



DESARROLLO DE MODELOS DE CARACTERÍSTICAS MEDIANTE TÉCNICAS DE MINERÍA DE DATOS

Autor:
Jose Miguel García García

Asesor:
German Arturo Urrego Giraldo

Informe Trabajo de Investigación Grupo de Investigación de Ingeniería y Tecnologías de las
Organizaciones y de la Sociedad –ITOS

Maestría en Ingeniería –Énfasis en Ingeniería de Software
Universidad de Antioquia
2017

Índice general

Lista de figuras	3
Lista de tablas	4
1. Resumen	5
2. Abstract	6
3. Introducción	7
4. Marco Teórico	8
4.1. Minería de Datos	8
4.2. Ingeniería de Líneas de producto	11
4.3. Modelos de características	12
5. Antecedentes	13
5.1. Lecturas recomendadas	14
6. Objetivos	16
6.1. Objetivo General	16
6.1.1. Objetivos específicos	16
7. Productos	17
8. Metodología	18
8.1. Revisión bibliográfica mediante mapeo sistemático de la literatura	19
8.1.1. Planificación de la revisión	20
8.1.2. Estrategia de Búsqueda	23
8.1.3. Escaneo de artículos	25
8.1.4. Análisis de palabras clave	27
8.1.5. Proceso de Extracción de Datos	27
8.2. Diseño del prototipo	27
8.2.1. Adquisición de la información	28
8.2.2. Diseño y elaboración	28
8.2.3. Demostración	28
8.2.4. Evaluación	28
8.3. Implementación del prototipo(Caso de aplicación)	28

9. Resultados y análisis	30
9.1. Revisión bibliográfica mediante mapeo sistemático de la literatura	30
9.1.1. Planificación de la revisión	30
9.1.2. Estrategia de búsqueda	54
9.1.3. Escaneo de artículos	56
9.1.4. Análisis de palabras claves	57
9.1.5. Proceso de extracción	58
9.2. Diseño del prototipo	59
9.3. Implementación del prototipo(Caso de aplicación)	60
10.Conclusiones y Trabajos Futuros	66
Conclusiones y Trabajos Futuros	66
10.1. Conclusiones	66
10.2. Trabajos futuros	66
Bibliografía	67

Índice de figuras

4.1. Proceso Industrial estándar para minar datos.	8
4.2. Modelo de características de ejemplo representando un gato.	12
5.1. Indicadores básicos de tenencia y uso de la información y la comunicación en las empresas Colombianas en 2013.	13
8.1. Metodología adoptada en el proyecto de investigación	18
8.2. Mapeo Sistemático de la Literatura propuesto por Petersen, <i>et al.</i>	20
8.3. Marco de trabajo propuesto en la revisión de documentos.	21
8.4. Estrategia de búsqueda propuesta por Zhang, H. <i>et al.</i> para identificar los estudios pertinentes en ingeniería de software [1].	24
9.1. Distribución de los documentos por tiempo y por tipo. Captura tomada de la aplicación library.	31
9.2. Captura de las sección de palabras claves en la aplicación library.	31
9.3. Captura de la sección donde se muestra la ubicación geográfica de autores según su universidad, tomada en la aplicación library.	35
9.4. Captura de la sección donde se muestra en orden descendente los documentos según su cantidad de citas, tomada en la aplicación library.	36
9.5. Captura tomada de la aplicación <i>Mendely</i> , en donde se presenta la selección de un documento con sus etiquetas.	57
9.6. Captura tomada de la aplicación <i>library</i> , en donde se presentan las formas de filtrar los datos y presentar los reportes personalizados.	58
9.7. Captura tomada de la aplicación <i>Mendely</i> , en donde se presenta la selección de un documento con sus etiquetas.	59
9.8. Modelos generados desde R en el código RandomForest.R.	65

Índice de tablas

8.1. Preguntas que apuntan a obtener las publicaciones con las técnicas de minería de datos en el proceso de ingeniería de líneas de producto.	22
8.2. Preguntas que apuntan a identificar y caracterizar las técnicas de minería de datos usadas en el proceso de la ingeniería de líneas de producto.	22
8.3. Preguntas que apuntan a evaluar las técnicas de minería de datos encontradas, considerando la creación automática de los modelos de características.	23
8.4. Preguntas que apuntan a identificar las mejoras y los enfoques novedosos en las técnicas de minería de datos usadas en la ingeniería de líneas de producto.	23
9.1. Cantidad de documentos encontrados en los diferentes Lugares de publicación a lo largo del tiempo	32
9.2. Resultados de la RQ13: ¿Qué técnicas de minería pueden ser explotadas en la ingeniería de líneas de productos?	50
9.3. Resultados de la búsqueda manual y automática sobre publicaciones relacionadas con las técnicas de Minería de Datos	55
9.4. Resultados del análisis de las variables.	62

Capítulo 1

Resumen

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 [2]. 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[3]. De esta forma, compañías del sector bancario[4], tecnológico [5] y manufacturero [6] han hecho uso de las técnicas de minería de datos para el desarrollo de sus productos mediante la ingeniería de líneas de producto. 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 Claves: Product line engineering, PMML, RIPS, Java, Python scikit-learn.

Capítulo 2

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 [2]. 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[3]. In this way, companies in the banking [4], technology [5] and manufacturing [6] sectors have made use of data mining techniques for the development of their products through product lines engineering. 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: Product line engineering, PMML, RIPS, Java, Phyton scikit-learn.

Capítulo 3

Introducción

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 [7]. 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 [8]. 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 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?

Capítulo 4

Marco Teórico

4.1. Minería de Datos

La minería de datos (MD) es el proceso de encontrar patrones y relaciones en los datos con el fin de realizar actividades descriptivas y predictivas. La MD 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 MD 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 [2]. El modelamiento de los datos mediante las técnicas de MD 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 para ser más competitivas en el mercado. 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 [2]. Por esta razón la MD 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 4.1.

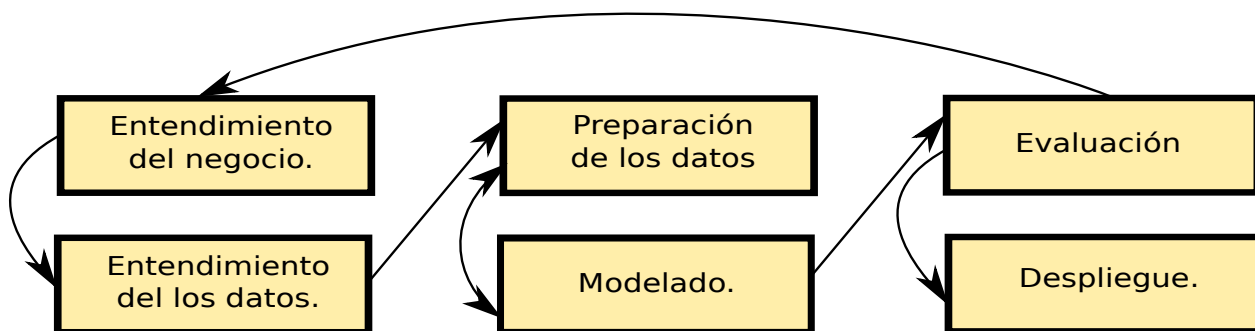


Figura 4.1: Proceso Industrial estándar para minar datos.

La Figura 4.1 ilustra el proceso empleado tradicionalmente en la MD, Cross Industry Standard Process for Data Mining (CRISP-DM)[13]. Es importante diferenciar las técnicas o funciones que se aplican para cada una de las etapas de este proceso:

- **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[13].

- **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[13].
- **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[13].
- **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[13].
- **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[13].
- **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[13].

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]. Los procedimientos que se realizan para analizar estos datos se pueden clasificar en las siguientes técnicas y funciones:

- **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[2]. 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.
- **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[13]. 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[2], por ejemplo, el valor de venta para una casa estará dado por la cantidad de baños multiplicado por su área.
- **Atributo importancia:** En el análisis de negocio es muy importante identificar las características en los datos que influyen el comportamiento de los beneficios económicos[13]. 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.
- **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[13].
- **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[13]. 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-medoids clustering, Hierarchical clustering, Kernel K-means, soft K-means, etc. La diferencia de los algoritmos radica en la función de distancia usada[2].

- **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[10].
- **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[13].
- **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[2].
- **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[9].
- **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[2, 10].
- **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[10]. 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.

Es importante considerar en esta sección que 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[14].

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 y los bosques aleatorios de Leo Breiman. 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 [2, 10]. 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 y bajo este enfoque se desarrolla *Random Forest* o árboles aleatorios[2, 10]. La formación de las áreas a partir de un conjunto de datos está dada por los valores X_1 y X_2 , se puede observar que la Ecuación 4.1 representa los futuros puntos que se ubicarán en las k regiones[2]. Esta es la representación de un árbol de regresión con k 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.

$$\hat{f}(X) = \sum_{m=1}^k c_m I\{(X_1, X_2) \in R_m\} \quad (4.1)$$

En la Ecuación 4.1 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[2], 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[2]. Por otra parte, Las líneas de producto también tienen un modelo muy similar a la Ecuación 4.1, 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.

4.2. Ingeniería de Líneas de producto

La ingeniería de líneas de producto (ILP) tiene como objetivo la producción de conjuntos de productos con más características comunes que diferentes, estas líneas de producto (LP) se han convertido en un paradigma viable para mejorar la productividad y la calidad de la producción en masa[15]. 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 LP, 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 LP [16–18].

Las LP protagonizan la etapa de diseño de los nuevos productos que consumen los clientes y la MD 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[5, 6]. La MD 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?)[6, 13]. 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[15]. 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[19]. Estos productos pueden ser software también, incorporar LP en el ciclo de vida de desarrollo de software (CVDS) mejora las estadísticas de mercadeo ampliando los beneficios de dos a siete veces[17].

4.3. Modelos de características

Un modelo de características (MC) representa la información de todos los posibles productos en una LP en términos de las características y las relaciones entre ellas [?]. En la Figura 4.2 se elaboro un ejmplo sencillo para mostrar la representación de las características de un gato sin embargo recomendamos a diferentes autores y sus representaciones[20–23].

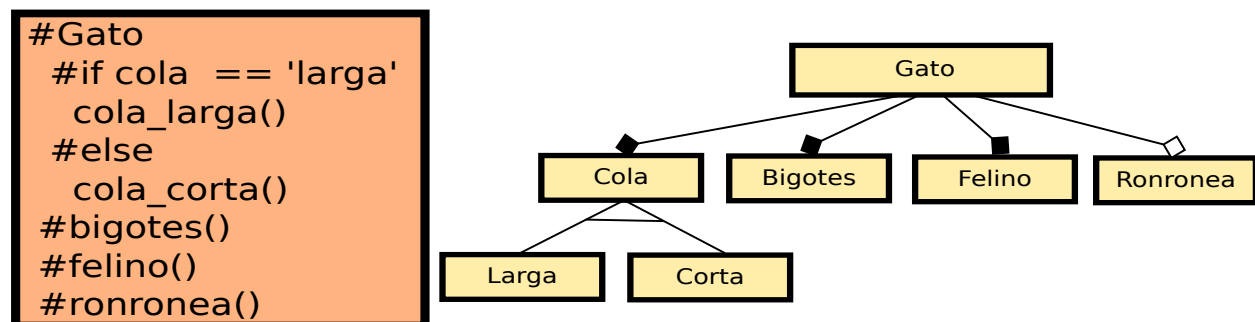


Figura 4.2: Modelo de características de ejemplo representando un gato.

En la literatura anterior se agrupan las características y sus relaciones de forma jerárquica, las relaciones pueden ser Obligatorio, Opcional, Alternativo, O, inclusión y exclusión[15].

- **Obligatoria:** Esto significa que el hijo asignado tiene que estar incluido en el producto.
- **Opcional:** Esto significa que el hijo asignado puede o no tener esta característica.
- **Alternativo:** Da a conocer que solo una característica en la jerarquía puede ser seleccionada.
- **O:** Esta relación significa que se puede seleccionar todas o ninguna de las características en la jerarquía. Requerida: Cuando una característica requiere a otra, esta no puede existir sin la presencia de la otra.
- **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 RECoVar[17, 24], VariaMos[25], la cual tiene como objetivo desarrollar familias de sistemas y también tiene herramientas para realizar operaciones sobre otros modelos. SPLOT (software product line online tool)[26], 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.

Capítulo 5

Antecedentes

En la actualidad las empresas en Colombia usan las computadoras y el Internet más que antes, se estima que de 8659 empresas el 99 % posee computador y está conectada a Internet [8]. En la Figura 5.1 se muestran los indicadores básicos de tenencia y uso de la información y la comunicación en las empresas.

Indicadores Básicos de Tenencia y Uso de Tecnologías de la Información y Comunicación en empresas

2013 Cifras Definitivas

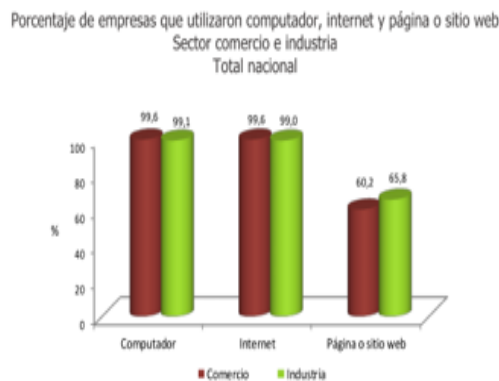


Figura 5.1: Indicadores básicos de tenencia y uso de la información y la comunicación en las empresas Colombianas en 2013.

El Ministerio de Tecnologías de la información y las comunicaciones ha invertido hasta \$373993 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].

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. Desde los años setenta los computadores se usaron en el análisis exploratorio de los datos y en las décadas posteriores fuimos testigos de un avance en el procesamiento y el almacenamiento de las computadoras, lo que permitió que grandes cantidades de datos fueran clasificados, almacenados y administrados de forma

eficiente por 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. Actualmente, el análisis computacional de los datos está ganando popularidad. Las enormes cantidades de datos son cada vez más comunes y aunque las personas encargadas de analizarlos siguen teniendo en cuenta las técnicas supervisadas, el descubrimiento de información no supervisado es 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 [10].

En este documento inicialmente se exploran varios ejemplos del uso de la minería de datos en las líneas de producto con el fin de evidenciar la pertinencia de este estudio, luego en los resultados se presenta el desarrollo del protocolo de la revisión sistemática de la literatura y se hace énfasis en nuestro ejemplo. En este apartado, Se partirá de un ambiente heterogéneo, donde existen diferentes compañías que desarrollan productos masivamente y pueden usar muchas técnicas de análisis de datos en común [27], como funciones descriptivas, predictivas y técnicas asociativas.

5.1. Lecturas recomendadas

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 [4].
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 [5].
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 [28]. 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 [6].

Teniendo en cuenta el desarrollo de los modelos de características, en la minería de datos pueden encontrarse características que están relacionadas de forma interesante y pueden ser modeladas mediante árboles de decisión [29] y reglas de asociación [30]. Además, se sabe que estas características pueden estar asociadas a un beneficio económico [31], 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 [5]. En los ejemplos anteriores los datos que son sometidos a un proceso de evaluación, generan un patrón de decisión que elimina la redundancia en la información. 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 [32].

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 [33] y shrinkage [16]. Por lo tanto, para plantear una estrategia que genere nuevos productos a partir de los datos históricos se presentan a continuación en el siguiente capítulo los siguientes objetivos de este proyecto de investigación.

Capítulo 6

Objetivos

6.1. Objetivo General

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

6.1.1. 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.

Capítulo 7

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, las aplicaciones que usan Big Data como fuente de datos y la Internet de las cosas, 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 dependiendo del contexto de negocio en el que se encuentren. Los modelos de características 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, el consumo y el medio ambiente cuando se puedan usar los productos derivados de esta investigación. En este proyecto se destaca que cada objetivo específico tiene como propósito generar un producto, el primer objetivo específico da como resultado el desarrollo del mapeo sistemático de la literatura (SMS), el segundo objetivo específico produce una lista de los métodos y herramientas candidatas dentro de la ciencia de los datos, haciendo énfasis en la minería de datos con los algoritmos de clasificación y agrupamiento, el tercer objetivo específico obtenemos un modelo de proceso y un modelo de producto, que muestre los elementos implicados y el flujo de trabajo que se debe seguir para convertir los datos en modelos de características, en el cuarto objetivo se desarrolla un método, este involucra el producto del objetivo anterior para generar un método que especifique las tecnologías y las actividades técnicas a desarrollar para la elaboración de los modelos de características, el quinto objetivo específico desarrollamos un ejemplo utilizando las tecnologías encontradas en el objetivo cuatro. Los productos en este proyecto describen los pasos necesarios para desarrollar el objetivo general.

Capítulo 8

Metodología

En la Figura 8.1 se presenta el esquema metodológico adoptado para la ejecución del proyecto, en el cual se pueden apreciar los principales elementos, ejes y fases que definen el desarrollo del proyecto de investigación.

