entre técnicas y funciones.

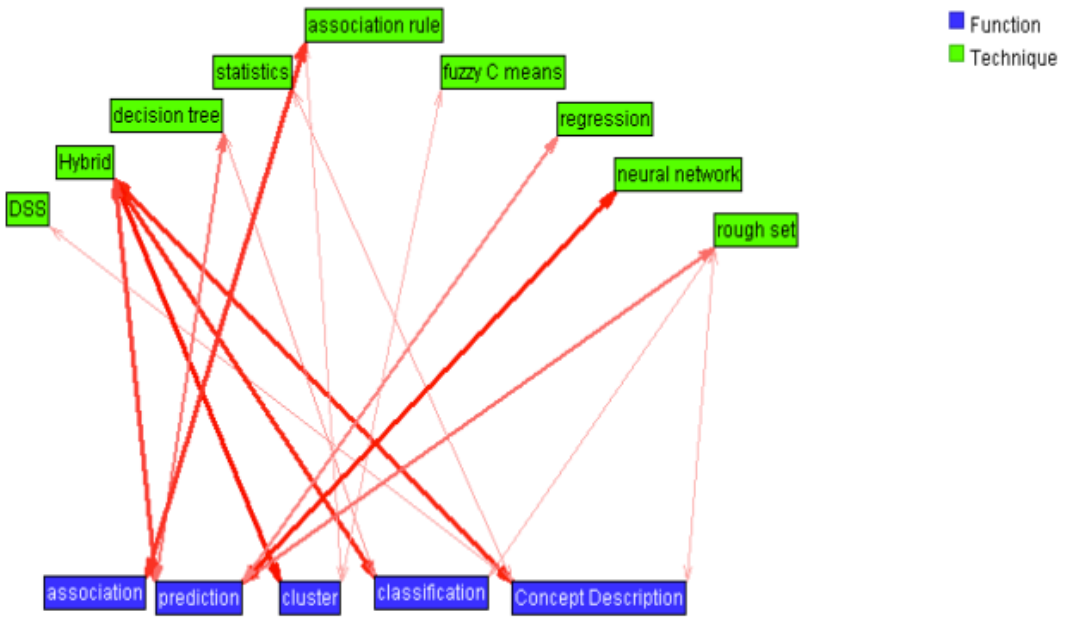


**Figura 3**: Funciones y técnicas en la minería de datos.

En búsqueda del propósito de construir modelos de características es necesario ampliar la información en los métodos que generen árboles y las soluciones a los

problemas de agrupamiento, penalización y clasificación. Entre los más importantes están los Árboles de regresión y clasificación (CART) y C4.5. Los árboles de clasificación son métodos que tratan de dividir o partir los datos desde el principio hasta el final, estos métodos dedican su esfuerzo a estimar esos puntos de inicio y fin en los datos y la distancia de la partición o en número de tamaño de la división [1]. En la Figura 4 se presenta la esencia de los algoritmos basados en árboles.

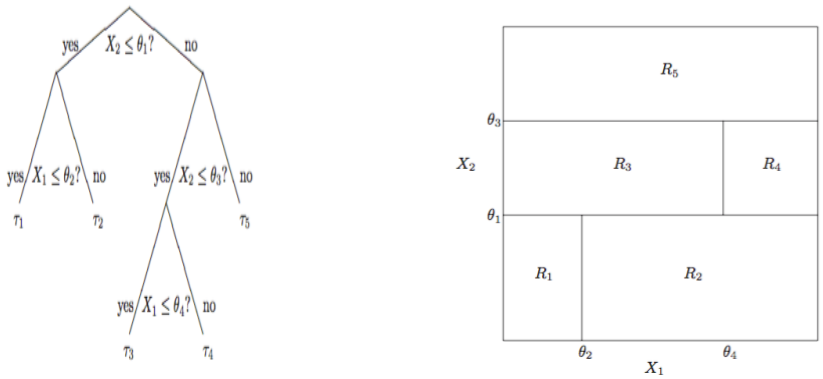


**Figura 4**: Ejemplo del funcionamiento de un algoritmo CART.

Un árbol es la representación de un conjunto de áreas, ahora bien, si los datos constituyen una nube de puntos, el árbol puede seleccionar las áreas en las cuales se divide dicha nube de puntos. Además de usar árboles, también es común implementar métodos predictivos como *Bootstrap* o *Cross validation* para ajustar los puntos a cada área como se muestra en las imágenes que componen la Figura 5.

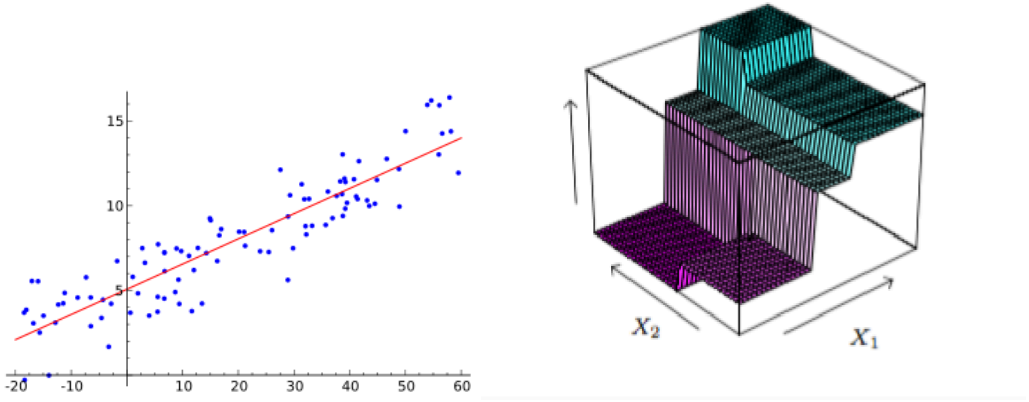


**Figura 5**: Áreas de ajuste para los datos.

En la Figura 5 la formación de las áreas a partir de un conjunto de datos está dada por los valores *X1* y *X2*, se puede observar que la Ecuación 1 representa los futuros puntos que se ubicarán en las cinco regiones que se muestran en la Figura 5 [13], [17].