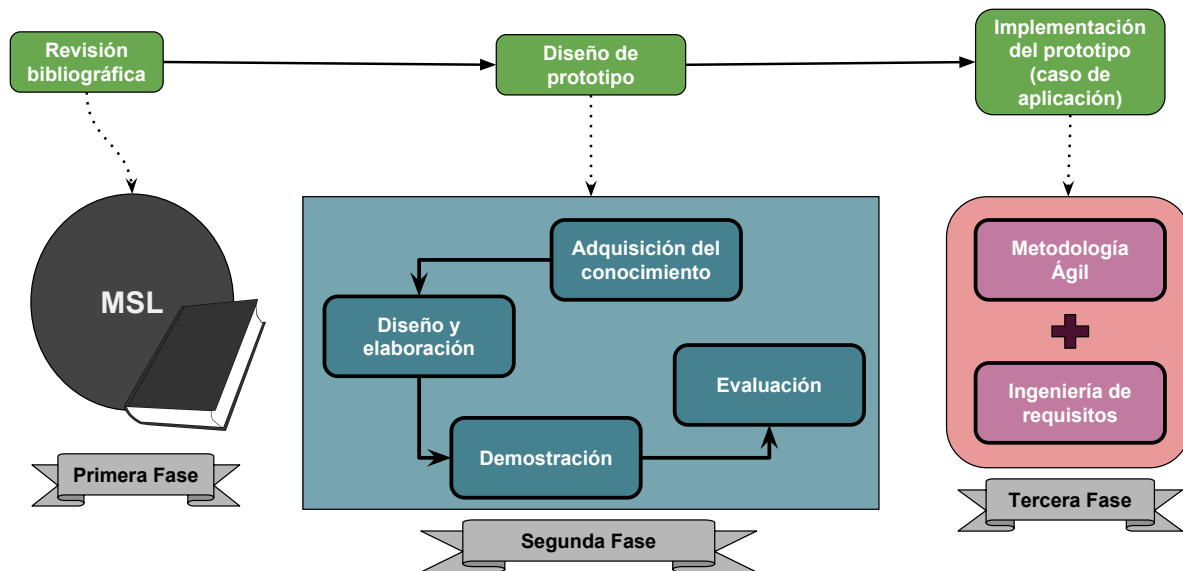


Figura 8.1: Metodología adoptada en el proyecto de investigación

- **Primera fase:** Con el fin de documentar las técnicas de minería de datos (MD) empleadas en la ingeniería de líneas de producto (ILP), se realizó una revisión de la literatura mediante la metodología *Systematic Mapping Studies* (SMS por sus siglas en inglés) o mapeo sistemático de la literatura (MSL), la cual consistió en categorizar y caracterizar un conjunto de documentos en diferentes dimensiones, mediante el desarrollo de las etapas presentadas a continuación. Es de anotar que el proceso de MSL consiste en las cinco etapas posteriores a la planeación, ya que esta depende de cada investigador.

1. En la primera etapa se realiza la planeación de la revisión.
2. En la segunda etapa se definieron las preguntas de investigación.

3. En la tercera etapa se ejecuto la búsqueda que daría respuesta a las preguntas planteadas anteriormente.
 4. La cuarta etapa consistió en el escaneo de la bibliografía, el cual puede ser manual o automático.
 5. La quinta etapa fue la clasificación de los estudios encontrados.
 6. La última etapa consistió en el mapeo de los datos encontrados, lo cual derivó en un documento llamado protocolo o informe final.
- **Segunda fase:** En esta fase se construyeron los modelos de proceso y de producto aplicando un enfoque o filosofía basado en la metodología Design Science [34] haciendo énfasis en el entendimiento de las diferentes capas en los sistemas de información. Posteriormente estos modelos fueron probados con el repositorios de datos Iris, con el propósito de construir un MC y, a partir de esta demostración, mejorar el método empleado.
 - **Tercera fase:** En esta fase se dejó constancia de la implementación del método seleccionado, teniendo en cuenta la dificultad que se presentó por la gran cantidad de datos. En las etapas de implementación y pruebas del ciclo de desarrollo del software se incluyeron el agilismo, también llamado metodologías ágiles [35], y las tecnologías en la nube para un control permanente y continuo. El propósito de estas etapas consistió en un desarrollo incremental basado en objetivos simples y fáciles de alcanzar, cuyo control de cumplimiento se desarrolló mediante entregas al tutor o al grupo de investigación y, con sus correcciones, se realizaron paulatinamente los avances en el desarrollo. El objetivo de esta fase fue un desarrollo de un código en plataformas abiertas que pueda ser usado por toda la comunidad académica.

8.1. Revisión bibliográfica mediante mapeo sistemático de la literatura

En esta sección se aplicaron las guías propuestas por Petersen, *et al.* [36] para el desarrollo del MSL. Sin embargo, a este proceso de mapeo que originalmente posee cinco etapas, le fue añadida una etapa previa de planeación de la revisión, como se muestra en la Figura 8.2, donde puede observarse también que cada etapa posee un producto derivado, el cual se mostrará en el desarrollo del documento.

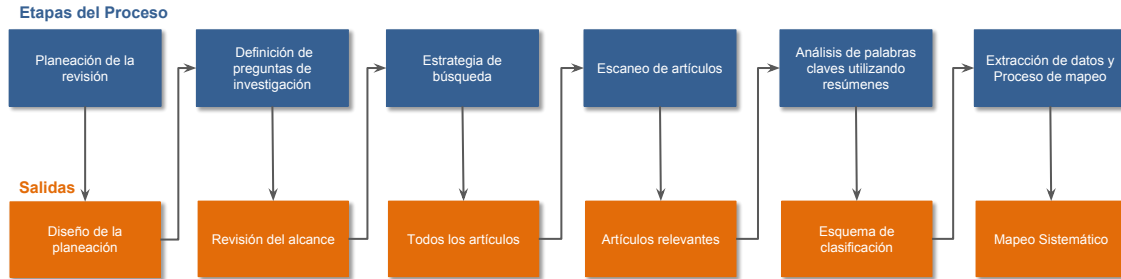


Figura 8.2: Mapeo Sistemático de la Literatura propuesto por Petersen, *et al.*

8.1.1. Planificación de la revisión

El proceso del MSL en su primera etapa define las preguntas de investigación teniendo en cuenta la planificación de la revisión. Para dar una idea del alcance y las metas del presente estudio, se propuso un método empírico usando el mismo marco de trabajo de Rolland, C. *et al.* [37] y Wieringa, R. *et al.* [38]. El estudio fue estructurado en siete conceptos claves o aspectos que giran entorno a la palabra clave: Minería de datos. El marco de trabajo se ilustra en la Figura 8.3 y es el insumo para realizar las preguntas de investigación que se desarrollaron por medio del MSL.

Aspecto 1. Definiciones sobre minería de datos: La MD es el proceso mediante el cual, haciendo uso de diferentes técnicas, se pretende encontrar patrones y relaciones en grandes cantidades de datos[2]; este aspecto, sin embargo, se centró en la definición de las técnicas de MD en el proceso de ILP que pudieran ser usadas para crear modelos de características.

Aspecto 2. Clasificación de las técnicas y métodos de la minería de datos: En este aspecto se realizaron definiciones de algunos criterios ilustrados por Hastie, T. *et al.* [10] y Caruana R. *et al.* [39] para comparar y extraer la información sobre las técnicas más apropiadas para desarrollar este estudio.

Aspecto 3. Ciclo de vida de la ingeniería de líneas de producto: En este aspecto se abordó el tema del ciclo de vida de la ILP como un proceso cuyo objetivo es crear grupos de productos que compartan ciertas características que logren satisfacer determinados nichos de mercado [40], teniendo en consideración la creación de productos personalizados de forma masiva.

Aspecto 4. Soporte automatizado: Dentro de la ILP es muy común representar un producto en la notación de MC teniendo en cuenta determinadas características y relaciones [15]. El propósito de este aspecto fue evidenciar la pertinencia de los MC dentro de los modelos de variabilidad, explicando el desarrollo incremental de los mismos y su mantenimiento automatizado a través del tiempo.

Aspecto 5. Acercamientos novedosos En este aspecto se identificaron los métodos más utilizados y los trabajos más novedosos en el área de la MD.

Aspecto 6. Técnicas de evaluación: Debido a que las técnicas de MD son evaluadas teniendo en cuenta su ajuste a los datos de prueba, este aspecto propuso identificar y exponer los métodos de evaluación de las técnicas de MD considerando el tipo de datos, el contexto y el dominio. Con el propósito de adoptar un método de evaluación.

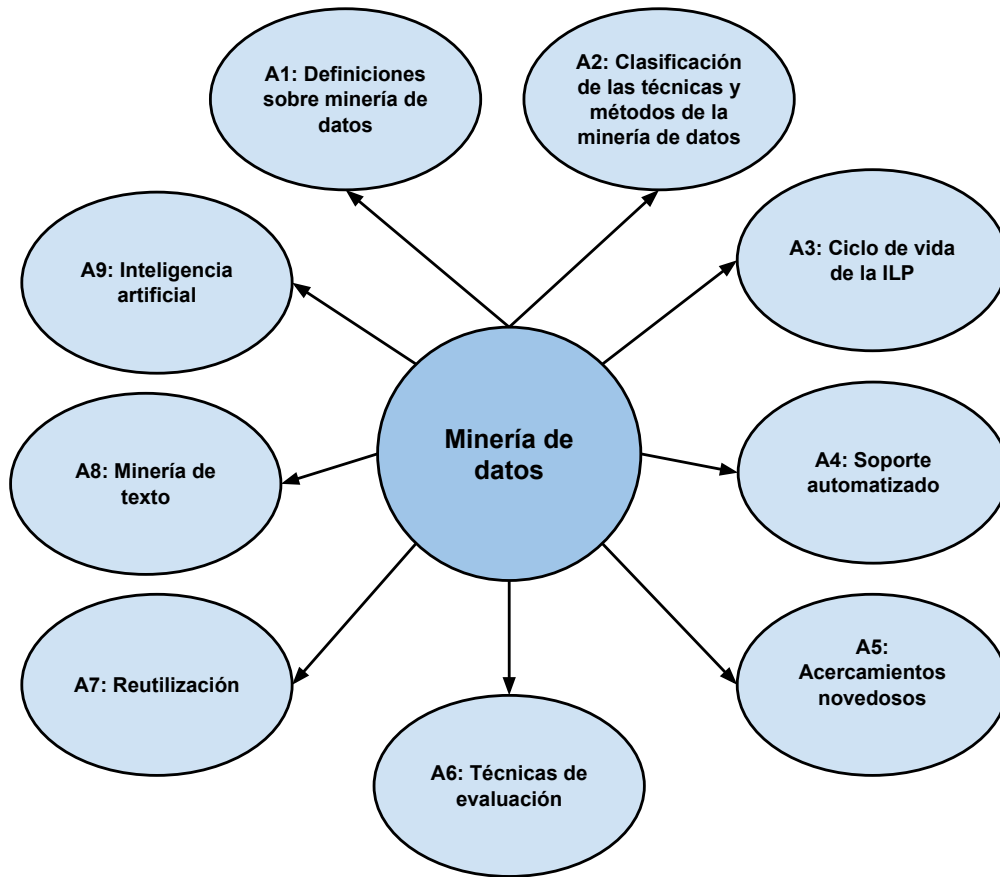


Figura 8.3: Marco de trabajo propuesto en la revisión de documentos.

Aspecto 7. Reutilización: En este aspecto se identificaron los elementos reusables de los códigos utilizados en los diferentes enfoques de la MD, los conceptos involucrados y sus productos.

Aspecto 8. Minería de texto: Este aspecto hizo hincapié en las herramientas o las diferentes técnicas de agrupamiento y su aplicación en grandes cantidades de datos. Desde el punto de vista de la construcción de MC, el interés consistió en revisar todas aquellas técnicas de agrupamiento que usaran árboles como estructuras jerárquicas.

Aspecto 9. Inteligencia artificial: Algunas técnicas de la MD son no supervisadas y bordean el tema de la inteligencia artificial. Con el propósito de cubrir la mayoría de documentos encontrados en el MSL, este aspecto se encargó de incluir el tema de la inteligencia artificial dentro del estudio.

La meta principal de esta sección fue establecer las preguntas de investigación que permitieran la clasificación de la literatura respecto a las técnicas de MD y la ILP utilizados en cualquier contexto. Para esto se establecieron preguntas basadas en metas más pequeñas que pretendieron orientar la búsqueda hacia las publicaciones que específicamente usaron las técnicas de MD en el proceso de ILP .

Las metas en las que se basó la búsqueda se enumeran a continuación:

Meta 1

Obtener publicaciones con información sobre las técnicas de MD en el proceso de ILP. Las preguntas propuestas para el cumplimiento de esta meta se muestran en la Tabla 8.1.

Tabla 8.1: Preguntas que apuntan a obtener las publicaciones con las técnicas de minería de datos en el proceso de ingeniería de líneas de producto.

ID	Pregunta de investigación	Motivación
RQ1	¿Cómo están distribuidos los estudios en el tiempo?	Representar la tendencia que ha tenido el trabajo de investigación en el tiempo.
RQ2	¿Qué tipo de documentos hablan del papel de la MD en la ILP?	Para los nuevos investigadores es interesante conocer cuáles son las revistas, conferencias, talleres o títulos de publicación que son protagonistas y abarcan la mayoría de los documentos.
RQ3	¿Qué distribución geográfica o cuáles autores son los más representativos en el presente estudio?	Con el propósito de considerar los grupos de investigación y las regiones interesantes para construir alianzas estratégicas y futuros trabajos.
RQ4	¿Qué categoría o escalafón de investigación tienen asignados los documentos encontrados en nuestro estudio?	Conociendo la categoría o el escalafón se puede establecer un grado de madurez en el tema de investigación como lo propone Wieringa, R. <i>et al.</i> [38].

Meta 2

Identificar y caracterizar las técnicas de MD utilizadas en el proceso de ILP. Las preguntas propuestas para el cumplimiento de esta meta se muestran en la Tabla 8.2.

Tabla 8.2: Preguntas que apuntan a identificar y caracterizar las técnicas de minería de datos usadas en el proceso de la ingeniería de líneas de producto.

ID	Pregunta de investigación	Motivación
RQ5	¿Qué algoritmos o técnicas son usadas en le ciclo de vida de la ILP para crear MC?	Identificar las técnicas apropiadas y los algoritmos para crear MC.
RQ6	¿Qué técnicas de la MD son usadas en el campo de la ILP?	Entender para qué se esta usando la MD en ILP con el fin de tener un entendimiento de su uso.
RQ7	¿Qué contexto de dominio está implementando MD e ILP?	Con el propósito de considerar las empresas y los sectores de la industria que pueden ser interesantes en la construcción de alianzas estratégicas y futuros trabajos.
RQ8	¿Cómo están definidas las técnicas de MD encontradas en los documentos?	Es importante la identificación de los procedimientos y los cálculos que se utilizan en las técnicas de MD.
RQ9	¿En cuál etapa del ciclo de vida de la ILP se usa la MD?	Asignar las técnicas encontradas a una etapa dentro del ciclo de vida de la ILP puede ser un marco de trabajo adecuado.

Continúa en la siguiente página

Tabla 8.2 – *Continuación de la página anterior*

ID	Pregunta de investigación	Motivación
RQ10	¿Qué tipo de productos se han derivado del uso de las técnicas de MD en la ILP?	Es importante conocer qué se obtiene después de aplicar las técnicas de MD en las diferentes etapas del ciclo de vida de la ILP.

Meta 3

Evaluar las técnicas de MD encontradas considerando la automatización de los MC. Las preguntas propuestas para el cumplimiento de esta meta se muestran en la Tabla 8.3.

Tabla 8.3: Preguntas que apuntan a evaluar las técnicas de minería de datos encontradas, considerando la creación automática de los modelos de características.

ID	Pregunta de investigación	Motivación
RQ11	¿Cómo se puede validar el uso correcto de las técnicas de MD en ILP?	Con los enfoques de MD que propone Hastie, T. <i>et al.</i> [10] se puede considerar que las técnicas de MD están siendo aplicadas o usadas para el propósito que fueron diseñadas.
RQ12	¿Los estudios encontrados en la literatura apoyan la automatización de técnicas de MD?	Caracterizar la aplicabilidad de las técnicas de MD considerando su replicabilidad y automatización.

Meta 4

Identificar enfoques de mejoras y nuevos enfoques en técnicas de MD utilizadas en la ILP. Las preguntas propuestas para el cumplimiento de esta meta se muestran en la Tabla 8.4.

Tabla 8.4: Preguntas que apuntan a identificar las mejoras y los enfoques novedosos en las técnicas de minería de datos usadas en la ingeniería de líneas de producto.

ID	Pregunta de investigación	Motivación
RQ13	¿Qué técnicas de MD pueden ser explotadas en la ILP?	Identificar las técnicas que son más importantes en el tema de interés.
RQ14	¿Entre los documentos encontrados son evidentes las propuestas innovadoras en la ILP?	Como una percepción subjetiva, realmente es enriquecedor para la investigación futura saber cuáles son las tendencias más atractivas.

8.1.2. Estrategia de Búsqueda

En el presente MSL como paso inicial se hizo uso de la búsqueda manual y la búsqueda automática, acompañada de un proceso de bola de nieve incremental desde adelante hacia atrás (*backwards snowballing process*)[41], en el cual se miraron los títulos de los documentos, luego los resúmenes y finalmente, para completar el proceso de búsqueda, se buscaron documentos similares en las referencias y en los lugares de publicación.

El estándar *Quasi-Gold*

Zhang, H. *et al.* [1] propuso un enfoque para identificar los estudios pertinentes en la ingeniería de software mediante una revisión sistemática de la literatura, dicho enfoque está ilustrado en la Figura 8.4.

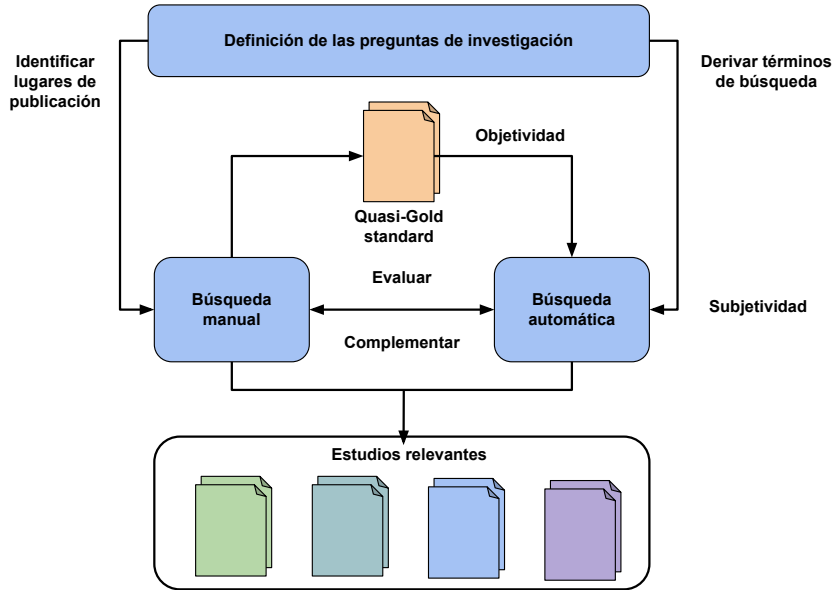


Figura 8.4: Estrategia de búsqueda propuesta por Zhang, H. *et al.* para identificar los estudios pertinentes en ingeniería de software [1].

Los anteriores conceptos fueron tomados en consideración teniendo en cuenta que un *gold standard* representa, con la mayor exactitud posible, el conjunto conocido de los estudios primarios identificados en una colección de acuerdo con las preguntas de investigación propuestas en una revisión sistemática de literatura. De esta forma, una estrategia de búsqueda perfecta capturaría, exactamente, el *gold standard* sin ningún resultado irrelevante. En consecuencia con lo anterior, Zhang, H. *et al.* introduce el término *quasi-gold standard (QGS)* para obtener e identificar los estudios pertinentes en la ingeniería de software. Un QGS es un conjunto de estudios conocidos que son encontrados con una búsqueda manual en fuentes relacionadas (donde fueron publicados) durante un periodo de tiempo determinado. El QGS puede considerarse como un *gold standard* en las condiciones en que estas restricciones (fuentes y tiempo) son aplicadas.

Los QGS son útiles para hacer un método más objetivo con el fin de planificar y probar cadenas de búsqueda manuales y automáticas. La idea principal es evitar perder estudios relevantes en la estrategia de búsqueda, por tanto los resultados de las búsquedas manuales y automáticas se comparan con el QGS para calcular la sensibilidad y precisión de los resultados de la búsqueda. Es deseable obtener una alta sensibilidad en lugar de una alta precisión, dado que, una alta sensibilidad significa que se abarcan mas estudios. La sensibilidad por encima del 80 % es óptima para los mejores resultados, pero los resultados entre el 70 % y el 80 % también son aceptables según Zhang, H. *et al.* [1]. Adicional al QGS, se utilizó otro método para ampliar la cobertura de la búsqueda y encontrar otros estudios pertinentes, este método se llama *snowballing* [41], y consistió en utilizar la lista de referencias de un documento o las citas del documento para identificar estudios adicionales sobre el tema. Hay dos maneras de hacer *snowballing*: *snowballing* hacia atrás y *snowballing* hacia adelante. En el presente proyecto se realizó *snowballing* hacia atrás, lo cual consistió en utilizar la lista de

referencias de los trabajos que ya habían sido seleccionado para obtener otros estudios relevantes sobre el tema.

Criterios de exclusión e inclusión

Con el fin de realizar una cuidadosa selección de los estudios pertinentes, los documentos obtenidos deben ser verificados, primero, con criterios de exclusión; si un documento cumple con alguno de los criterios de exclusión, entonces será excluido. Los trabajos restantes se verifican con todos los criterios de inclusión para que sean aceptados.

Criterios de Exclusión:

- **CE1:** Documentos como reportes técnicos, lecciones aprendidas, borradores, estudios que describen eventos, afiches o pósteres y trabajos sin publicar que presenten MC.
- **CE2:** Documentos como artículos de opinión (*Opinion articles or position papers*).
- **CE3:** Documentos que no expliquen el uso de los métodos propuestos por la MD.
- **CE4:** Documentos que presenten MC, pero que no estén aplicados a la ILP.

Criterios de Inclusión:

- **CI1:** Documentos (artículos, actas o memorias en conferencias, capítulos de libros o reportes de investigación) que presenten MC.
- **CI2:** Documentos que presenten MC en el proceso de ILP.
- **CI3:** Documentos que fueron publicados preferiblemente durante la última década del siglo XXI.
- **CI4:** Documentos que fueron publicados por profesores o personajes destacados.

8.1.3. Escaneo de artículos

Para seleccionar los trabajos pertinentes, se siguió el proceso descrito a continuación, el cual consta de tres etapas:

- Se realizó una búsqueda manual, utilizando los lugares de publicación que se enumeran . La búsqueda manual es útil para mejorar las cadenas de búsqueda utilizadas en la búsqueda automatizada y el QGS.
- Se realizó una búsqueda automatizada, mediante el uso de los motores de búsqueda enumerados en la sección siguiente sección.
- Finalmente, el proceso de *Snowballing* en cada una de las primeras etapas, se realizó en tres rondas:
 - Selección de trabajos según su título y palabras clave, utilizando los criterios de inclusión.
 - Selección de trabajos de acuerdo a su resumen de los trabajos seleccionados en la primera ronda. Aquí se utilizaron tanto los criterios de inclusión como los criterios de exclusión.
 - Selección del conjunto definitivo de trabajos revisando su texto completo utilizando todos los criterios de inclusión y exclusión.

Tratamiento a los Documentos Repetidos

Un documento duplicado es uno que se recupera de varias fuentes de búsqueda (es decir, bibliotecas digitales y lugares de publicación) por lo que se tiene más de una copia de la misma. El conjunto final de estudios pertinentes no debe incluir todas las copias de un documento duplicado. En consecuencia, los documentos duplicados se purgarán de modo que las copias adicionales se excluyan comparando la búsqueda manual y la búsqueda automática. Se llaman estudios repetidos a los artículos sobre el mismo estudio que se publican en varios lugares. Los estudios repetidos pueden contener los mismos autores, una combinación de los nombres del autor o la lista de autores con algunas variaciones.

Búsqueda Manual

Antes de hacer una búsqueda automatizada, debe realizarse una búsqueda manual. La búsqueda manual se utiliza para identificar los estudios pertinentes. También contribuye a la elaboración del *QGS* [1]. Este proceso consiste en una navegación manual de los lugares de publicación relevantes presentados a continuación. Antes de presentar la búsqueda automática y los sitios de búsqueda, estas dos búsquedas se apoyan en estudios que bien pueden ser considerados como *QGS* o contribuyen a la elaboración de este proceso, por lo tanto, se seleccionaron los siguientes documentos: [20, 42–49] y se considera usar sus lugares de publicación y el de sus referencias para este trabajo de investigación.

Artículos

1. *Information Sciences*
2. *IEEE Transactions on Software Engineering*
3. *Empirical Software Engineering*
4. *Applied Soft Computing*
5. *ACM Computing Surveys*

Conferencias, Talleres y Memorias

1. *International Software Product Line Conference*
2. *International Conference on Information Technology*
3. *International Workshop on Product Line Approaches in Software Engineering (PLEASE)*
4. *Proceedings of the International Conference on Software Product Line - SPLC*
5. *Proceedings of the IEEE*
6. *Proceedings of the Annual ACM Symposium on Applied Computing - SAC*
7. *Proceedings of the international conference on Machine learning - ICML*
8. *Proceedings of the International Conference on Software and System Process - ICSSP*

9. *Proceedings of the International Symposium on Computers in Education - SIIE*
10. *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering - EASE*
11. *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining - KDD*
12. *Proceedings of the International Conference on Computer Systems and Technologies - CompSysTech*
13. *Proceedings of the Euromicro Conference Series on Software Engineering and Advanced Applications - SEAA*

Búsqueda Automática

La búsqueda automática se basa en obtener estudios relevantes usando las librerías digitales, esto se realiza de acuerdo a unas cadenas de búsqueda que se adaptan de acuerdo al motor que usa cada una de las librerías digitales. Estas búsquedas son más efectivas que las manuales pero el rendimiento depende de la calidad de la cadena de búsqueda, la capacidad de los motores de búsqueda y la diversidad o la popularidad de los documentos pertenecientes a la materia. Para mejorar la búsqueda automática, se hizo uso de las bibliotecas digitales y los índices, que proveen un gran repositorio de documentos y valiosas fuentes de publicación. Para la búsqueda automática seleccionaremos dos bases de datos digitales *Scopus* y *Web of Science*. *Scopus* es largamente usada por que contiene indexadas la mayoría de las editoriales importantes como *Elsevier*, *IEEE*, *Springer*, *Wiley-Blackwell*, entre otras. *Web of Science* también es muy útil por *Inspec*. La búsqueda automática se completó con los resultados obtenidos en *Google Scholar*, *Nature* y *Science*.

8.1.4. Análisis de palabras clave

En este trabajo de investigación, sólo se definió la siguiente cadena de búsqueda basada en los estudios recomendados, las facetas y el consejo del investigador principal, y se examinaron los resultados considerando todos los criterios de inclusión y el proceso de *Snowballing and Screening of papers*. (*“product line*” OR “product famil*” OR variability OR “product platform*”*) *AND* (*“data mining” OR “Artificial Intelligence” OR “clustering”*) *AND* (*“feature*” OR variabili* OR characteristic**)

8.1.5. Proceso de Extracción de Datos

Para la extracción de datos se usó la siguiente tecnología: *Excel and Google Sheets*, *Splunk Enterprise* & *Mendeley Desktop* teniendo en cuenta la información asociada con cada meta del proyecto de investigación.

8.2. Diseño del prototipo

La ciencia puede ser considera como una actividad que contribuye al entendimiento de un fenómeno. El fenómeno en este proyecto de investigación parte de una serie de comportamientos que pueden entenderse como unos patrones y la posibilidad de clasificarlos, Con el objetivo de producir un artefacto(software) que no existía se consideró *Design Science* [34] como una estrategia para el desarrollo de la investigación, una publicación o una patente. El prototipo de esta investigación

partirá de la construcción de los modelos de características considerando los datos públicos de Iris y luego los de RIPS. En este proyecto de investigación se siguió la metodología propuesta por [34] y se resumió el proceso en: adquisición de la información, diseño y elaboración, demostración y evaluación.

8.2.1. Adquisición de la información

Desde La Facultad Nacional de Salud Pública “Héctor Abad Gómez” de la Universidad de Antioquia nos fue entregada una gran colección de datos pertenecientes a Los Registros Individuales de Prestación de Servicios de Salud (RIPS), lo primero que se realizó fue guardar esta información y realizar un respaldo, posteriormente se seleccionaron los primeros 66262 datos correspondientes al 0,1%. Debido a que no se contó con los recursos computacionales suficientes para realizar el procesamiento sobre la totalidad de los datos. En esta etapa se llevó a cabo una revisión preliminar de la información y se establecieron los posibles factores y variables que pudieran ser estadísticamente significativos en la construcción de los modelos de características.

8.2.2. Diseño y elaboración

Para la construcción de los modelos y los diseños se realizaron entregas en Gráficos vectorizados Redimensionables *Scalable Vector Graphics (SVG)* exportados al Formato de Documentos Portable *Portable Document Format (PDF)* mediante *InkScape*. En cada entrega se iniciaba una discusión sobre los supuestos en el diseño del método, y debido a que era una etapa temprana del proyecto se buscaron resultados conceptuales que se ajustaran a los mejores paradigmas encontrados en la literatura. Además se realizaron preguntas sobre ¿qué tan efectivo es el modelo planteado?, ¿Soporta operaciones de complejidad elevada?, ¿qué operaciones soporta?, ¿es escalable?, ¿Cuáles son las salidas y entradas en cada una de las etapas del modelo? En cada sesión se afinó el diseño hasta tener los requisitos para el desarrollo.

8.2.3. Demostración

Esta etapa de la metodología llamada *Design Science*, se basó en demostraciones simples de las tecnologías, para tener pequeños artefactos, como en nuestro caso *Code Snippets* para el desarrollo conceptual del modelo. Su uso utilizó el *Hortonworks sandbox*, los datos RIPS y los datos de Iris, y con esta forma de trabajar se establecieron reglas, restricciones existentes en la tecnología y se definieron cuáles partes del diseño era necesario afinar.

8.2.4. Evaluación

Para esta etapa en el *Design Science* se hicieron una serie de pruebas de cada parte del modelo propuesto, estas pequeñas pruebas llamadas *micro-evaluations* establecieron las pruebas formales al modelo después de tener una versión estable. En esta etapa también se establecieron los criterios de evaluación, en nuestro caso probar el ajuste de los datos y realizar el contraste entre los modelos de características al final del proceso planteado en el modelo.

8.3. Implementación del prototipo(Caso de aplicación)

Como primera etapa de este desarrollo se tuvo el reconocimiento de los datos. En donde La Facultad Nacional de Salud Pública “Héctor Abad Gómez” puso a disposición del Grupo de Investigación de Ingeniería y Tecnologías de las Organizaciones y de la Sociedad (ITOS) los Registros

Individuales de Prestación de Servicios de Salud (RIPS). Luego se descubrieron las características técnicas indispensables para la adquisición de la información dado el peso de los datos, el siguiente paso fue el preprocesamiento de los datos, para el cual fue indispensable que el dueño del negocio (la persona que tiene conocimiento del contexto y el dominio) acompañe y valide los resultados, para efectos del desarrollo inicial se tomaron aproximadamente los primeros 600 mil registros de la base de datos (0.01 %) mediante el comando *split* en *macOS Sierra*. En el preprocesamiento de los datos se desarrolla un *script* en *python* (*numpy*, *pandas*, *sklearn*). En este *Python Script* se tuvieron como insumo inicial los datos que el cliente obtiene tras hacer un vaciado a archivo desde la base de datos que fue facilitada y, su posterior división para efectos del caso de aplicación. El resultado del *Python Script* son los datos limpios para aplicar correctamente el algoritmo de minería de datos. Esta etapa se conoce en la literatura como *ETL*, y es definido como el proceso de extraer, transformar y cargar los datos en un volumen mas grande o una infraestructura de datos distribuidos, con el fin de facilitar la operaciones de las técnicas de inteligencia de negocio, minería de datos, aprendizaje de maquina, metaheurísticas e inteligencia artificial en los datos estandarizados o normalizados. La segunda etapa de este proceso fue la elección de las técnicas a usar sobre los datos preparados, se tomo el bloque de datos siguientes como pruebas y el actual como los datos como aprendizaje y se considero una selección de los factores. Se continuo con esta parte de la etapa con la ayuda de Caruana *et al.*[39] y Hastie *et al.* [10], fue necesario tener las habilidades y la experiencia suficiente en reconocer cual es la técnica mas apropiada en el descubrimiento de las características en los RIPS, cuidando sobre todo la clasificación del tipo de variables y los meta datos. Para ilustrar esta etapa presentamos un algoritmo desarrollado en *R* (*randomForest*, *MASS*, *rpart*, *caret*), Este *R - Script* tiene como salida la selección de las características que mas aportan a la suma de cuadrados de los datos. Para el desarrollo de la tercera y ultima etapa se realizo un *Java POJO* mediante el reconocimiento de una semántica basada en la lógica relacional, que tenga en cuenta los modelos de analítica descriptiva que se pueden generar en el formato *Predictive Model Markup Language (PMML)* y los Modelos de características desarrollados en el formato *Simple XML Feature Model format (SXFM)*, que finalmente con la salida del *R - script* se dispondrán los datos en un orden y un formato en que cual, las técnicas de minería se conviertan en la entrada de los software de diseño asistidos por computadora *Computer-aided design (CAD)*, que se utilizan en el desarrollo de los Modelos de Características. Para escalar la solución a el 100 % de los datos se propuso en el diseño del modelo una infraestructura de *Big Data* que pueda soportar la gran cantidad de datos en las industrias actuales.

Capítulo 9

Resultados y análisis

Considerando los grandes volúmenes de información que poseen las industrias, los cuales muchas veces están dispersos y crecen rápidamente como silos de datos sin ser aprovechados [9], tuvimos en cuenta qué el uso de los modelos de características en el proceso de ingeniería de líneas de productos, puede usarse para la creación masiva de productos hechos a la medida, sin descuidar la calidad, la reutilidad y la variabilidad. En el proceso de ELP algunos productos serán penalizados [22], los modelos de características actuales siguen siendo grandes y complejos de configurar [12]. Es decir, generar o derivar productos desde la ingeniería de líneas de producto y considerar la ciencia de los datos para identificar o seleccionar que características toman protagonismo, o impactan positiva o negativamente los intereses de los nichos de mercado [10], sigue siendo un aspecto en constante desarrollo. Por esta razón se planteo el primer objetivo en este proyecto y en él se realizo el siguiente estado del arte.