$$\hat{f}(X) = \sum_{m=1}^{5} c_m I\{(X_1, X_2) \in R_m\}$$

**Ecuación 1**: Representación de un árbol de regresión con 5 regiones. Donde se estiman los valores de $f(X)$, los cuales están dados por los tamaños ($c_m$) de las divisiones de los datos de entrada en el intervalo (I) cuando estos pertenecen a una región ($R_m$) determinada.

En la Ecuación 1 se satisface la condición de asignación de los puntos a unas regiones señaladas como ramas de un árbol, donde la interpretabilidad puede verse reducida a medida de qué el árbol crezca o tenga más ramas. Un punto de discusión es la estimación del corte del árbol, esta decisión es delicada cuando se puede ver afectada por la variabilidad de las observaciones [13], por ejemplo, la detección de spam en los correos electrónicos es un escenario en donde se usan estos modelos, aunque en este ejemplo la vida de una persona no se vea comprometida por la mala asignación de su correo, estos métodos también apoyan la selección de los pacientes que sufren de una enfermedad en el corazón [13]. Por otra parte, Las líneas de producto también tienen un modelo muy similar a la Ecuación 1, en la cual para definir familias de productos es necesario generar un árbol partiendo de la asignación de las características a las regiones que serán los segmentos del mercado.

## Ingeniería de Líneas de producto

La ingeniería de líneas de producto (*PLE*) tiene como objetivo la producción de conjuntos de productos con más características comunes que diferentes, estas líneas de producto (PL) se han convertido en un paradigma viable para mejorar la productividad y la calidad de la producción en masa [18]. La producción en masa, es el legado de la revolución industrial y está definida como la producción de un conjunto estandarizado de un mismo producto, en donde la personalización de este se convierte en un reto interesante para las compañías de cualquier tipo, inclusive para las empresas de desarrollo de software. Las características en los productos son el insumo en la línea de producto, su razón de ser. Un producto tiene diversas características que pueden representarse en un modelo de características, por esta razón los modelos de características son comúnmente la representación de las líneas de producto [19], [20].

Las líneas de producto protagonizan la etapa de diseño de los nuevos productos que consumen los clientes y la minería de datos se especializa en extraer la información sobre la tendencia de los clientes en el mercado, la integración de estas dos tecnologías proporciona a las empresas información útil para el desarrollo de nuevos productos que se adapten a las necesidades de la sociedad [3] , [23]. La minería de datos puede solucionar una gran cantidad de problemas relacionados con el entendimiento de datos científicos (¿Cuáles son las causas raíces de un error?) y datos de negocios de cualquier dominio (¿Cuál es el producto que compran más sus clientes?) [14], [5]. Saber qué se puede producir, con la certeza de ser

comprado es lo que motiva a las compañías a crear productos de manera masiva y mejorar su producción para hacerla cada vez más rápida y flexible [18]. Estas tecnologías aplican para cualquier tipo de dominio y es de vital importancia para la salud del planeta reciclar estos datos, porque las empresas al saber la demanda de sus productos y servicios pueden afinar sus procesos de producción, ahorrando dinero y recursos [24]. Estos productos pueden ser software también, incorporar líneas de producto en el ciclo de vida de desarrollo de software (SDLC) mejora las estadísticas de mercadeo ampliando los beneficios de dos a siete veces. En la figura 6 e observa el SDLC con las adecuaciones para soportar el reúso y la variabilidad [20].

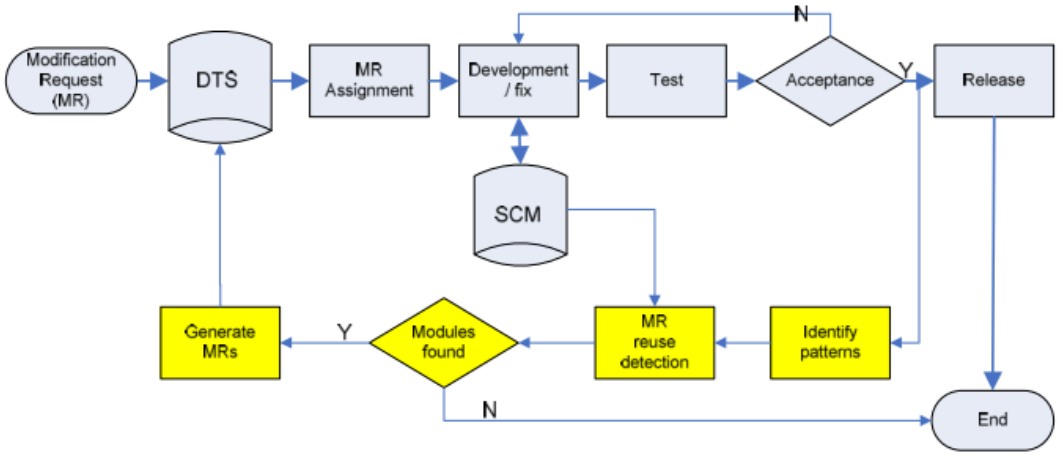


**Figura 6**: Ciclo de vida del desarrollo de software.

Antes de la salida a producción se realiza una identificación de patrones y se detectan los componentes que se pueden reutilizar y generar como nuevos módulos teniendo en cuenta la variabilidad y el reúso de los componentes.