9.1. Revisión bibliográfica mediante mapeo sistemático de la literatura

9.1.1. Planificación de la revisión

Considerando que un árbol es una estructura jerárquica, toda técnica de minería que use este enfoque puede ser interesante en la construcción de MC. Nos interesamos particularmente en buscar las técnicas de minería de datos que ofrecían este producto final, se estableció entonces que fue mucho mas conveniente haber buscado técnicas de minería y aplicarlas al contexto del proyecto de investigación, donde en particular, teníamos una gran cantidad de datos, buscábamos predecir una clase, los tipos de datos son categóricos y continuos, el numero de clases es conocido y perteneciente al dominio de la salud dentro del contexto de la información sobre la facturación. Así definimos los aspectos que se dan origen a las preguntas de investigación

Desarrollo de las preguntas de investigación

La meta principal de esta sección fue la clasificación de la literatura que contiene las técnicas de MD en el contexto de la ILP, con el objetivo de desarrollar el método para la construcción de los modelos de características. Para alcanzar este gran objetivo propusimos las siguientes preguntas basadas en metas mas pequeñas.

Con *Splunk*^{®1} desarrollamos los gráficos para el análisis del estado del arte, *Splunk*[®] hace

¹  Library.zip

mas accesibles, útiles y valiosos los datos obtenidos a partir del gestor bibliográfico *Mendeley Ltd.* exportados en el formato de archivo *Research Information Systems (RIS)*, . Teniendo en cuenta las preguntas y la motivación de la metodología, se desarrolló el siguiente Mapeo Sistemático de la Literatura, *Systematic Mapping Studies (SMS)*. En la gráfica 9.1 presentamos los resultados que ayudaron a solucionar la primera pregunta de investigación planteada en la metodología.

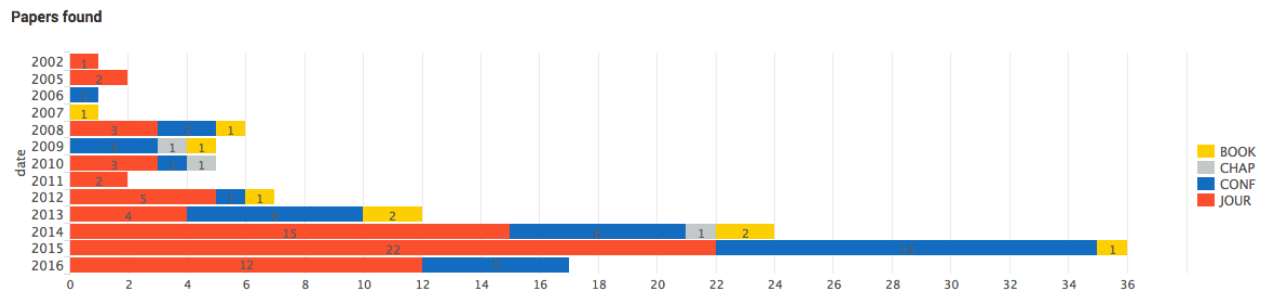


Figura 9.1: Distribución de los documentos por tiempo y por tipo. Captura tomada de la aplicación library.

RQ1:¿Cómo están distribuidos los estudios en el tiempo? En la gráfica 9.1, se mostró una tendencia en los últimos tres años, el protagonismo de los documentos encontrados en las memorias de las conferencias y talleres es notorio seguido de las revistas indexadas. Para los investigadores es interesante saber que hay una demanda en los temas que abarcan la ingeniería de software, el reuso y la variabilidad. De la misma manera, todo lo que se considera en *Big Data & IoT* esta tomando fuerza. Por otra parte, no se descarta la idea de la simplicidad y la rapidez de publicación, como criterio de elección para el lugar, como se muestra en la tabla 9.1, puede ser mucho mas fácil para un investigador realizar un aporte en las memoria de las conferencias a desarrollar un capitulo en un libro. Por esta razón nos preguntamos.

RQ2: ¿Qué tipo de documentos hablan del papel de la minería de datos en la ingeniería de líneas de producto? Los tipos de documentos encontrados varían en el tiempo pero presentan una tendencia a ser artículos de revistas indexadas y memorias en las conferencias, se quiso resaltar las palabras claves mas populares en la gráfica 9.2, que nos ayudarán en la construcción de nuestro *Quasi-Gold*, como también los diferentes lugares de publicación presentes en la tabla 9.1

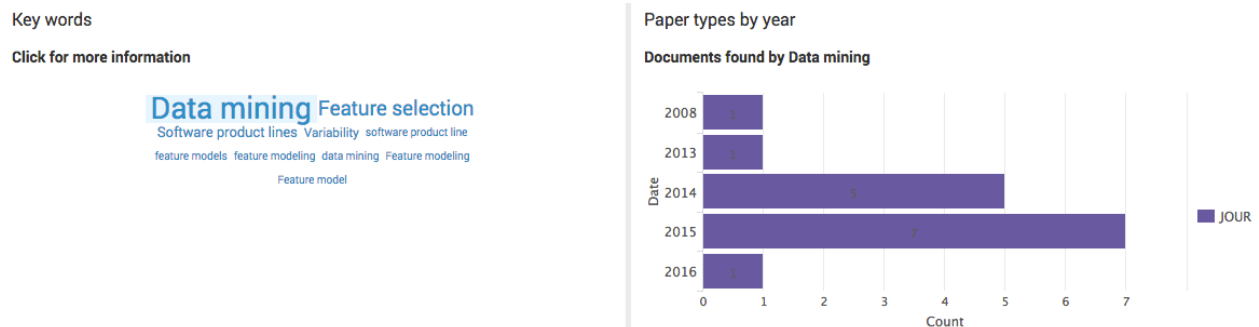


Figura 9.2: Captura de las sección de palabras claves en la aplicación library.

Tabla 9.1: Cantidad de documentos encontrados en los diferentes Lugares de publicación a lo largo del tiempo

Lugar	Cantidad
Expert Systems with Applications	8
Journal of Mechanical Design	4
Proceedings of the 19th International Conference on Software Product Line - SPLC '15	4
Proceedings of the Tenth International Workshop on Variability Modelling of Software-intensive Systems - VaMoS '16	4
Information Sciences	3
Information and Software Technology	3
Proceedings of the Ninth International Workshop on Variability Modelling of Software-intensive Systems - VaMoS '15	3
2008 12th International Software Product Line Conference	2
2013 4th International Workshop on Product Line Approaches in Software Engineering (PLEASE)	2
ACM Computing Surveys	2
Applied Soft Computing	2
Empirical Software Engineering	2
IEEE Transactions on Software Engineering	2
Knowledge-Based Systems	2
Neurocomputing	2
Proceedings of the 18th International Software Product Line Conference on Companion Volume for Workshops Demonstrations and Tools - SPLC '14	2
Proceedings of the Eighth International Workshop on Variability Modelling of Software-Intensive Systems - VaMoS '14	2
Requirements Engineering	2
2009 17th IEEE International Requirements Engineering Conference	1
2009 International Conference on Information and Multimedia Technology	1
2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery	1
2011 37th EUROMICRO Conference on Software Engineering and Advanced Applications	1
2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)	1
2014 10th International Conference on Natural Computation (ICNC)	1
2014 40th EUROMICRO Conference on Software Engineering and Advanced Applications	1
2014 Second IEEE Working Conference on Software Visualization	1
2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)	1

Continued on next page

Tabla 9.1 – *Continued from previous page*

Lugar	Cantidad
2015 IEEE 22nd International Conference on Software Analysis Evolution and Reengineering (SANER)	1
2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)	1
2015 IEEE Second International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)	1
Advanced Engineering Informatics	1
Advances in Product Family and Product Platform Design	1
Applied Computing and Informatics	1
Applied Energy	1
CEUR Workshop Proceedings	1
Collaboration and Technology	1
Computer Standards & Interfaces	1
Computer-Aided Design	1
Computers & Chemical Engineering	1
Computers and Electronics in Agriculture	1
Construction and Building Materials	1
Control Engineering Practice	1
DYNA	1
Elements	1
Engineering Applications of Artificial Intelligence	1
Evolving Software Systems	1
Information Systems	1
International Journal of Electrical Power & Energy Systems	1
International Journal of Forecasting	1
Journal of Computing and Information Science in Engineering	1
Journal of Retailing and Consumer Services	1
Journal of Software: Evolution and Process	1
Journal of Systems and Software	1
Journal of Testing and Evaluation	1
Journal of the Brazilian Computer Society	1
Journal of the Royal Statistical Society: Series B (Statistical Methodology)	1
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	1
Nature	1
Pattern Recognition	1
Performance Evaluation	1
Proceedings - 40th Euromicro Conference Series on Software Engineering and Advanced Applications SEAA 2014	1
Proceedings of the 16th International Conference on Computer Systems and Technologies - CompSysTech '15	1

Continued on next page

Tabla 9.1 – *Continued from previous page*

Lugar	Cantidad
Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12	1
Proceedings of the 18th International Software Product Line Conference on - SPLC '14	1
Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering - EASE '15	1
Proceedings of the 19th International Conference on Software Product Line	1
Proceedings of the 2009 ACM symposium on Applied Computing - SAC '09	1
Proceedings of the 2012 International Symposium on Computers in Education (SIIE)	1
Proceedings of the 2015 International Conference on Software and System Process - ICSSP 2015	1
Proceedings of the 23rd international conference on Machine learning - ICML '06	1
Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16	1
Proceedings of the IEEE	1
Proceedings of the Ninth International Workshop on Variability Modelling of Software-intensive Systems	1
Proceedings of the Seventh International Workshop on Variability Modelling of Software-intensive Systems - VaMoS '13	1
Proceedings of the Tenth International Workshop on Variability Modelling of Software-intensive Systems	1
Proceedings of the {ER} 2009 Workshops ({CoMoL} {ETheCoM} {FP-UML} {MOST-ONISW} {QoIS} {RIGiM} {SeCoGIS}) on Advances in Conceptual Modeling - Challenging Perspectives	1
Remote Sensing of Environment	1
Scientific Reports	1
Software Product Lines in Action: The Best Industrial Practice in Product Line Engineering	1
Statewide Agricultural Land Use Baseline 2015	1
The 7th International Conference on Information Technology	1

Continuamos con el análisis de los autores mas representativos de nuestros resultados y, nos preguntábamos.

RQ3: ¿Qué distribución geográfica o cuales autores son los mas representativos en nuestro estudio? Con el propósito de conocer las personas que pueden estar interesadas en nuestro proyecto de investigación, descubrimos la siguiente distribución geográfica para los autores mas populares de nuestro MSL como se muestra en la figura 9.3.

Mathieu Acher de la Université de Rennes 1, Rennes, Francia con cinco documentos en los cuales destaca un capítulo del libro *Lecture Notes in Computer Science*[50]. Desde Martin Becker de Fraunhofer-Platz 1, Kaiserslautern, Alemania que destaca con Bo Zang de la University of

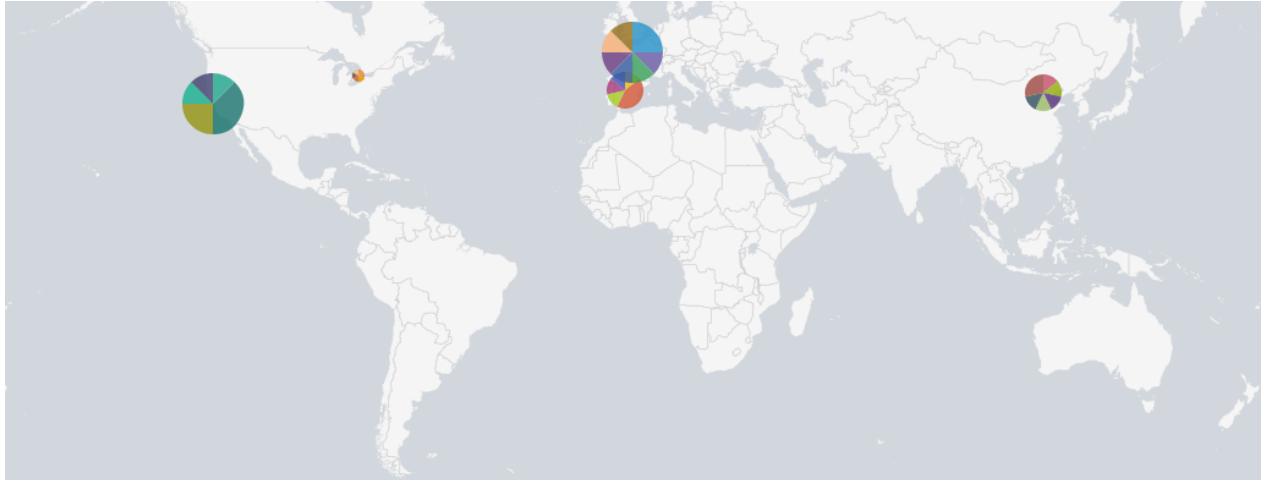


Figura 9.3: Captura de la sección donde se muestra la ubicación geográfica de autores según su universidad, tomada en la aplicación library.

Kaiserslautern, Alemania por *RECoVar*[17] y sus diversas contribuciones en los talleres sobre variabilidad en la ingeniería del software. También tenemos a Guillaume Bécancourt de Université de Rennes 1 en Rennes, Francia con un artículo muy interesante *On breaking the curse of dimensionality in reverse engineering feature models*[51], Rafael Capilla de Rey Juan Carlos University en Madrid, España propone administrar la variabilidad en la ingeniería de software desde su libro *Systems and Software Variability Management*[11], Krzysztof Czarnecki de University of Waterloo, en la provincia de Ontario en Canadá trabaja en las conferencias y talleres sobre minería de datos, Trevor Hastie de Stanford University en California, U.S.A., es un contribuyente excepcional al aprendizaje de máquina, con diferentes libros de una calidad sin parangón como *The elements of statistical learning*[10] se convirtió en el texto guía para muchos estudios y cursos, Raúl Mazo y Camille Salience de Université Paris 1 Panthéon-Sorbonne, París, Francia con artículos como *Reusable knowledge in security requirements engineering: a systematic mapping study*[44] y memorias en conferencias como VariaMos[25], protagonizan las regiones y las personas con las cuales podemos construir alianzas estratégicas y futuros trabajos.

Con el fin de establecer un grado de madurez para el tema de investigación como lo propone Roel Wieringa, *et al.* [38]. Proponemos la siguiente pregunta de investigación.

RQ4: ¿Qué categoría o escalafón de investigación tienen asignado los documentos encontrados en nuestro estudio? Basados en la cantidad de citas como se muestra en la figura 9.4, pudimos determinar algunos estudios importantes.

Descubrimos que en el campo de la MD y el *machine learning*, existe una comunidad vibrante de investigadores, es decir que hay personas muy apasionadas por generar conocimiento y, un ejemplo de la madurez de estos trabajos es *Regularization and variable selection via the elastic net*[52], donde se parte desde los avances de Leo Breiman y Trevor Hastie. En esta publicación se recuerda que desde la suma de cuadrados se pueden seleccionar los valores que más aportan en la varianza de las observaciones, se pasa por los métodos de selección de las variables como lo hace *Lasso* o *LARS algorithm* para proponer un acercamiento a los problemas que ocasiona la maldición de la dimensionalidad: Donde se tienen más factores que muestras. Y como darle solución mediante la penalización de los coeficientes de los factores de que no aporten tanto a la suma de cuadrados. Siguiendo con el tema de la MD y teniendo en cuenta que todo trabajo de investigación necesita justificar las herramientas que se usan o se proponen, es común que un ingeniero busque la mejor

RQ4: What category of research have assigned the documents found?

Click for go to paper

DOI	titulo	Citations
10.1111/j.1467-9868.2005.00503.x	Regularization and variable selection via the elastic net	3225
10.1016/j.is.2010.01.001	Automated analysis of feature models 20 years later: A literature review	506
10.1145/1143844.1143865	An empirical comparison of supervised learning algorithms	133
10.1007/978-3-642-29044-2	Experimentation in Software Engineering	123
10.1016/j.cad.2004.05.006	Product portfolio identification based on association rule mining	109
10.1145/2580950	A Classification and Survey of Analysis Strategies for Software Product Lines	93

« prev 1 2 3 4 5 6 7 8 9 10 next »

Figura 9.4: Captura de la sección donde se muestra en orden descendente los documentos según su cantidad de citas, tomada en la aplicación library.

herramienta, por esta razón consideramos a Rich Caruana y a Alexandru Niculescu-Mizil en *An empirical comparison of supervised learning algorithms*[39] donde se descubren las justificaciones para usar *randon forest* en la solución de los problemas con aprendizaje supervisado como el que se desarrollo en este proyecto de investigación con los RIPS. Por otra parte, en el campo del desarrollo de software tenemos el inconveniente de que casi todos los enfoques novedosos son publicados en eventos industriales como *Summit's*, por ejemplo en el caso de Apple esta *The Apple Worldwide Developers Conference (WWDC)* o el *Google I/O* en Google y en el caso de los videojuegos tenemos el *E3 - Electronic Entertainment Expo* para la presentación de estos avances. Sin embargo, las editoriales nunca están demasiado lejos cuando se trata de publicar documentos que cuenten como se generan productos personalizados de forma masiva, problema por el cual los modelos de características son la opción mas popular en los últimos años y por esto no fue raro encontrar una gran comunidad siguiendo los trabajos de David Benavides, *et al.*, como ejemplo tenemos *Automated analysis of feature models 20 years later: A literature review*[15]. En este trabajo de investigación se encontró que la ILP se usa en numerosos contextos de negocio igual que la MD, por esta razón identificamos estas coyunturas.

RQ5: ¿Qué algoritmos o técnicas son usadas en el ciclo de vida de la ingeniería de líneas de producto para crear modelos de características? Considerando el *framework* propuesto por Jianxin (Roger) Jiao en *Advances in Product Family and Product Platform Design*[16] sabemos que la ILP es un proceso que se le conoce por un ciclo de vida y este puede aplicarse en cualquier contexto. La literatura muestra la transformación del ciclo de vida de ILP a travez del tiempo, donde inicialmente se planteo en el mundo del software(*Software Product Lines in Action*[40]), pero como toda organización puede modelar sus procesos de negocio como un software, este tomo protagonismo y ahora se usa en la mayoría de los industrias y con el la ILP, esto permitió que atravez de la ILP surgieran los modelos de características como respuesta al modelado de la variabilidad de los productos en las industrias, esta pregunta se detiene en este punto, acá, proponemos hacer un alto y contemplar como las técnicas de minería de datos se usan para crear modelos de características.

- **Data Mining-Driven Product Design:** Iniciamos con *Advances in Product Family and Product Platform Design*[16] donde Timothy W Simpson y Jianxin Roger Jiao nos recordaron el campo emergente *data mining-driven product design* el cual tiene como objetivo incorporar a los grandes volúmenes de datos en el proceso de diseño de los productos, por ejemplo, emplear *clustering* para el agrupamiento de las funciones de los productos con el fin de crear familias de productos, usar *Naive Bayes Classification* para la creación de productos, escogiendo entre muchas características las combinaciones de ellas que se consideran novedosas, utilizar *data mining Fuzzy c clustering techniques* como una estrategia de identificación de plataformas en

el diseño de una familia de productos o proponer un *Data Mining framework* para extraer el conocimiento con el fin de crear plataformas de productos, entre otros.

- **Feature-Oriented Modeling Approach:** Yang Guanzhong y Zhang Yaru en *A feature oriented modeling approach for embedded product line engineering*[53] propone describir las características de los productos de manera detallada mediante un proceso de modelado de características que inicialmente se divide en dos partes, análisis del dominio y análisis de requerimientos, con esto se pasa a un modelo conceptual de dominio que permitirá el análisis de las características. Este análisis se extraen las relaciones entre características y su variabilidad e inmutabilidad, luego, se construye el modelo de características y se valida. En este documento se tienen en cuenta las propuestas tradicionales como *Feature Oriented Domain Analysis (FODA)*, *Feature-Oriented Domain Model(FODM)* & *Feature Oriented Reuse Method (FORM)* para proponer, que en las líneas de producto embebidas se pueden extender de las características de los productos un conjunto de conceptos, características y atributos. Con todo esto en mente, la vista de un modelo de características quedaría dividida en tres partes, modelo de características funcionales, modelos de características no-funcionales y modelo de características del sistema. Las relaciones entre estos modelos especificará la consistencia o la flexibilidad en la variabilidad de la línea de productos. En *Code Smells Revisited: A Variability Perspective*[54] Wolfram Fenske y Sandro Schulze describen que la variabilidad que tiene de un modelo de características y la detección de errores en el software, se puede considerar con FODA, el enfoque de anotaciones y el de los mecanismos basados en composiciones.
- **Evolution of Software Product Lines:** Al analizar la evolución de la línea de productos podemos obtener una colección de modelos de características útiles para la creación de nuevos productos, en *Evolution of Software Product Lines*[55] Goetz Botterweck y Andreas Pleuss proponen realizar una ingeniería inversa para la creación de modelos de características iniciando por ejemplo de características sin estructura, arquitecturas o incluso descripciones informales de los productos. El proceso que se describe en el documento anterior es complejo pero los autores lo resumen como un proceso de migración, que se compone de varios procesos mas pequeños, como el análisis de la línea de producto, la planeación de su futuro y finalmente la implementación de su evolución. Luego, como todo proceso es necesario chequear la consistencia de la línea de productos, con esto Alcimir Rodrigues Santos, *et al.* presentan un resumen sistemático sobre las estrategias para chequear la consistencia en las LPS[20] considerando también su evolución. El kernel de Linux a sido estudiado ampliamente en el contexto de la ILPS, por ejemplo Rothberg Valentin *et al* en *Feature Models in Linux*[56] muestra como el funcionamiento de los componentes está supeditado a los casos de uso. El modelo de características perteneciente al kernel de Linux es una colección de mas de 15200 características que crece cada día, y en su evolución fueron creados los modelos de características que representan su línea de productos. En este documento se presenta un marco de trabajo (*FMDiff*) que realiza operaciones de división y comparación para clasificar y analizar los cambios en el kernel, luego se emplea *Kconfig extractor dumpconf from the undertaker suite* para generar el modelo de características según la arquitectura en *Rigi Standard Format (RSF)*. Finalmente cuando se cuenta con el modelo de características en el formato RSF se pueden seguir haciendo las comparaciones y darle continuidad al proceso de evolución.
- **Ingeniería inversa de la variabilidad y las configuraciones:** Esta técnica obedece al *framework: RECoVar*[17] propuesto por Bo Zhang y Martin Becker que tiene como objetivo

general el incremento de la productividad en la línea de productos de software y librarse de la erosión en los artefactos durante la evolución de la LPS, este framework incluye dos enfoques el primero es la extracción del modelos de varibilidad basado en el código, mientras que el segundo identifica las correlaciones complejas en las características presentes en la configuración de la LPS. para el proceso de minería de datos usan *the Orange tool* y en los trabajos futuros no descartan la idea de usar los modelos de características probabilísticos (*Probabilistic Feature Model (PFM)*) para la extracción de las configuraciones que involucren la correlación entre características.

- **The Feature Mining process:** La ingeniería inversa para la adquisición de un modelo de características realizada de manera manual puede consumir mucho tiempo, ser propensa a errores y requerir mucho esfuerzo, por esta razón Ra'Fat AL-Msie'Deen, *et al.*[23] proponen que si se necesita explotar las variantes existentes en un software para la construcción de una línea de productos de software, se debe realizar un *Formal Concept Analysis (FCA)* y luego extraer su variabilidad mediante la combinación de la similaridad léxica y estructural de sus *object-oriented building elements(OBE's)*. Para realizar esto Ra'Fat AL-Msie'Deen, *et al.*[23] usaron una *Dependency Structure Matrix (DSM)*, para establecer las dependencias entre las clases(*OBE's*), para analizar las dependencias y extraer los *OBE's* usaron *Eclipse Java Development Tools (JDT)*, presentaron los *OBE's* con la librería *JDOM* en el formato *XML*, para aplicar *FCA* usaron *Eclipse eRCA platform* y para la similaridad léxica crearon su propia técnica de *Latent semantic analysis (LSA)* basada aparentemente en un algoritmo de *Cluster K-means*, luego representa su modelo de características en un *Eclipse plug-in for Feature-Oriented Software Development*. llamado *FeatureIDE* o en *ArgoUML-SPL*. También podemos considerar que en la construcción automática de los modelos de características en donde se tiene en cuenta la maldición de la dimensionalidad o la alta dimensionalidad[57], se pueden aplicar algoritmos de reducción de dimensión con la motivación de sintetizar la información las variables y obtener la información relevante desde las matrices de configuración como se muestra en *On breaking the curse of dimensionality in reverse engineering feature models*[51] por Jean Marc Davril, *et al.* En *Feature model augmentation with sentiment analysis for product line planning*[18] debido a la compilación de grandes volúmenes de información sobre las preferencias de los clientes, Feng Zhou y Jianxin (Roger) Jiao realizaron un análisis de sentimientos sobre las preferencias de los clientes al comprar en una tienda en línea, con el objetivo de incorporar los gustos de los clientes en el proceso de construcción de un modelo de características. Como se ve también en *Mining customer knowledge for product line and brand extension in retailing*[58], este método que se propone en el documento anterior también buscaba mejorar la eficiencia y la calidad en la planeación de la línea de productos, en lugar de hacerlo por el camino tradicional que se indica en la ingeniería de requisitos.
- **The domain analysis process:** Aunque esta es una etapa dentro del ciclo de vida del desarrollo de software, esencial en el desarrollo de la línea de productos, Negar Hariri, *et al.*[59] propone usar el descubrimiento de información en las bases de datos para la creación de modelos de dominio que bien se pueden procesar para crear modelos de características. En este trabajo se propone usar algoritmos de agrupamiento difusos, que combinados con técnicas de extracción de reglas de asociación como *kNN* pueden recomendar características que deberían estar presentes en la construcción de futuros productos.
- **Composition Operators:** Como el proceso de análisis del dominio, presentado anteriormente, estas operaciones sobre los modelos de características (MC) no fueron en sí usadas desde su concepción para la creación de dichos modelos; fue con la aparición del

problema de la alta dimensional en los datos, en el desarrollo de la ingeniería de líneas de producto que se construían MC grandes y monolíticos, este proceso fue muy incomodo, propenso a errores y muy costoso para las diferentes partes interesadas en el desarrollo de la línea de productos. Con el fin de manejar la alta dimensionalidad los MC fueron separados y *composed* por varias operaciones con el objetivo de no perder la integridad cuando sean separados o unidos nuevamente. Esta filosofía de dividir y vencer dio a conocer dos operaciones basicas: *insert & merge* que se comparan al detalle en *Comparing Approaches to Implement Feature Model Composition*[50] en donde Mathieu Acher propone considerar otras dos operaciones *diff & refactoring* nosotros pensamos que la derivación de los modelos de características mediante sus operaciones pueden generar nuevos elementos y con ellos, nuevos MC.

- **Variability management:** Es una actividad que dentro de la ingeniería de líneas de productos maneja el conocimiento racional para la toma de decisiones entre las diferentes partes interesadas en un proyecto de ILP, con esto nos referimos a que la *Variability management* es una rama de la ciencia que se basa en los argumentos, las razones y las justificaciones detrás de cada decisión que toman los interesados. Anil Kumar Thurimella y Bernd Bruegge proponen la metodología *issue-based variability management methodology (IVMM)*[60] que combina preguntas, opciones y criterios.
- **The multi-level feature model:** En *Cross domain web information extraction with multi-level feature model*[32], Qian Chen propone que para manejar las características efectivamente, se debe hacer énfasis en las relaciones que aparezcan en la configuración del dominio, con estas relaciones se propone crear un modelo que permita la extracción de características considerando los cambios en el dominio. Esto incrementa la adaptabilidad y el reuso de las características en el modelo. Este es un modelo teórico de una estructura jerarquica que puede derivar modelos de características.
- **Automated feature model configuration:** Mohsen Asadi, *et al.* en *Toward automated feature model configuration with optimizing non-functional requirements*[12] establece la idea de emplear tecnicas de planeación para solucionar un problema de configuración, en este documento se transforman los modelos de características, mediante una tecnica de planeación en inteligencia artificial conocida como *Hierarchical Task Network (HTN) Planning* para que automáticamente seleccione las características mas relevantes para los interesados. En este documento se establece que el proceso de configuración del árbol de características estará dado por los requisitos funcionales y no-funcionales que se persiven en las necesidades de los interesados.
- **Extracción de características desde el lenguaje natural:** Noor Hasrina Bakar, *et al.* propone una completa colección de las técnicas usadas para extraer las características en el lenguaje natural de los requisitos en la ingeniería de líneas de producto de software[49], este estado del arte es relevante debido a que en él, se considera que las técnicas usadas para el descubrimiento y agrupamiento de las características ya sea desde la minería de datos, las técnicas automáticas y manuales, aún no están disponibles totalmente para la comunidad. Por otra parte, muchas de las técnicas carecen de una validación o de un caso practico. En consecuencia, muchos interesados en el tema aún siguen extrayendo las características mediante la ingeniería de requisitos.
- **Antologias:** En el documento de Lamia Abo Zaid, *et al.* *Applying semantic web technology to feature modeling*, se parte de que introducir y administrar la variabilidad en los productos

de software por ejemplo, no es una tarea trivial, pueden existir muchas variaciones de los productos y mediante diferentes notaciones se puede representar la misma información, por esta razón representar los modelos de características para un sistema puede llegar a ser muy complicado. Una antología representa el conocimiento de cierto dominio mediante clases, propiedades y restricciones. Considerando lo anterior en este documento se propone usar un framework para la construcción de modelos de características usando *web ontology language* (OWL).

- **Variability modeling Techniques:** En *A survey of variability modeling in industrial practice*[61], Thorsten Berger, *et al.* presenta la respuesta a tres preguntas que se planteó en el anterior *survey*, ¿Que notaciones en el modelado de la variabilidad son usadas? ¿cuales son las escalas de los modelos industriales? ¿Que beneficios y desafíos se perciben con el modelado de la variabilidad? Y aunque desde el 2012 el *Object Management Group (OMG)* estableció un estándar para el modelamiento de la variabilidad, la comunidad ha desarrollado varios acercamientos que pretenden adaptarse a las diferentes necesidades academicas e industriales sin seguir el estándar con rigor, por consiguiente, hay una diversa variedad de notaciones en la definición de los modelos de características en la ingeniería de líneas de producto².
- **Holistic feature modeling:** Jaejoon Lee, *et al.* en *A holistic approach to feature modeling for product line requirements engineering*[62] nos presentan la reconciliación entre la ingeniería de requisitos y la ILPS, con un enfoque que considera los problemas cotidianos y los modela mediante metas y atributos de calidad, para luego separar las características en los espacios del problema y la solución, finalmente crear el modelo de características que represente los productos de la LP. Usar modelos de metas y de características en el espacio de la solución y el problema respectivamente es un gran enfoque que proporciona múltiples puntos de vista de la LP, sistematicamente este enfoque puede establecer que las características pertenecen al espacio de la solución y los requerimientos al espacio del problema.
- **Temporal Market-Driven Responses:** En el proceso de extracción, carga y transformación que esta explico en *Advances in Product Family and Product Platform Design - Quantifying the Relevance of Product Feature Classification in Product Family Design*[63] se establece que en la creación de plataformas de producto y familias de producto (Termino para referirse tambien a las líneas de producto), a lo largo del tiempo deben aparecer, desaparecer y reaparecer características en los productos de acuerdo a las necesidades del mercado y siempre será interesante modelar que características se incluyen y cuales no en la próxima generación de la familia de productos.
- **Configuración interactiva de modelos de características:** Dadas las grandes proporciones que puede tener un modelo de características partimos de que la configuración inicial del modelo será el producto mas sencillo que cumpla con las reglas del contexto para el cual es diseñado, luego podríamos configurar productos que acarrearán mas características y por lo tanto mas complejidad. En *On lazy and eager interactive reconfiguration*[64], Mikolás Janota propone que *eager* provee información sobre el usuario y *lazy* busque en cientos de miles de características, con esta configuración interactiva se pueden sugerir productos para públicos determinados, es necesario aclarar que con la sugerencia de estos productos no se desea limitar al consumidor, este enfoque solo quiere establecer configuraciones por defecto para agilizar sus decisiones.

²<http://www.omgwiki.org/variability/doku.php>

- **Matrices de configuración y descripción de los productos:** Las descripciones de los productos son usualmente representadas en estas tablas conocidas en la literatura como *configuration matrix*, este formato describe los productos con el objetivo de documentarlos y diferenciarlos. Esta representación sirve de intermediario para obtener modelos de características, en *Synthesis of attributed feature models from product descriptions*, Guillaume Bécan, *et al.* desarrollan un algoritmo basado en los marcos de trabajo de *FAMA & FODA* para crear modelos de características llamados *attributed feature models (AFMs)* desde la descripción de los productos. Como forma de validación se espera que no haya pérdida de información cuando se represente la descripción del producto en forma de matriz o en la forma de modelo de características.

Por otro lado también es interesante para el desarrollo del proyecto de investigación justificar el uso de la minería de datos en PLE, por esta razón nos hicimos el siguiente cuestionamiento.

RQ6: ¿Qué técnicas de minería de datos son usadas en el campo de la ingeniería de líneas de producto? Actualmente, diversas tecnologías almacenan los datos del proceso de ILP, estos datos son valiosos y han sido aprovechados para tomar decisiones. En este estudio exploramos los diferentes métodos, técnicas, algoritmos y modelos que se han usado en la construcción de líneas de producto.