## Modelos de características

Un modelo de características representa la información de todos los posibles productos en una línea de productos en términos de las características y las relaciones entre ellas [20]. La Figura 7 muestra un ejemplo del modelo de características donde se expone un teléfono celular.

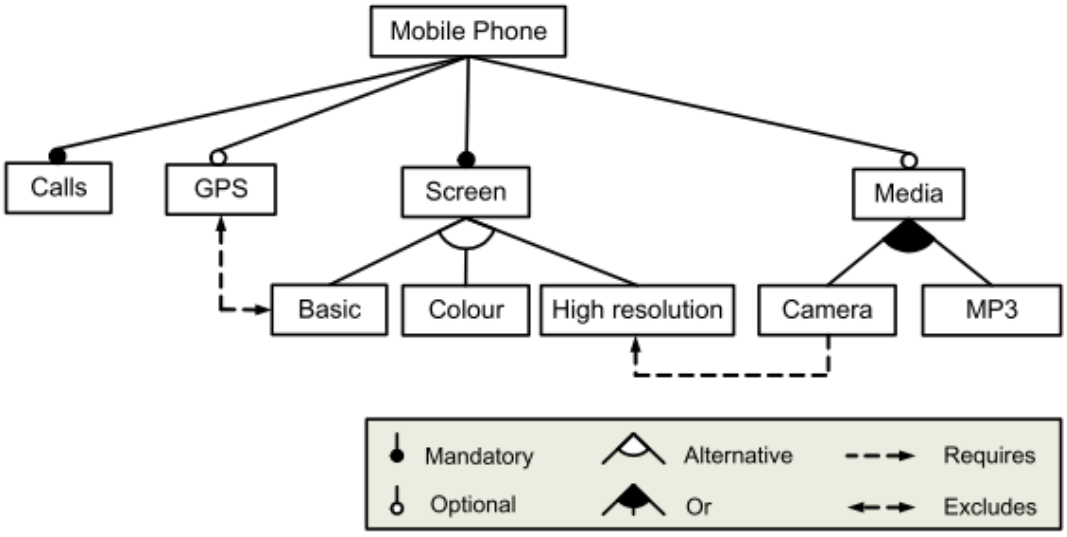


**Figura 7**: Modelo de características que representa un celular.

En la imagen anterior se agrupan las características del teléfono celular y se relacionen entre ellas de forma jerárquica, las relaciones pueden ser Obligatorio, Opcional, Alternativo, O, inclusión y exclusión [20].

1. **Obligatoria**: Esto significa que el hijo asignado tiene que estar incluido en el producto.
2. **Opcional**: Esto significa que el hijo asignado puede o no tener esta característica.
3. **Alternativo**: Da a conocer que solo una característica en la jerarquía puede ser seleccionada.
4. **O**: Esta relación significa que se puede seleccionar todas o ninguna de las características en la jerarquía.
5. **Requerida**: Cuando una característica requiere a otra, esta no puede existir sin la presencia de la otra.
6. **Exclusión**: Si una característica excluye a otra da a entender que las características seleccionadas no pueden ser parte del mismo producto.

La comunidad ha creado herramientas de ingeniería de software asistido por computadora, *Computer Aided Software Engineering* (CASE) para el modelamiento y la configuración de los modelos de características, aunque están en constante desarrollo, se conocen algunas muy populares como VariaMos [21], la cual tiene como objetivo desarrollar familias de sistemas y también tiene herramientas para realizar operaciones sobre otros modelos. SPLOT [22] (*software product line online tool*), la cual es una herramienta online para la configuración de características y la derivación de productos a partir de diagramas de características y modelos de variabilidad. La intención de esta investigación no será competir con estas herramientas sino complementarlas y extender su funcionalidad a la adquisición de datos que ofrecen las técnicas de agrupamiento dentro de la minería de datos.

## 8.2. ESTADO DEL ARTE

En la actualidad las empresas en Colombia usan las computadoras y el Internet más que antes, se estima que de 8.659 empresas el 99% posee computador y está conectada a Internet [7]. En la Figura 8 se muestran los indicadores básicos de tenencia y uso de la información y la comunicación en las empresas.

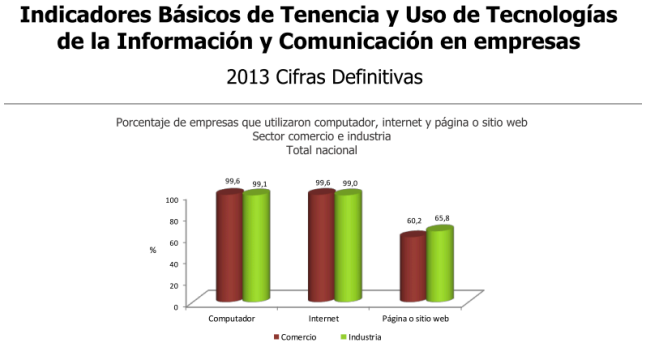


**Figura 8**: Indicadores del uso del computador, el internet y las páginas web en Colombia.

El Ministerio de Tecnologías de la información y las comunicaciones ha invertido hasta $373.993 millones de pesos colombianos hasta marzo del 2014 solo en el proyecto de conectividad de alta velocidad, el cual busca que el 100% de los municipios del país tengan acceso a Internet de alta velocidad [25].

Hace 50 años el análisis estadístico multivariado hubiera usado métodos lineales para descubrir el conocimiento almacenado en los datos históricos que se generan por las aplicaciones y el Internet, lamentablemente esta cantidad de datos hubiera sido un problema en esa época. Actualmente, el análisis computacional de los datos está ganando popularidad, desde los años setenta la interactividad de los computadores apoyo el análisis exploratorio de los datos, en las décadas posteriores, fuimos testigos de un avance en el procesamiento y el almacenamiento de las computadoras. Grandes cantidades de datos fueron clasificados, almacenados y administrados de forma eficiente gracias a los paquetes interactivos de software estadísticos, esto facilitó el análisis avanzado de los datos sin mucho esfuerzo, los procesos de extracción, transformación y carga fueron más comunes, los datos de máquina y la internet dieron inicio a nuevas disciplinas como la minería de datos y el aprendizaje estadístico.
Las enormes cantidades de datos son cada vez más comunes y las personas encargadas de analizarlos siguen teniendo en cuenta las técnicas supervisadas, aunque el descubrimiento de información no supervisado sea la nueva tendencia. Como consecuencia, la estadística multivariada incluye nuevas técnicas desde las ciencias de la computación, muchas de ellas aún en su etapa de desarrollo. Los orígenes de estas técnicas son algoritmos derivados del modelado, la optimización y el razonamiento probabilístico, el desarrollo constante de las comunidades en el

área han madurado y convertido la minería de datos, el aprendizaje estadístico y la inteligencia artificial en un excelente marco de trabajo [1].

En este documento se exploran varios ejemplos del uso de la minería de datos en las líneas de producto. Se partirá de que en un ambiente heterogéneo, donde existen diferentes compañías que desarrollan productos masivamente, se pueden usar en común muchas técnicas de análisis de datos [27], [28], como funciones descriptivas, predictivas y técnicas asociativas. Los ejemplos presentados a continuación son compañías que quisieron adoptar este nuevo paradigma para incluirlo en el software que desarrolla sus productos:

**1. Creación de un sistema de puntuación (Banco de Irán).**

Por lo general cuando se quiere aspirar a un crédito en una entidad financiera, los usuarios son puestos bajo un sistema de puntuación, estos son algoritmos dentro de los métodos clasificatorios, que preparan los datos del usuario o los alistan para un proceso posterior. Estos decrementos en la cantidad de datos a procesar, reducen el costo de cómputo que puedan generar las grandes cantidades de datos. Su aplicación es sencilla y su actualización también lo es. La información de los usuarios es ingresada a un modelo que propone el mejor producto crediticio según el comportamiento del usuario, su historial, sus productos o sus relaciones con otras entidades financieras. Mientras exista variabilidad se puede ver un sistema como una línea de productos y la minería de datos puede extraer las variabilidades de un sistema [26].

**2. Planeando nuevas generaciones de productos (Apple).**

Para planear nuevas generaciones de productos, es decir planear la existencia de un determinado producto en el mercado, se deben tener en cuenta todas las características que lo conforman, el color, el precio y el tamaño. El iPhone es un producto que quiere preservar su estado en el mercado, ser reconocido a lo largo del tiempo por ser novedoso y agrupar las mejores características; los usuarios generan los datos que ayudan a conocer sus necesidades y este conocimiento impulsa el desarrollo de nuevos productos como las tablets y de nuevos paradigmas como los Modelos variables de estados dinámicos. Para implementarlos se necesita conocer el historial de las ventas y encontrar su tendencia, según la línea de producto se modela el sistema como un modelo de canibalización, este modelo consume los datos extraídos, busca las características que generan mayor beneficio económico y desecha las que no. Habiendo definido el modelo de canibalización es el turno de la iteración hacia adelante de Monte Carlo, con este algoritmo se originan las generaciones necesarias para que los productos sean competitivos en el mercado [23].

**3. Desarrollando productos con técnicas de minería de datos (cámara digital).**

Con los crecientes avances en la tecnología, los ciclos de vida de desarrollo de los productos deben hacerse cada vez más rápido, generando mayores ingresos, construyendo productos de mayor calidad, reduciendo el costo de producción y orientando los productos a las necesidades del cliente. Es en estas necesidades donde están las pistas para el mejoramiento del ciclo de vida del proceso de

desarrollo de nuevos productos [3]. En el proceso de construcción de una cámara digital se pregunta ¿qué quiere o necesita un cliente de una cámara?
¿Cuáles características son más importantes que otras? ¿Puede integrarse el diseño de los productos con lo que saben los clientes? ¿Cómo pueden estas reglas ayudar a mejorar el diseño de una cámara? Antes de que un producto sea diseñado, muchas compañías tienen en cuenta los datos almacenados en bases de datos multidimensionales para responder las preguntas anteriores y validar sus hipótesis. Además, usan técnicas de la minería de datos que en el contexto de la clasificación, estimación, segmentación y descripción preparan los datos para ser minados, es decir para extraer conocimiento y por ejemplo en el caso de la cámara digital, planear la construcción de una línea de productos [3], [29].

Transversal a estos ejemplos se desean mostrar, sin entrar en detalle, las técnicas usadas para modelar los datos y producir las líneas de producto. En este documento se mostrarán algunas de las técnicas de minería de datos que se aplican a la fase de análisis del dominio en el ciclo de vida de la ingeniería de líneas de producto [27], como se presenta en la Figura 9.