- **SimpleKMeans E-learning Web Miner (EIWM) example:** En *Software Product Line Engineering for e-Learning Applications : A Case Study*[65], Diego Pablo Sánchez, *et al.* muestra que dentro del contexto de la educación, las plataformas educativas como Moodle son muy usadas por diversas instituciones y que el reuso y la variabilidad de este software genera las condiciones adecuadas para la construcción de una línea de productos. En este documento se mostró como a partir de *EIWM* se pudo generar una línea de productos considerando los gustos de los estudiantes por una particular actividad dentro de un curso; se generaron nuevos cursos de acuerdo al descubrimiento de los gustos por cierto material en estudiantes y profesores. Por ejemplo, *E-learning Web Miner* usa *EM algorithm (Expectation Maximization)* combinado con *SimpleKMeans* para agrupar a los estudiantes de acuerdo a la similaridad en sus gustos. Por otra parte tenemos a Shu-Hsien Liao, *et al.* con *Mining customer knowledge for product line and brand extension in retailing*[58] y sus esfuerzos para entender las complicadas y sensibles demandas de los clientes sobre los catálogos de productos de Carrefour en Taiwan mediante minería de datos. Carrefour que es una empresa inmersa en el negocio del *Retailing* que inició una línea de productos con su propia marca para impactar de forma estratégica un nicho de mercado. Ellos decidieron extraer el conocimiento que se almacena en sus bases de datos. Una vez aplicaron la metodología CRIPS desde sus bases de datos relacionales hasta conseguir un *datawarehouse*, continuaron con la metodología de algoritmo *K-means* para implementar cluster analysis, no obstante según [58] algoritmos como *genetic algorithms*, *neural nets*, *decision trees*, *regression*, *etc.*, se pueden implementar para lograr un análisis del dominio que produzca una línea de productos para *Retailing*. Sin embargo sin un análisis inteligente de los productos que consumen los clientes no es correcto afirmar que el cliente está satisfecho con los productos que compra.
- **Domain analysis and reuse environment (DARE):** El análisis del dominio es una actividad que pretende identificar, analizar, organizar y modelar las características comunes en un dominio en particular. Negar Hariri, *et al.*[59] propusieron dos algoritmos para automatizar la tarea anterior, el primer algoritmo es *Fast Algorithms for Mining Association Rules* y usa minería de reglas de asociación de manera no-supervisada para descubrir la afinidad de las

características a lo largo de los productos y enriquecer el perfil de los mismos. El segundo algoritmo es *k-nearest neighbor* el cual toma los productos con sus perfiles aumentados y ayuda a los interesados en el proyecto de ILP recomendando las características en los productos, en el proceso de ingeniería de requisitos.

- **Conjunctive and disjunctive association rule mining:** En la construcción de modelos de características probabilísticos, Krzysztof Czarnecki, *et al.* presenta un procedimiento en *Sample Spaces and Feature Models: There and Back Again*. [66] que aplica *conjunctive and disjunctive association rule mining* para encontrar patrones de características recurrentes, y luego generar modelos de características probabilísticos, recordemos que los modelos de características son la representación de la LP, entonces los modelos de características probabilísticos buscan extender la funcionalidad de los modelos de características básicos con la inclusión de restricciones suaves, las restricciones suaves son configuraciones en los productos que no son obligatorias para el usuario. Las reglas de asociación son muy usadas por las herramientas de construcción de software basadas en ingeniería inversa como se explico en la pregunta anterior en *the feature mining process*. Las reglas de asociación conjuntas buscan relaciones de tipo *AND* en los productos, es decir que producto implica otro, esto se conoce también como *frequent itemsets*. Las reglas de asociación disyuntas buscan relaciones de tipo *OR*, es decir, que productos estarán bajo una categoría y se pueden agrupar. En *Product portfolio identification based on association rule mining*, Jianxin Jiao y Yiyang Zhang [31], propusieron que la identificación de los productos iniciales en una línea de producto o el diseño incremental del catalogo de productos, se desarrollara a través de una identificación de los productos mediante *association rule mining system (ARMS)*, donde se hizo la selección de las variables en los productos que mas se ajusten a la estrategia de mercadeo, desde los registros de las ventas pasadas y las descripciones de los productos. Estas practicas investigativas y industriales tiene como objetivo identificar la configuración optima en los productos que generan la mayor ganancia y satisfacción en los clientes, como ejemplo en [31] se presento un ejemplo de una compañía que produce motores de vibración para celulares, en donde la arquitectura ARMS diferencio las necesidades del cliente desde los requisitos funcionales y el dominio de funciones. La identificación del portafolio de producto involucró la realización de un agrupamiento de requisitos funcionales para cada necesidad del cliente. Estos agrupamientos se realizaron mediante *fuzzy clustering analysis* y el mecanismo de mapeo se realizo mediante reglas de asociación.
- **Relational Machine Learning:** Maximilian Nickel, *et al.* en *A Review of Relational Machine Learning for Knowledge Graphs* [3] presentaron un resumen sobre como los modelos estadísticos pueden ser entrenados con grafos voluminosos, con el proposito de que cuando el modelo prediga un nuevo hecho, este se convierta en un nuevo nodo del grafo. En este documento se exploraron la factorización tensorial y las *multiway neural networks* y se discute como este conocimiento combinado puede originar una LP teniendo en cuenta el costo computacional que es inherente al trabajar con gráficos.
- **Feature Selection and Classification Algorithms:** Adam Woznica, *et al.* en *Model mining for robust feature selection* [67], presentan un framework para la creación de modelos de características. En la creación de un modelo de características Adam Woznica, *et al.* estableció que los algoritmos de selección de características más apropiados seguían los siguientes modelos: $\mathbf{w} = (w_1, \dots, w_p)^T, w_l \in \mathbb{R}$ teniendo en cuenta que \mathbf{w} le asignaría un peso a cada característica, en la literatura este problema de clasificación es conocido como *feature weightings*, $\mathbf{r} = (r_1, \dots, r_p)^T, r_l \in \mathbb{N}^+$ teniendo en cuenta que r_l le asignaría un escalafón a

cada característica, en la literatura este problema de clasificación es conocido como *feature rankings*, $\mathbf{s} = (s_1, \dots, s_p)^T$, $s_l \in \{0, 1\}$ teniendo en cuenta que $s_l \in \{0, 1\}$ representa la ausencia o presencia de cada característica, en la literatura este problema de clasificación es conocido como *feature subset*. En este documento se propone detectar la mejor estrategia para describir un producto con sus características se insinuaron técnicas para *Mining Multi-label Data*, y con ello proponer o crear modelos de características en una LP.

- **On-demand clustering framework:** Es importante mencionar el uso de los métodos de agrupamiento en la construcción de una línea de productos y también en el proceso de ingeniería del software, Nan Niu y Steve Easterbrook, en *On-Demand Cluster Analysis for Product Line Functional Requirements*[68] compartieron sus observaciones y experiencias al encontrar que desde los requisitos funcionales se puede lograr una mejor comunicación con los interesados en un proyecto de desarrollo de software, darse cuenta de las preocupaciones de los interesados en la etapa de análisis de ingeniería de línea de producto, proveer de un contexto y bajar el lenguaje de abstracción entre los interesados y los ingenieros. Con el análisis que se realiza con ayuda de las técnicas de minería que se muestran en el documento anterior, el diseño de la línea de producto se vuelve mas atractivo tras descubrir que características y atributos son indispensables desde los requisitos funcionales del software. El hecho de construir modelos de dominio desde los requisitos funcionales es un acercamiento de lo poderoso que pueden llegar a ser las técnicas de minería de datos en el contexto de la ILP.
- **Bagging:** En la ingeniería de líneas de productos las funcionalidades de estos productos son abstraídas como características, encontrar la relación de estas características implica un esfuerzo, en el que por ejemplo Pavel Valov, *et al.* en *Empirical comparison of regression methods for variability-aware performance prediction*[69] realizó una comparación en donde se descubrió que el *Bagging* de Leo Brieman es un acertivo y robusto selector de características, que en este caso en particular identifica como los diferentes factores influyen en el rendimiento de un producto. Falta.
- **Inteligencia artificial:** Como la inteligencia no puede ser desligada de la capacidad de aprender, no podemos desligar la inteligencia artificial del aprendizaje de maquina. Para aprender desde los datos, es necesario usar la minería de datos, el conocimiento en bases de datos, la inteligencia de negocio o las tecnologías emergentes como *BigData & Internet of things*. [10]

En este estudio fue relevante considerar otros estados del arte, como *Intelligent software product line configurations: A literature review* de Uzma Afzal, *et al.*[70]. En este estudio se presenta a la inteligencia artificial como un conjunto de modelos o implementaciones, que están diseñados para simular las acciones racionales de los seres humanos, del mismo modo se presenta porque estas las técnicas son relevante en la ingeniería de líneas de producto.

- **Lógica:** La lógica es un lenguaje matemático que captura el concepto de raciocinio, su sabores mas comunes son la lógica proposicional, lógica de primer orden y lógica descriptiva. La lógica es usada para representar las relaciones entre los diferentes objetos de un dominio. Por ejemplo las relaciones entre los objetos de un dominio sean completamente ciertos (1) o completamente falsos (0).

Por otra parte, la lógica difusa (*Fuzzy Logic*) es un tipo especial de lógica que establece que desde la lógica binaria que se planteó anteriormente, se pueda partir de un rango de valores $[0 - 1]$, para definir las relaciones de los objetos del dominio como parcialmente ciertas (0,8) o parcialmente falsas (0,2).

- *Knowledge-based reasoning*: La representación del conocimiento por sus siglas en ingles (*KR*) representa de forma simbólica la inferencia y el raciocinio, esta viene en tres sabores el deductivo el abductivo y el inductivo. Con esto se trata de establecer si un producto por ejemplo pertenece o no a una clase.
- *Ontological modeling and reasoning*: En la informática una ontología representa una entidad y sus interacciones, es una colección de conceptos, los componentes clásicos de una ontología son las clases, los atributos y las reglas. En la ingeniería de líneas de productos esta representación provee la variabilidad y el reuso de los componente, estas reglas son codificadas como *Semantic Web Rule Language (SWRL)* el cual es un estándar proveído por el consorcio *W3C* para el análisis del dominio del conocimiento.
- *Constraint satisfaction problem*: Los algoritmos que resuelven los problemas de satisfacción de restricciones o *CSP* por sus siglas en ingles, son usados en el modelamiento del espacio del problema en la ingeniería de líneas de productos, explícitamente para administrar la variabilidad, detectar modelos de características inválidos, realizar *debugging* y detectar errores en las características. Un *CSP* tiene tres componentes claves, la definición de las variables, el dominio de cada una y un conjunto de restricciones o limitaciones, estos algoritmos deben buscar por ejemplo las configuraciones de un producto dado un vendedor (Variable), en una región (Dominio) y con un rango de precios (Restricción o limitante).
- Optimización Los algoritmos de optimización tiene como objetivo la selección del mejor candidato en un listado de los posibles mejores. Los algoritmos genéticos por ejemplo pertenecen al dominio de la computación evolutiva para optimizar la solución de un problema. La idea que se tiene en la ingeniería de líneas de producto por ejemplo es la de agentes autogestionados, operando independientemente, de forma inteligente y paralela, para que cada agente pueda generar una solución de un problema multimodal.
- *Swarm intelligence techniques*: *Swarm intelligence* es un enfoque relativamente nuevo en el dominio de la optimización y los algoritmos genéticos. La naturaleza de la optimización ayuda a enfrentar los problemas que involucran una gran cantidad de características, como la configuración de los modelos de características en la ILP. por ejemplo en [70] se propuso usar *Artificial Bee Colony (ABC)*, *Particle Swarm Optimization (PSO)* y *Ant Colony Optimization (ACO)* para justificar la búsqueda de la solución de un espacio multimodal, para la optimización de la configuración de productos en una LP de software.
- *Predictive analytics*: El análisis predictivo, (*PA*) por sus siglas en ingles, es una rama de la inteligencia artificial que se enfoca en el procesamiento inteligente de los datos de negocio, con el fin de extraer propuestas de valor que aumente el rendimiento. En la ILP, el análisis predictivo es usado para encontrar inconsistencias que puedan existir con las características, con esto se pretende clasificar patrones en la selección de características que generen inconsistencias y predecir la aparición de estos patrones en futuros modelos de características.

Con el propósito de considerar las Empresas y los sectores de la industria en las cuales se realizaron trabajos interesantes proponemos la siguiente pregunta de investigación.

RQ7: ¿Qué contexto de dominio está implementando minería de datos e ingeniería de líneas de producto? En esta pregunta, proponemos algunos dominios interesantes que además de implementar líneas de productos también hacen uso de las técnicas de minería de datos. Como resultados representativos de esta sección recomendamos *Intelligent software product*

line configurations: A literature review [?], un documento que nos resume esta pregunta y adicional habla sobre las aplicaciones de la inteligencia artificial en la ingeniería de líneas de producto de software, las posibles vias de investigación, el impacto y el poder predictivo de las técnicas última generación.

Junto con Michael J. Corl, Michael G. Parsons y Michael Kokkolaras en *Advances in Product Family and Product Platform Design*[63] se pueden explorar las aplicaciones en el dominio de los productos de empresas con un enfoque militar (barcos) y ademas presentamos una serie de ejemplos interesantes que se identificaron en la literatura, los cuales hablan sobre el desarrollo de tecnologías para automóviles y teléfonos celulares.

Adicionalmente se recomiendan los 36 documentos encontrados por en el estudio sistemático de Larissa Rocha Soares, *et al.* en *Analysis of non-functional properties in software product lines: A systematic review*[46] especificamente la pregunta de investigación 1.3 sobre los dominios cubiertos.

- **Desarrollo de nuevas cámaras digitales):** En la búsqueda de la cámara que los clientes necesitan, Jae Kwon Bae & Jinhwa Kim,[6] se plantean, ¿exactamente qué necesitan los clientes o qué esperan de las cámaras digitales? y para resolver esta pregunta de investigación recurrieron a la metodología que desarrolla el algoritmo C5,0 para extraer mediante minería de datos en material para el desarrollo de la primera fase de análisis en una línea de producto: el diseño del producto y los estudios de mercado.
- **Desarrollo de nuevos teléfonos celulares:** En la identificación de un portafolio de productos de vibradores de celulares que cumpla con las necesidades del cliente en cuanto a la funcionalidad, Jianxin Jiao & Yiyang Zhang, en *Product portfolio identification based on association rule mining*[31], proponen que los requerimientos del cliente se integren como el insumo inicial en el desarrollo de la metodología de reglas de asociación. Debido a que los clientes esperan una mejora en los sistemas actuales de clasificación y adquisición de la ingeniería de requisitos la cual por el momento es propensa a errores, este desarrollo implica que otras técnicas[71] y los sistemas expertos puedan ser implementados en el futuro en el ciclo de vida de las líneas de producto. No obstante en el diseño de las nuevas plataformas de productos, el objetivo del cliente se ha convertido en hacer coincidir las expectativas del diseño con las del rendimiento de los nuevos teléfonos celulares, para conseguir el producto de menor costo y de características competitivas [72].
- **Desarrollo de nuevos automóviles:** Como veremos con las tablets, en los telefonos celulares también en el dominio de los autos se han presentado estudios que consideran el analisis de sentimientos en una de las etapas de la ingeniería de líneas de producto, en *Automated Discovery of Lead Users and Latent Product Features by Mining Large Scale Social Media Networks*[73], se toman los datos del *Twitter* y *Facebook* que pertenecen a 2,547 teléfonos inteligentes de 33 compañías diferentes y 29 de los mejores y los peores atomóviles reportados por los consumidores, con estos datos se crearon sistemas de puntaje para identificar el mejor y el peor producto dado el comportamientos de los usuarios en las redes sociales durante todo el ciclo de vida de los productos, luego se identificaron las características mas fuertes y débiles, para ayudar a los diseñadores a entender porque un producto es considerado bueno o malo. Con el uso de las redes sociales los investigadores pudieron inferir las expectativas de los clientes con el fin de predecir la percepción del mercado sobre los prototipos de los productos.
- **Planificación de la nueva generación de las *Kindle Fire HD tablets*:** En *Feature model augmentation with sentiment analysis for product line planning*[18] Feng Zhou & Roger Jiao proponen descubrir la información sobre las preferencias de los clientes en los productos

Kindle dado un conjunto de características del idioma inglés, los investigadores pronosticaron sentimientos para determinadas características en los productos, las características que sean recomendadas por los sentimientos de los usuarios se usaran para el diseño incremental del modelo de características.

Con esto, no solo se estaría diseñando una nueva generación de productos, también se están considerando las preferencias de los clientes y produciendo en el modelo de características una característica nueva que involucra el gusto de la comunidad como un porcentaje positivo o negativo, estas reseñas en los productos comúnmente se muestran con estrellas en las plataformas de compras web, en donde el consumidor puede interactuar con el dueño del producto y pedirle explícitamente que quiere ver en las nuevas versiones.

En la identificación de las técnicas de minería de datos tomamos en cuenta la información arrojada por las preguntas anteriores y para iniciar el proceso de adopción fue útil preguntarnos

RQ8: ¿Cómo están definidas las técnicas de minería de datos encontradas en los documentos? En el contexto de este trabajo de investigación, se encontraron interesantes las técnicas que se definen generalmente como parte de un proceso de extracción de conocimiento, en donde para cada problema se puedan recomendar varias técnicas como se muestra en Scikit Learn[74]. También según las recomendaciones de Jaroslav Pokorný[75], para nuestro caso particular usamos *clustering*, dado el sabor de las bases de datos usadas(*backup* RDBMS) y el enfoque del modelo que estaría en formato XML.