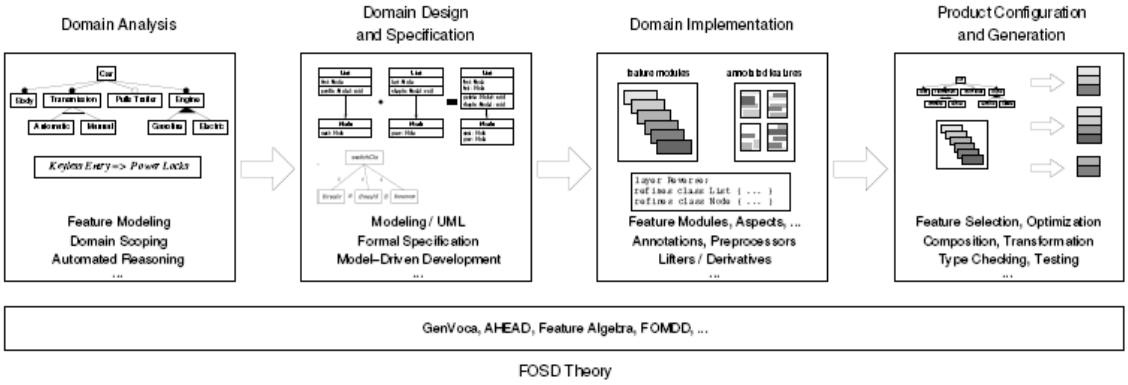


**Figura 9**: Marco de trabajo para la creación de líneas de producto.

Las actividades mostradas en la Figura 9 son propias del análisis del dominio dentro del marco de trabajo de la PLE. En la minería de datos pueden encontrarse características que están relacionadas de forma interesante y pueden ser modeladas mediante árboles de decisión y reglas de asociación [29], [9], [30]. Además, se sabe que estas características pueden estar asociadas a un beneficio económico, y con esta asociación se puede usar un proceso de canibalización o de clasificación, desde el punto de vista de la información que está relacionada con las características específicas que logran el beneficio económico [3].

En los ejemplos anteriores los datos, cuando son sometidos a un proceso de evaluación, generan un patrón de decisión que elimina la redundancia en la información [31]. Cuando se tienen las reglas de agrupamiento se inicia el proceso de construcción del modelo de características, el cual, acompañado de un marco de trabajo, finalmente genera las estrategias de diseño de los nuevos productos [3].

Los ejemplos mencionados en el desarrollo de este estado del arte están orientados a la extracción de las características por medio del proceso *CRISP*, en el cual los datos históricos que son usados actualmente tienen muchos formatos y es preciso adaptar este proceso a la extracción del conocimiento de forma novedosa. La oportunidad de innovar depende del mejoramiento de la arquitectura *CRISP* para que incorpore un nuevo proceso de descubrimiento de las características mediante nuevos algoritmos de agrupamiento, debido a que normalmente no se conocen las clases o la cantidad de grupos de las mismas y que lo más popular es determinar este número mediante *cross validation* y *shrinkage* [32], [33]. Por lo tanto, para plantear una estrategia que genere nuevos productos a partir de los datos históricos se presentan a continuación los siguientes objetivos de la propuesta de investigación.

## 9. OBJETIVOS DEL PROYECTO

### Objetivo General

Desarrollar un método que emplee la minería de datos en la extracción de datos históricos para construir modelos de características.

### Objetivos específicos

1. Documentar los conceptos de líneas de productos, modelos de características, métodos y herramientas de minería de datos mediante una búsqueda sistemática para tener una visibilidad del estado del arte.
2. Adoptar el uso de las técnicas de extracción encontradas en diferentes repositorios de datos (históricos de datos, bases de datos de catálogos de productos, logs de procesos de configuración).
3. Elaborar un modelo de proceso y un modelo de producto para la construcción del modelo de características a partir de los diferentes repositorios de datos.
4. Construir el método para la elaboración de modelos de características a partir de repositorios de datos utilizando técnicas de minería de datos
5. Implementar el método encontrado de minería de datos en la plataforma Variamos.

## 10. IMPACTO Y PRODUCTOS

Al finalizar esta investigación se tiene como producto un desarrollo de software, el cual generará modelos de características a partir de datos históricos; la industria aprovechará sus datos históricos para generar líneas de producto de una forma innovadora, es decir, modelará un diagrama de características usando las técnicas de minería de datos. En el futuro este será un reto con la presencia del Internet de las cosas, y la implementación de Big Data, en donde cada producto tendrá una conexión a Internet (más datos que minar). Se espera que los productos no sean simples objetos con sensores informando sobre los cambios en el medio ambiente, sino que en realidad estos puedan contener el conocimiento y reaccionar

adecuadamente las características dependiendo del contexto de negocio en el que se encuentren. Las líneas de producto le dan a la industria la posibilidad de ver gráficamente todos los posibles productos que se pueden generar. Con la incorporación de un método de minería de datos para la creación de modelos de características el afinamiento de las líneas de producto resultantes al usar este método dará a conocer productos con las características que los consumidores prefieren, y podemos asegurar el impacto positivo en las industrias cuando puedan mejorar sus beneficios al usar los productos derivados de esta investigación.

## 11. METODOLOGÍA

En la Figura 10 se expone brevemente el esquema metodológico adoptado para el desarrollo del proyecto, en el cual se pueden apreciar los principales elementos y ejes que definen el desarrollo metodológico del trabajo.

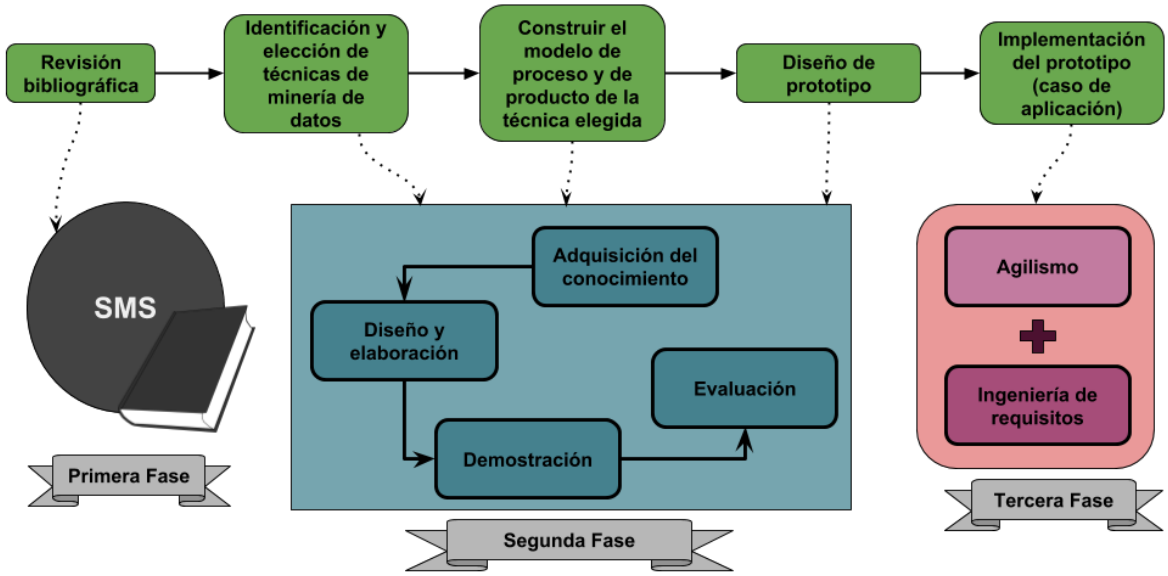


**Figura 10**: Metodología.

**Fase 1:**
Para documentar las técnicas de minería de datos se usará *systematic mapping studies*, al usar estudios sistemáticos (SMS) se reduce el sesgo en la investigación [34], el resultado es un estudio completo, repetible y auditable. El SMS genera un conjunto de documentos caracterizados y categorizados en una gran cantidad de dimensiones y características que consta de 5 etapas:

1. En la primera etapa se definen preguntas de investigación.

2. En la segunda etapa se ejecuta la búsqueda que dará respuesta a las preguntas planteadas anteriormente.
3. La tercera etapa consiste en el escaneo de la bibliografía el cual puede ser manual o automático.
4. La cuarta etapa es la clasificación de estudios encontrados.
5. La última fase consta del mapeo de los datos encontrados que da como resultado un documento llamado protocolo o informe final [35].

Con los hallazgos realizados, se podrán seleccionar más fácilmente las técnicas de minería de datos que se aplican en la elaboración de los modelos de características.

Fase 2:
En esta fase se elaboran los modelos de proceso y de producto aplicando un enfoque o filosofía basado en la metodología *Design Science* [32], posteriormente se Integran los modelos en un método que permita la construcción del modelo de características a partir de los diferentes repositorios de datos. También se realizan demostraciones para mejorar el método.

Fase 3:
Consta de hacer la implementación del método seleccionado, teniendo en cuenta las dificultades por la gran cantidad de datos, el ciclo de desarrollo, incluyendo sus pruebas y ajustes se desarrollará mediante el agilísimo o las metodologías ágiles [36] y las tecnologías en la nube, para un control permanente e incremental. Con esta última fase no se busca hacer reuniones diarias ni desarrollar una planeación semanal, lo que se pretende es realizar un desarrollo incremental basado en objetivos simples y fáciles de alcanzar que se desarrollaran con entregas al tutor o el grupo de investigación y con sus correcciones se realizan los avances incrementales en el desarrollo.
Se tiene como objetivo un desarrollo en código y plataformas abiertas que pueda ser usado por toda la comunidad académica.

## 12. CRONOGRAMA

En la Tabla 1 se presenta el cronograma planteado para el desarrollo de las actividades propuestas para el cumplimiento de los objetivos del proyecto.

| **Actividades/meses** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Realización del estudio sistemático (SMS): se obtiene un protocolo, un documento con la bibliografía categorizada y clasificada. | | | | | | | | | | | | |
| Adquisición del conocimiento: en esta actividad se obtiene el método a usar después de hacer la revisión bibliográfica y los primeros desarrollos o pruebas de los métodos. | | | | | | | | | | | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Diseño y elaboración: Construcción e integración de los modelos de proceso y de producto, en esta actividad se obtiene el método para la producción de modelos de características. | | | | | | | | | | | |
| Demostraciones: para esta actividad se desea probar la funcionalidad del método adoptado se puede tener como objetivo la creación de los modelos de procesos y producto | | | | | | | | | | | |
| Evaluación / implementación: Se desea obtener un prototipo funcional: mediante el agilismo se realizan las pruebas y se hacen mejoras que paulatinamente generan un producto maduro. | | | | | | | | | | | |

**Tabla 1**: Cronograma de actividades

## 13. BIBLIOGRAFÍA


[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, vol. 1. New York, NY: Springer New York, 2009.