En este trabajo de investigación se considero la metodología propuesta por Conrad S. Tucker en *Quantifying the Relevance of Product Feature Classification in Product Family Design*[16], en donde se definieron los modelos de características como árboles de decisión de series de tiempo. Esta metodología se basa en *Data Mining-Driven Product Design* y el framework del descubrimiento de conocimiento de bases de datos, la metodología consta de los siguientes pasos: Adquisición de datos → Selección y limpieza de datos → transformación de los datos → minería de datos y descubrimiento de patrones → interpretación y finalmente evaluación. En esta investigación también encontramos que desde *Data Mining-Driven Product Design*, se definen las técnicas de la minería de datos como la representación de los cambios en los modelos de características a través del tiempo[18], ¿cual de estas técnicas es la que mejor define el comportamiento de los gustos de los clientes?[4] y ¿cual es la mejor técnica que define este tipo de problemas en donde se involucran los modelos de características?, estas son preguntas interesantes que aún no se pueden responder porque se manejan muchos datos que pertenecen a muchos dominios y en donde no se conocen la cantidad de categorías.

Sin embargo en los documentos consultados en este trabajo nos encontramos con el siguiente enfoque, donde tendremos un conjunto de datos que describen una variable respuesta, mediante una función de clasificación.

$f(x) = \{(X|Y)\}$ $X, Y \in R_n$ es decir $X_1, X_2, \dots, X_p \rightarrow R_1, R_2, \dots, R_j$. Esta función considera las regiones presentes donde aparecerán las diferentes clases que se considera estimar, y como detalle técnico, las regiones serán divididas en los segmentos para pruebas y entrenamiento.

Los estimados para hallar la clase en la región (R_j), serán en un principio *the most commonly concurring class* por tratarse de datos categóricos. Considerando que, los datos observados pertenecen a determinado conjunto de observación (X_p), dado el comportamiento del conjunto de variables Y_k , Entonces, podemos definir la probabilidad de que una observación pertenezca a una clase k en la m -ésima región P_mk , dado que, la fracción de las observaciones de entrenamiento que no pertenecen a la más común de las clases esta dada por $E = 1 - \max_k(P_mk)$.

La anterior medida (*the most commonly concurring class*) nos insinúa un poco la distribución de Bernoulli, la cual no es muy sensible, pero con el tiempo esta función se ha descrito mejor como el

Gini-index, $G = \sum_{k=1}^K P_m k(1 - (\hat{P}_m k))$, en donde un menor valor significa que el nodo donde se encuentran las observaciones contiene valores de esa la clase k en esa región $P_m k$.

Por otra parte, hay mejores interpretaciones de este problema, por ejemplo, la entropia - cruzada D , en este caso, es la medida que indica la proporción de incertidumbre que se tiene sobre una clase k en esa región $P_m k$, $D = \sum_{k=1}^K P_m k(\log(\hat{P}_m k))$, es decir no se tiene incertidumbre de pertenecer a una clase k en esa región $P_m k$.

Aunque estas medidas nos ofrecen mucha información para determinar las decisiones en un solo árbol de clasificación, siempre es interesante conocer las regiones en donde las clases k son representadas por las observaciones $P_m k$, $D \sim 0$ sii $\hat{P}_m k \sim 0 \rightarrow D \sim G$ con los índices anteriores el árbol podará los nodos que no ofrezcan mucha información. Sin embargo con esta medida no se pueden tomar decisiones. El error de clasificación fue la medida más encontrada en la mayoría de los ejercicios académicos e industriales.

RQ9: ¿En cuál etapa del ciclo de vida de la ingeniería de líneas de producto se usa la minería de datos? Considerando que la documentación de la variabilidad y la reutilización de los requisitos con que se construyen los artefactos en cualquier dominio, según Klaus Pohl y Thorsten Weyer[76] son el espíritu que alimenta las técnicas de minería, con esta pregunta se busca identificar las estrategias que incorporan la minería de datos en las diferentes etapas del ciclo de vida de la ingeniería de líneas de producto. Se recomienda el trabajo propuesto por Daniela Castelluccia y Nicola Boffoli [77], en donde se muestran las diferentes facetas del ciclo de vida de la ingeniería de líneas de producto: ingeniería del dominio, derivación de productos, configuración de productos, validación, verificación y mantenimiento de la línea de producto. junto con diferentes técnicas de minería de datos.

En este trabajo de investigación consideramos el framework[76] con el objetivo de automatizar la construcción de los modelos de características e identificar puntos de mejoras y optimización. Al usar la minería de datos en la etapa de ingeniería del dominio por ejemplo[55] el objetivo es encontrar el material necesario para desarrollar estrategias de mercadeo interesantes en el moldeamiento de familias de productos, como se muestra en *Platform Valuation for Product Family Design*[16]. Por otra parte, Bo Zhang & Martin Becker en RECoVar[17], proponen que desde la fase de evolución (mantenimiento), donde se hace *refactoring*, es decir, que desde los puntos de mejora, se haga ingeniería inversa o se apliquen técnicas de minería, con esta filosofía, se deberían crear elementos nuevos basados en los arquetipos presentes en el espacio de la solución. El diagrama de características del dominio es un elemento que está presente en la etapa temprana del proceso de ILP (ingeniería del dominio), es usado por los diseñadores para identificar la importancia de los actores y la funcionalidad. El descubrimiento de estos patrones en *On-Demand Cluster Analysis for Product Line Functional Requirements*[68], muestra que mediante una prueba de concepto con *overlapping partitioning cluster(OPC)* se pueden identificar además del agrupamiento de las funcionalidades y las entidades mas importantes, productos sugeridos y las metas que se deben cumplir para satisfacer a los inversionistas o a los interesados.

En las *Engineering Change (EC)* donde se consigna la administración de los cambios en los productos mientras se optimiza un diseño y se depositan en sus registros los costos, el tiempo, la cantidad de cambios y si el cambio necesita ser pronosticado o controlado, se propone una solución interesante[78], para la predicción de los cambios, está esta basada en un modelo (FSB) que combina los conceptos: funcion, comportamiento y estructura. El modelo se aplica a un caso practico de un motor disel en donde explica explícitamente todas las dependencias de los elementos del producto, permite modelar y calcular todas la solicitudes relevantes de cambio, mejora la compresión del origen de los cambios, es escalable a diferentes niveles, es decir permite cierta granularidad en el producto y es flexible, puede presentar la información de los resultados en diferentes niveles, Todas

estas características del modelo de vinculación FBS pueden ayudar a controlar y contrarrestar la propagación del cambio y reducir la incertidumbre y el riesgo en el diseño.

Después de aplicar las técnicas de minería de datos en las diferentes etapas del ciclo de vida de la ingeniería de líneas de producto se tienen algunos componentes que pueden ser relevantes para este trabajo de investigación.

RQ10: ¿Qué tipo de productos se han derivado del uso de las técnicas de minería de datos en la ingeniería de líneas de producto? Los productos que se generán a partir del uso de las técnicas de minería de datos son modelos en su mayoría, estos son archivos o expresiones matemáticas que sin un contexto no dicen mucho, pero con esta pregunta se exploran estas diferentes propuestas con el fin de remitirse a ellas en un futuro, cuando se tenga un contexto de interés.

Considerando los conjuntos de datos en los modelos de características del repositorio de SPLOT[26] con el formato SXFM y PML con el formato fmlbdd, Guillaume Bécan, *et al.* evaluaron mediante heurísticas y algoritmos de agrupamiento lo que se define como *Ontologic-aware Feature Model Synthesis*[79], en este trabajo se desarrolla un analizador sintáctico de ontologías basado en FAMILIAR[80] (WebFML), con este desarrollo se quiere apoyar las tareas del mantenimiento en la el proceso de la ILP y extraer información de *generic ontologies* como Wordnet y *open collaborative based initiatives* como Wikipedia. En *Synthesis of attributed feature models from product descriptions*[81], Guillaume Bécan *et al.* se propone un algoritmo derivado de las técnicas de minería de datos que procesa una estructura jerárquica de características o grupos de ellas, descubre componentes del dominio, características y las relaciones entre ellas, para descripciones de productos alojadas en *The Best Buy dataset*, para este producto, el límite de restricciones está limitado por la gran dimensionalidad de los datos.

La extracción de características para proponer productos que se ajusten a las necesidades de los consumidores, es un tema que se ha desarrollado desde diferentes contextos en este trabajo de investigación, por sí mismo, es un producto que se origina después de aplicar técnicas de minería de datos. Por ejemplo, en *Cross domain web information extraction with multi-level feature model*[32], Qian Chen, *et al.* desde dos algoritmos de minería como *Learning information extraction rules for semi-structured and free text* & *A hierarchical approach to wrapper induction* pueden extraer de diferentes dominios, las características más relevantes, las características que están compuestas o dependen de otras o las que están menos relacionadas con el dominio, estas características son llamadas sub características o características atómicas. Con esta información se pueden generar soluciones a problemas específicos, o en este caso representar la información mediante *multi-level feature models*, con el objetivo de adaptar un modelo a los intereses cambiantes de una organización. En la identificación de la información en la fase de evolución dentro de la ILP, Christoph Seidl, *et al.* en *Capturing variability in space and time with hyper feature models* presenta *Hyper Feature Models (HFM)* para explicar diferentes elementos de la arquitectura OSGi en Eclipse, por otra parte la ingeniería de líneas de producto de software está ampliamente desarrollada en la academia[82] y satisfactoriamente implementada en la industria. Basados en la selección de características los desarrollos de los proyectos de software, identifican los requisitos no funcionales de diferentes maneras[46], este producto ha generado espacios para los analistas del negocio puedan considerar el análisis que proveen diferentes métodos para eliminar el sesgo en la dirección de los proyectos o las investigaciones.

En la búsqueda de nuevos productos se considera también de forma implícita el trabajo de los investigadores, al redactar como un producto en las diferentes revistas y talleres todo el material bibliográfico sobre las técnicas de minería de datos en la ingeniería de líneas de producto, estos productos aunque no se consideran todos en este trabajo de investigación, se mencionan si se requiere profundizar en esta pregunta en el futuro. Trevor Hastie, *et al.*[10] pudieron concluir que

las técnicas de minería de datos están siendo aplicadas o usadas con diferentes propósitos y en diferentes dominios, sin embargo hasta el momento esto se demostraron mas no se han mostrado estudios sobre la validación de las técnicas de minería de datos. En la siguiente pregunta se quiere explorar o mencionar algunos autores y trabajos importantes que se encontraron en el ámbito de la validación.

RQ11: ¿Cómo se puede validar el uso correcto de las técnicas de minería de datos en la ILP?

Los problemas de selección de características tienen como objetivo pronosticar si existe o no una característica dependiendo de los datos observados. en *A causal feature selection algorithm for stock prediction modeling*[83], por ejemplo, Xiangzhou Zhang, *et al.* proponen una selección de características con el propósito de predecir el inventario para conocer las características de los productos(*observational data-based causal analysis to stock prediction*). En este estudio, se usa el metodo de ventana deslizante que divide los datos en pruebas y entrenamiento, es natural para los inversionistas hacer predicciones sobre los datos mas recientes, pero eso, para los investigadores es un reto tener un *set* de datos de prueba y entrenamiento con pocos datos, por esta razón es recomendable una estrategia de particionamiento, usar los datos mas recientes para capturar las características de interés, establecer estos datos como entrenamiento y luego buscar datos similares en los conjuntos de datos pasados para realizar las pruebas. En este estudio se prueban el modelo sobre 9 diferentes *datasets* donde se destacan la evaluación de las medidas de exactitud, precisión, adecuación y estabilidad para los algoritmos de clasificación. Finalmente se pueden evaluar la rentabilidad de los modelos como un *benchmark* con el *Shanghai Composite Index* en los periodos que se muestran en los datos.

En *A Data-Driven Network Analysis Approach to Predicting Customer Choice Sets for Choice Modeling in Engineering Design*[84] Mingxian Wang & Wei Chen, predicen las elecciones de los consumidores mediante un modelo de análisis de red, que mediante la segmentación del mercado se puede analizar la heterogeneidad de los clientes, primero se examinaron la similaridad de las asociaciones entre productos, luego se presentan los resultados de los modelos despues de entrenar y probar, estos resultados se miden de acuerdo a su medida de exactitud, al comparar dos matrices formadas por las elecciones predecidas contra el conjunto de elecciones, finalmente se le llama a este estimador el *hit-rate*. Con esta técnica se pueden entender las preferencias de los clientes, con la meta de mejorar el diseño de los productos.

Con *Rialto: A Knowledge Discovery suite for data analysis*[85] de Giuseppe Manco, *et al.* se describe una plataforma de descubrimientos de conocimiento basada en KDD, que tiene como principal objetivo la elaboración de un marco de trabajo para proveer principios de diseño, detectar las características básicas y asegurar la eficacia de las decisiones en el diseño. En *Supporting Domain Analysis through Mining and Recommending Features from Online Product Listings*[59], Negar Hariri, *et al.* recomiendan las matrices de productos con recomendaciones sobre las características (*feature pool*) muy parecido a los enfoques vistos anteriormente[58, 67], Negar Hariri, *et al.* comparan y recomiendan algunos algoritmos mediante *m-fold cross-validation*, evaluar de manera cuantitativa los experimentos y de manera cualitativa el análisis[86]. Los resultados que fueron obtenidos en el sistemas de recomendaciones propuesto en este documento, tiene la capacidad de ayudar en el análisis de dominio[16] a crear modelos de variabilidad (se expone el caso de COSS software), adicionalmete los ejemplos que se muestran en el documento ilustran el proceso, sus beneficios y limitaciones.

Complementando, con *Wind power forecasting using the k-nearest neighbors algorithm*[87], Ekaterina Mangalova & Evgeny Agafonov, se puede escoger un algoritmo de agrupamiento, por su interpretabilidad, *k-nearest neighbors algorithm* puede soportar datos cíclicos, como periodos de tiempo (años, meses, días), no tiene problemas con la inclusion de nuevos datos(robustez), de manera

similar a los casos anteriores en este documento se trabaja con una partición de los datos para probar y entrenar el modelo. En este documento se tiene como objetivo pronosticar la operación de la planta eólica, para evitar usar centrales de reserva cuando sea innecesario.

Con el objetivo de automatizar el desarrollo de los Modelos de características y caracterizar la aplicabilidad de las técnicas de minería de datos surge la siguiente pregunta.

RQ12: ¿Los estudios encontrados en la literatura apoyan la automatización de técnicas de minería de datos?

Como respuesta al tema de la automatización se pensó en como hacer pruebas automáticas a todo un proceso partiendo desde la ingeniería de líneas de producto. En donde, se descubrió el aporte de Beatriz Pérez Lamancha, *et al.* en el cual se reconocen los diferentes caminos, sugerencias y marcos de trabajo a seguir para reconocer e implementar pruebas en la ILPS, es de notar que los estudios encontrados evocan las validaciones como una prueba del progreso en su investigación y como parte fundamental del método científico, es decir que una validación es un paso importante a tener en cuenta en el proceso de minería de datos[58], y por consiguiente necesita ser automatizado como parte de todo el producto final.

Es necesario tomar de ejemplo la “tubería” de integración y entrega continua, descargar, construir, probar, desplegar y actualizar y saber que en cada dominio en donde se aplican las técnicas de MD será el mismo proceso y eso hará posible una automatización general, tenemos como ejemplo numerosas metodologías para aplicar las técnicas de minería de datos que independiente de que su contexto sea, análisis de imágenes[?], sistemas expertos[21] o líneas de productos [15], siempre imitan a KDD, en donde el componente de minería de datos es una parte que puede variar o innovar de acuerdo al dominio o el contexto en nuestro caso la construcción de modelos de características. Identificamos las mejoras y los enfoques novedosos en las técnicas de minería de datos y se propuso si pueden ser usadas en la ingeniería de líneas de producto o en técnicas similares[48].

RQ13: ¿Qué técnicas de minería pueden ser explotadas en la ingeniería de líneas de productos?

Para describir con facilidad que técnicas fueron usadas en la ILP según este trabajo de investigación, la Tabla 9.2 nos ayudará a obtener las publicaciones con las técnicas de minería de datos en el proceso de ILP. Además se presentan los hallazgos y se identifican las principales técnicas según el documento y su dominio.

Tabla 9.2: Resultados de la RQ13: ¿Qué técnicas de minería pueden ser explotadas en la ingeniería de líneas de productos?