[2] M. Nieke, C. Seidl, and S. Schuster, “Guaranteeing Configuration Validity in Evolving Software Product Lines,” in *Proceedings of the Tenth International Workshop on Variability Modelling of Software-intensive Systems - VaMoS ’16*, 2016, pp. 73–80.

[3] J. K. Bae and J. Kim, “Product development with data mining techniques: A case on design of digital camera,” *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9274–9280, Aug. 2011.

[4] D. Liang, C.-F. Tsai, and H.-T. Wu, “The effect of feature selection on financial distress prediction,” *Knowledge-Based Syst.*, vol. 73, pp. 289–297, Jan. 2015.

[5] S. Barak and M. Modarres, “Developing an approach to evaluate stocks by forecasting effective features with data mining methods,” *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1325–1339, Feb. 2015.

[6] M. de T. de la I. y las Comunicaciones, “INFORME RENDICIÓN DE CUENTAS Ministerio de Tecnologías de la Información y las Comunicaciones – Fondo de Tecnologías de la Información y las Comunicaciones,” p. 177, 2014.

[7] D. S. Ardila, “Indicadores Básicos de Tenencia y Uso de Tecnologías de la Información y Comunicación en empresas 2013 Cifras Definitivas,” Bogotá, 2015.

[8] X. Z. Wang and C. McGreavy, “Automatic Classification for Mining Process Operational Data,” *Ind. Eng. Chem. Res.*, vol. 37, no. 6, pp. 2215–2222, Jun. 1998.

[9] Y. Elovici and D. Braha, “A decision-theoretic approach to data mining,” *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 33, no. 1, pp. 42–51, Jan. 2003.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009.

[11] R. Capilla, J. Bosch, and K.-C. Kang, *Systems and Software Variability Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[12] M. Asadi, S. Soltani, D. Gasevic, M. Hatala, and E. Bagheri, “Toward automated feature model configuration with optimizing non-functional requirements,” *Inf. Softw. Technol.*, vol. 56, no. 9, pp. 1144–1165, Sep. 2014.

[13] A. J. Izenman, *Modern Multivariate Statistical Techniques*, vol. 64, no. 9–12. New York, NY: Springer New York, 2008.

[14] M. F. Hornick, E. Marcadé, and S. Venkayala, “Solving Problems in Industry,” in *Java Data Mining*, vol. 87, no. 62, Elsevier, 2007, pp. 25–49.

[15] M. F. Hornick, E. Marcadé, and S. Venkayala, “Data Mining Process,” in *Java Data Mining*, Elsevier, 2007, pp. 51–83.

[16] A. K. Choudhary, J. A. Harding, and M. K. Tiwari, “Data mining in manufacturing: A review based on the kind of knowledge,” *J. Intell. Manuf.*, vol. 20, no. 5, pp. 501–521, 2009.

[17] D. H. Moore, “Classification and regression trees, by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Brooks/Cole Publishing, Monterey, 1984,358 pages, $27.95,” *Cytometry*, vol. 8, no. 5, pp. 534–535, Sep. 1987.

[18] D. Benavides, S. Segura, and A. Ruiz-Cortés, “Automated analysis of feature models 20 years later: A literature review,” *Inf. Syst.*, vol. 35, no. 6, pp. 615–636, Sep. 2010.

[19] T. W. Simpson and J. R. Jiao, *Advances in Product Family and Product Platform Design*. New York, NY: Springer New York, 2014.

[20] M. Jiang, J. Zhang, H. Zhao, and Y. Zhou, “Maintaining software product lines — an industrial practice,” in *2008 IEEE International Conference on Software Maintenance*, 2008, pp. 444–447.

[21] R. Mazo, J. C. Muñoz-Fernández, L. Rincón, C. Salinesi, and G. Tamura, “VariaMos,” in *Proceedings of the 19th International Conference on Software Product Line - SPLC ’15*, 2015, pp. 374–379.

[22] G. Bécan, M. Acher, B. Baudry, and S. Ben Nasr, “Breathing ontological knowledge into feature model synthesis: an empirical study,” *Empir. Softw. Eng.*, vol. 21, no. 4, pp. 1794–1841, Aug. 2016.

[23] C.-Y. Lin and G. E. Okudan, “Planning for multiple-generation product lines using dynamic variable state models with data input from similar products,” *Expert Syst. Appl.*, vol. 40, no. 6, pp. 2013–2022, May 2013.

[24] X. Lian and L. Zhang, “Optimized feature selection towards functional and non-

functional requirements in Software Product Lines," in *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 2015, pp. 191–200.

[25] M. de T. de la I. y las Comunicaciones, "INFORME RENDICIÓN DE CUENTAS Ministerio de Tecnologías de la Información y las Comunicaciones – Fondo de Tecnologías de la Información y las Comunicaciones," p. 177, 2014.

[26] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *J. Retail. Consum. Serv.*, vol. 27, pp. 11–23, Nov. 2015.

[27] T. Thüm, S. Apel, C. Kästner, I. Schaefer, and G. Saake, "A Classification and Survey of Analysis Strategies for Software Product Lines," *ACM Comput. Surv.*, vol. 47, no. 1, pp. 1–45, Jun. 2014.

[28] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[29] P. Li, X. Wu, X. Hu, and H. Wang, "Learning concept-drifting data streams with random ensemble decision trees," *Neurocomputing*, vol. 166, pp. 68–83, Oct. 2015.

[30] G. Ritschard, "CHAID and Earlier Supervised Tree Methods," 1st ed., Université de Genève, Ed. Genève: Département d'économétrie, 2010, p. 30.

[31] Q. Chen, W. Zhu, C. Ju, and W. Zhang, "Cross domain web information extraction with multi-level feature model," in *2014 10th International Conference on Natural Computation (ICNC)*, 2014, pp. 780–784.

[32] V. Vaishnavi and B. Kuechler, "Design Science Research in Information Systems," *2004*, 2013. .

[33] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103, no. 9–12. New York, NY: Springer New York, 2013.

[34] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Inf. Softw. Technol.*, vol. 64, pp. 1–18, Aug. 2015.

[35] T. MAREW, J. KIM, and D. O. O. H. BAE, "Systematic Functional Decomposition in a Product Line Using Aspect-Oriented Software Development:: a Case Study," *Int. J. Softw. Eng. {&} Knowl. Eng.*, vol. 17, no. 1, pp. 33–55, 2007.

[36] D. M. Alaimo, *Proyectos ágiles con #Scrum : flexibilidad, aprendizaje, innovación y colaboración en contextos complejos*, 1st ed. Ciudad Autónoma de Buenos Aires: Kleer, 2013.

## 14. PRESUPUESTO

**Tabla 2: Presupuesto global (en miles de $).**

| RUBROS | FUENTES | | TOTAL |
|---|---|---|---|
| | ***Nombre del Grupo de Investigación*** | ***Otras fuentes: indique cual*** | |
| PERSONAL | **ITOS** | **Propias** | 46272 |
| EQUIPOS | | **Propias** | **2000** |
| MATERIAL BIBLIOGRÁFICO | | | **300** |
| PUBLICACIONES, PATENTES O REGISTROS DE SOFTWARE | | | **4000** |
| SERVICIOS TÉCNICOS | | | **1000** |
| MATERIALES | | | **2000** |
| VIAJES | | | **8000** |
| **TOTAL** | | | **63572** |

**Tabla 3: Descripción de los gastos de personal (en miles de $).**

Descripción de los gastos de personal (en miles de $).

- El valor de la hora de Germán Urrego es 82.000$
- El valor de la hora de José Miguel García es 40.000$

| INVESTIGADOR / ASESOR/ AUXILIAR | FORMACIÓN ACADÉMICA | FUNCIÓN DENTRO DEL PROYECTO | DEDICACIÓN (h/sem) | RECURSOS | | TOTAL |
|---|---|---|---|---|---|---|
| | | | | **Grupo** | **Otras fuentes** | |
| German Urrego | Doctor | Director | 2 | ITOS | Propias | 7872 |
| Jose Garcia | Ingeniero Informático | Estudiante | 20 | ITOS | Propias | 38400 |
| **TOTAL** | | | | | | 46272 |

**Tabla 4: Descripción de los equipos (que se planea adquirir o están en uso) (en miles de $).**

| EQUIPO | JUSTIFICACIÓN | RECURSOS* | | TOTAL |
|---|---|---|---|---|
| | | **Grupo/Facultad** | **Otras Fuentes** | |
| Computadora portátil | Equipo para realizar informes, avances y el desarrollo del proyecto mismo. | | Propias | 2000 |

| | | | |
|---|---|---|---|
| **TOTAL** | | | 2000 |

**Tabla 5: Descripción de los Materiales bibliográficos (que se planea adquirir o están en uso) (en miles de $).**

| ITEM | JUSTIFICACIÓN | RECURSOS* | | TOTAL |
|---|---|---|---|---|
| | | Grupo/Facultad | Otras Fuentes | |
| Libros impresos o accesos a bibliotecas | Con el propósito de tener una base bibliográfica para futuras propuestas | ITOS | Propias | 300 |
| **TOTAL** | | | | 300 |

**Tabla 6: Descripción de las publicaciones patentes o registros de software (que se planea adquirir o están en uso) (en miles de $).**

| ITEM | JUSTIFICACIÓN | RECURSOS* | | TOTAL |
|---|---|---|---|---|
| | | Grupo/Facultad | Otras Fuentes | |
| Trámites para la publicación de un artículo | Con el propósito de publicar un artículo en una revista y lo que implica | ITOS | Propias | 4000 |
| **TOTAL** | | | | 4000 |

**Tabla 7: Descripción de los servicios técnicos (que se planea adquirir o están en uso) (en miles de $).**

| ITEM | JUSTIFICACIÓN | RECURSOS* | | TOTAL |
|---|---|---|---|---|
| | | Grupo/Facultad | Otras Fuentes | |
| Configuración y acceso a servidores | Servicios requeridos para realizar las demostraciones de la bibliografía | ITOS | Propias | 1000 |
| **TOTAL** | | | | 1000 |

**Tabla 8: Descripción de los materiales (que se planea adquirir o están en uso) (en miles de $).**

| ITEM | JUSTIFICACIÓN | RECURSOS* | | TOTAL |
|---|---|---|---|---|
| | | Grupo/Facultad | Otras Fuentes | |
| Materiales para las comunicaciones orales | Con el propósito de realizar charlas sobre el tema | ITOS | Propias | 2000 |

| | | | |
|---|---|---|---|
| **TOTAL** | | | 2000 |

**Tabla 9: Descripción de los viajes (que se planea adquirir o están en uso) (en miles de $).**

| ITEM | JUSTIFICACIÓN | RECURSOS* | | TOTAL |
|---|---|---|---|---|
| | | Grupo/Facultad | Otras Fuentes | |
| Pasantía académica en new york | Con el propósito de crear nuevas relaciones y afianzar la segunda lengua | ITOS | Propias | 8000 |
| **TOTAL** | | | | 8000 |