Técnica	Dominio	Objetivo	Documento
<i>AI</i>	<i>SPL</i>	En una línea de productos de software los árboles de decisión y los modelos estadísticos, creados con nuevas técnicas de inteligencia artificial como enjambres, se pueden aplicar a cualquier otro dominio, para el diseño y la configuración de nuevos productos	[49, 70, 88]
<i>ANN</i>	<i>ERP</i>	Se estima una función usando algoritmos en WEKA en especial, <i>Artificial Neural Networks</i> para estimar costos en un <i>Enterprise Resource Planning</i> software.	[89]

Continúa en la siguiente página

Tabla 9.2 – *Continuación de la página anterior*

Técnica	Dominio	Objetivo	Documento
<i>asset mining</i>	<i>LEADT</i>	Con el propósito de localizar funciones en el código legado mediante LEAD (<i>Location, Expansion, And Documentation Tool</i>) las LPS se construyen con <i>Colored Integrated Development Environment</i> , para dirigir la atención de los desarrolladores en la implementación de las nuevas características del producto.	[90]
<i>Clustering</i>	<i>catalog services</i>	Facilitar el descubrimiento de la variabilidad de la LP mediante el agrupamiento de características en los productos legados.	[68]
<i>Clustering</i>	<i>Domain analysis</i>	Extraer características de los comentarios en los productos, incluyendo <i>GOOGLE Apps MarketPlace</i> , para posteriormente realizar un análisis del dominio, efectuar una comparación sobre la mejor manera de extracción y recomendar las mejores características.	[59]
<i>Clustering</i>	<i>KDD</i>	Ayudar a desarrollar la nueva generación de software <i>e-learning</i> , considerando el proceso de la extracción de conocimiento en bases de datos, con ayuda de Weka en la etapa de derivación de productos. En el proceso se busca evitar errores humanos, asegurar la calidad del producto y reducir los costos asociados a la fijación de estos errores humanos.	[65]
<i>DR</i>	Telefonía	Reducir el tamaño de los modelos de características, dado que cuando se tiene un espacio de solución muy amplio, verificar cada una de las soluciones puede convertirse en un problema polinomial no determinista, o en la minería de datos se discute como acabar con la maldición de la dimensionalidad mediante técnicas de reducción de dimensión .	[68]

Continúa en la siguiente página

Tabla 9.2 – Continuación de la página anterior

Técnica	Dominio	Objetivo	Documento
<i>DSL</i>	FODA	Traducir un modelo de características a una especificación mas formal ,para resolver los problemas heredados por el lenguaje humano. Mediante <i>Domain specific language</i> yse puede estructurar el lenguaje para que soporte operaciones y definiciones basadas en Java.	[91]
<i>ELM</i>	<i>YouTube</i>	Se aplican algoritmos de aprendizaje estadístico entre ellos <i>extreme machine learning</i> para hacer un análisis de sentimientos sobre una bases de datos de <i>YouTube</i> . <i>Enterprise Resource Planning</i> software.	[89]
<i>Feature selection</i>	NFP	Se presenta un <i>Systematic review</i> para adquirir las propiedades no funcionales en la ingeniería de líneas de producto de software .	[92]
<i>Full fledged</i>	BOM	Sistema para automatizar el mapeo de la cantidad de partes en el <i>Bill of Material</i> .	[92]
<i>Fuzzy clustering</i>	Cátalogo de productos	Identificar, evaluar y seleccionar los productos en el portafolio.	[31]
<i>Fuzzy modeling</i>	Diseño de productos	La presentación formal de las características técnicas de los prductos, centrado la atención en las necesidades del cliente.	[28]
<i>Lazy and eager</i>	Linux	Miniminar esfuerzos en la configuración de los modelos de características y encontrar cambios en el software con respecto al tiempo.	[56, 64]
<i>Neural networks</i>	Manufactura	Revelar las aplicaciones de la minería de datos en la manufactura.	[14]
<i>Neural networks</i>	Cracking	Comparar los diferentes enfoques de la minería de datos, para resolver un problema de clasificación en el proceso de craqueo catalítico del liquido de una refinería.	[93]

Continúa en la siguiente página

Tabla 9.2 – Continuación de la página anterior

Técnica	Dominio	Objetivo	Documento
<i>Neural linguistic programing</i>	Lucha contra el crimen	Dar a conocer la complicada labor de clasificar los datos de los criminales, una labor sin precedente en la que se recuperan grandes cantidades de datos y se evidencia en un estudio sistemático.	[94]
<i>Sentiment analisis</i>	<i>Social media data</i>	Se investigan dos dominios: celulares y automóviles con el objetivo de recolectar datos sobre las preferencias de manera oportuna y rentable.	[73]
<i>v-algorithm</i>	<i>Software</i>	Se define la línea de proceso de software para dos dominios de las cámaras de video y el desarrollo de software, para identificar maneras alternativas de realizar ciertas actividades que no eran tomadas en consideración cuando se desarrollaba la fase de especificación inicial del modelo en la línea de proceso de software.	[95]

El uso de las técnicas de minería de datos para la extracción de características y construcción de líneas de producto esta enmarcado como una de muchas técnicas que se puede usar para la creación de líneas de producto[96], en este caso, modelos de características[97]. Sin embargo, hay nuevos métodos detrás de esta nueva generación de líneas de productos, su administración, validación, adopción, desarrollo, análisis y sus futuras direcciones y propuestas[98], por esta razón se propuso la siguiente pregunta.

RQ14: ¿Entre los documentos encontrados son evidentes las propuestas innovadoras en la ingeniería de línea de productos? Tradicionalmente, la minería de datos ha sido usada por los estadísticos, los analistas de datos y los científicos con Ph.D.'s en el aprendizaje de la máquina[13], la minería de datos es ahora un punto de innovación relevante, esta tecnología es más accesible a un público más amplio y por esta razón fue importante incorporar las nuevas técnicas en el proceso de la ILP.

En *Using fuzzy modeling for consistent definitions of product qualities in requirements*[28] Jean-Marc Davril, *et al.* parten desde el dominio de la construcción de cámaras fotográficas para presentar el enfoque que diez años atrás Klaus Pohl y Thorsten Weyer[76] habia propuesto: que desde una matriz de configuración de productos y la documentación del dominio, se pueden construir las dependencias de las características de los productos, luego mediante la metodología *fuzzy modeling* se logra construir los modelos de características centrados en la necesidad del cliente para posteriormente facilitar la búsqueda de los productos por las características mas atractivas. Similar a estos trabajos se encontraron los de Jae Kwon Bae, *et al.*[6], orientados al dominio de las ventas por menudeo, que complementando con ingeniería inversa dan las primeras pistas sobre una posible aplicación de un motor de las recomendaciones o sugerencias que configure

automáticamente productos o servicios basados en las preferencias de los consumidores. Anteriormente se hacia referencia a que la mayor ventaja de la ILP debe ser la construcción automática de los productos de acuerdo a la especificación de unas características. Ahora se propone la entrega de valor mediante productos a la medida. Como ejemplo presentamos que en *Epsilon Generation Language*[65], se puede proponer elementos nuevos de acuerdo a las características que sugieran los datos de los usuarios, es decir adaptar el producto a los requerimientos de un usuario puede ahorrar muchos costos asociados a los errores humanos en el proceso de levantamiento de requisitos, en lugar de realizar los desarrollos mediante código generado por platillas. Si los lenguajes de especificación de características son definidos como reglas entendibles por una maquina, y son estandarizados, se tendría un proceso interesante.

Dentro del Análisis de Dominio Orientado a Característica (de su siglas en ingles FODA), el modelo de características es un artefacto ampliamente utilizado para expresar los puntos en común y la variabilidad de todos los productos de la línea de productos en términos de características y sus relaciones[19].

Una “característica” se define como un aspecto prominente o distintivo, el cual es visible por el usuario en un sistema de software. La selección de características es un paso esencial para derivar un producto individual con requerimientos funcionales específicos (RFs) que satisfacen algunos requisitos no funcionales (RNFs). El algoritmo IVEA[19] se encuentra novedoso por seleccionar características que satisfacen los RFs y que buscan la optimización de múltiples RNFs. Se puede utilizar para cobijar las preferencia del usuario y mejorar la conformidad de las soluciones.

En *Sample Spaces and Feature Models: There and Back Again*[66], Krzysztof Czarnecki, *et al.* exponen varios algoritmos de extracción de datos bien conocidos (*Conjunctive association rule mining & Disjunctive association rule mining*) que dan origen a los modelos de características probabilistas, en donde se especifica la probabilidad de contener o no una característica. Las aplicaciones de la minería de datos para la implementación de sistemas expertos[31], el diseño dirigido por técnicas de agrupamiento[68], las redes neuronales con tenseones (TensorFlow), los campos aleatorios de markov y el aprendizaje relacional para la creación de bases de conocimiento[3], los modelos de *elastic net*[52] y la extracción de características mediante clasificación binaria[99], el descubrimiento de patrones por medio de gráficos[100], el análisis de redes sociales[100], la programación inductiva y la minería de datos relacional[101], los algoritmos de canibalización de características o los enfoques para planear líneas de producto mediante modelos dinámicos de estado variable[5], son enfoques que dentro de un marco de trabajo[67] para LP con minería de modelos de características se encontraron novedosos.

Por otra parte el afinamiento de los modelos de características[69, 102], le da al modelamiento de la variabilidad las herramientas suficientes para aceptar los retos en otros dominios[103], la alta dimensionalidad de los datos puede no ser un problema, pues si se encuentran los valores óptimos en cuanto a la cantidad de características, ya es transparente el método de minería a usar[69]. Las herramientas de modelado[24, 25] pueden consumir estos modelos de características y generar mejores soluciones a la configuración y derivación de productos. Sin embargo hay herramientas[102] que aún no han sido explotadas en todos los dominios, dejando un camino incompleto a la estandarización de un formato o una herramienta para el modelado de la ingeniería de líneas de producto.

9.1.2. Estrategia de búsqueda

Se realizo un proceso de búsqueda manual y automático en las bases de datos de la tabla 9.3 y se obtuvieron en total 397805 documentos. En la tabla 9.3 se presentan los primeros hallazgos y

se identifican las principales fuentes de información según la cantidad de resultados en la búsqueda de nuestro concepto principal.

Tabla 9.3: Resultados de la búsqueda manual y automática sobre publicaciones relacionadas con las técnicas de Minería de Datos

Cadena de busqueda	Base de Datos	Resultados	Fecha
“data mining”	Springel	134,419	15 de enero de 2015
“data mining”	Scopus	92,904	15 de enero de 2015
“data mining”	IEEE	90,094	15 de enero de 2015
“data mining”	Web of Science	72,457	15 de enero de 2015
“data mining”	Science Direct	37,615	15 de enero de 2015
“data mining”	Wiley Online Library	19,475	15 de enero de 2015
“data mining”	Current Contents	14,128	15 de enero de 2015
“data mining”	ACM DL	10,494	15 de enero de 2015
“data mining”	DOAJ	4,501	15 de enero de 2015
“data mining”	Aps Journals	2,742	15 de enero de 2015
“data mining”	Esmerald	2,340	15 de enero de 2015
“minería de datos”	Ebsco	1,871	15 de enero de 2015
“data mining”	Cambridge	1,605	15 de enero de 2015
“data mining”	ACS Publications	1,323	15 de enero de 2015
“data mining”	Dialnet	508	15 de enero de 2015
“data mining”	ASME	475	20 de enero de 2015
“minería de datos”	Dialnet	270	20 de enero de 2015
“data mining”	Access Engineering	264	20 de enero de 2015
“data mining”	sciELO	142	20 de enero de 2015
“data mining”	ASTM	107	20 de enero de 2015
“minería de datos”	DOAJ	90	20 de enero de 2015
“minería de datos”	Science Direct	40	20 de enero de 2015
“minería de datos”	Scopus	35	20 de enero de 2015

En la tabla 9.3 se proveen los resultados resumidos de nuestro concepto principal en los diferentes indexadores de información científica. Aproximadamente el podium de esta tabla es liderada por los lugares más sobresalientes, en donde es recomendable indexar un futuro documento. El objetivo de esta sección es establecer las diferencias que caracterizaron esta investigación en términos de requerimientos en la estrategia de búsqueda, calidad de la evaluación de los documentos, el proceso de búsqueda y sus resultados, las metas de este mapeo sistemático de la literatura partieron desde el desarrollo de los modelos de características mediante técnicas de minería de datos y los retos que sobresalen en la unión de la ingeniería de líneas de producto con el concepto principal en el que giraron las facetas, esta información acá contenida resume las mejores prácticas [104] y propone una guía de desarrollo que no está exenta de actualización, compromiso de la sección de trabajos futuros.

Se seleccionó *Scopus* como lugar de búsqueda por sus extensibles servicios web y la cantidad de documentos que se obtuvieron. La cadena de búsqueda definida fue:

Cadena de búsqueda para Scopus 1. *TITLE-ABS-KEY("feature model") AND PUBYEAR >2000 AND (LIMIT-TO (EXACTKEYWORD, "Product Line Engineering") OR LIMIT-TO (EXACTKEYWORD, "Support Vector Machines") OR LIMIT-TO (EXACTKEYWORD, "Data Mining") OR LIMIT-TO (EXACTKEYWORD, "Variability Modeling") OR LIMIT-TO (EXACTKEYWORD, "Feature Selection") OR LIMIT-TO (EXACTKEYWORD, "Software Product Line (SPLs)") OR LIMIT-TO (EXACTKEYWORD, "Neural Networks") OR LIMIT-TO (EXACTKEYWORD, "Clustering Algorithms") OR LIMIT-TO (EXACTKEYWORD, "Feature Diagrams"))*

Después de definir una búsqueda automática se complementó con una búsqueda manual[39]. Más adelante los documentos fueron categorizados de forma empírica como: *i* Azul - Estudios exploratorios que presentaron un enfoque medianamente interesante para el caso, los cuales son documentos que abordan parcialmente el tema. *ii* Gris - Documentos que presentaron conceptos y definiciones que son básicas para entender el ‘*technical know-how*’ del dominio y los ejemplos que se desarrollarían en la investigación, *iii* Naranja - Documentos con el contenido de los dominios de la ILP y la MD, en estos documentos se habla mucho sobre selección de características o MC. *iv* Personal - Son documentos que se recomendaron para su estudio, basados en la evaluación de los modelos de MD y sus desarrollos como procesos de negocio. *v* Púrpura - En este filtro se ubicaron los documentos con las propuestas más innovadoras o los documentos de opinión y futuros desarrollos, también se consideraron algunos enfoques que son interesantes pero aún muy complejos de desarrollar. *vi* Rojo - Acá se ubicaron los documentos que enmarcaban un caso práctico en una industria de renombre, también se contemplaron los MC en la ILPS. *vii* Verde - Son documentos que se usaron como ejemplo para desarrollar el mapeo sistemático de los estudios en la literatura. Aplicando los criterios de exclusión y revisión, que se establecieron en la metodología, sobre los documentos encontrados en la tabla 9.3, se seleccionaron 129 documentos que se consideraron significativos, luego, fueron analizados teniendo en cuenta a He Zhang, *et al.*[1] y se eligieron finalmente 119 documentos.

Para el proceso de gestión de los documentos: búsqueda, colección, selección, organización y colaboración se usó el software *Mendely*, luego se exportaron los registros de la colección en el formato RIS (*Research Information Systems*) para indexarlos con *Splunk* y generar las gráficas y las estadísticas de este trabajo de investigación.

9.1.3. Escaneo de artículos

Luego de obtener una muestra de los artículos que son significativos para el trabajo de investigación, se leen los títulos y las palabras claves de los resultados de cada una de las búsquedas en la tabla 9.3, y siguiendo el detalle de la metodología en la sección de metodología en la página 25, cuando se encontraba un documento interesante, este era coleccionado en *Mendeley* como se muestra en la Figura 9.5.

Es importante que la información, es decir los metadatos que acompaña al documento, sea ingresada de forma rápida y eficaz, en este trabajo de investigación se usó el DOI (*Digital Object Identifier*) para consultar los metadatos de un documento y para poder almacenarlo correctamente con el formato RIS en la aplicación *Mendeley* y más adelante en la aplicación *library* como se muestra en la Figura 9.6.

Con este proceso en mente, cada vez que se encontró un documento de interés se pasaba por las diferentes aplicaciones, considerando el DOI como la clave primaria por hacer una analogía y como



Figura 9.5: Captura tomada de la aplicación *Mendely*, en donde se presenta la selección de un documento con sus etiquetas.

dato principal para el seguimiento de los documentos y el control de los documentos repetidos. Para complementar el proceso de escaneo de artículos se usó el *script* en *Python* 9.1.

Listing 9.1: El Snippet de código “doigenerator.py” se usa para consultar las citas generales de un documento mediante el *webservices* de *Scopus*.

```

1 import requests
2 import time
3
4 file =open ("//Applications/Splunk/etc/
5           apps/library/info/sms.ris", "r")
6
7 for line in file:
8     if line.find('DOI-') != -1:
9         r=requests.get("https://api.elsevier.com/content/abstract/
10           citations?doi="+line[6:]+
11           "&apiKey=7f59af901d2d86f78a1fd60c1bf9426a
12           &httpAccept=application%2Fjson")
13         data=r.json()
14         print data
15         time.sleep(1)

```

En el *script* 9.1 se observa que en la línea 4 se hace uso de un archivo ³ que posteriormente en el ciclo es consultado en el campo del DOI, este parámetro se pasa al *webservices* en la línea 9 junto con un *token* de autenticación que nos permitió acceder a los metadatos que se necesitaban para complementar la información en la aplicación *library*. Para replicar este ejercicio será necesario remitirse a la página de *Elseviery* obtener un nuevo *token* para la autenticación y uso del *webservices*.

9.1.4. Análisis de palabras claves

Como se mostró en la Figura 9.2 en el *podium* encontramos, 15 documentos con la palabra clave: Data mining, 9 documentos con la palabra clave: Feature selection y 5 documentos con la palabra

³  sms.ris

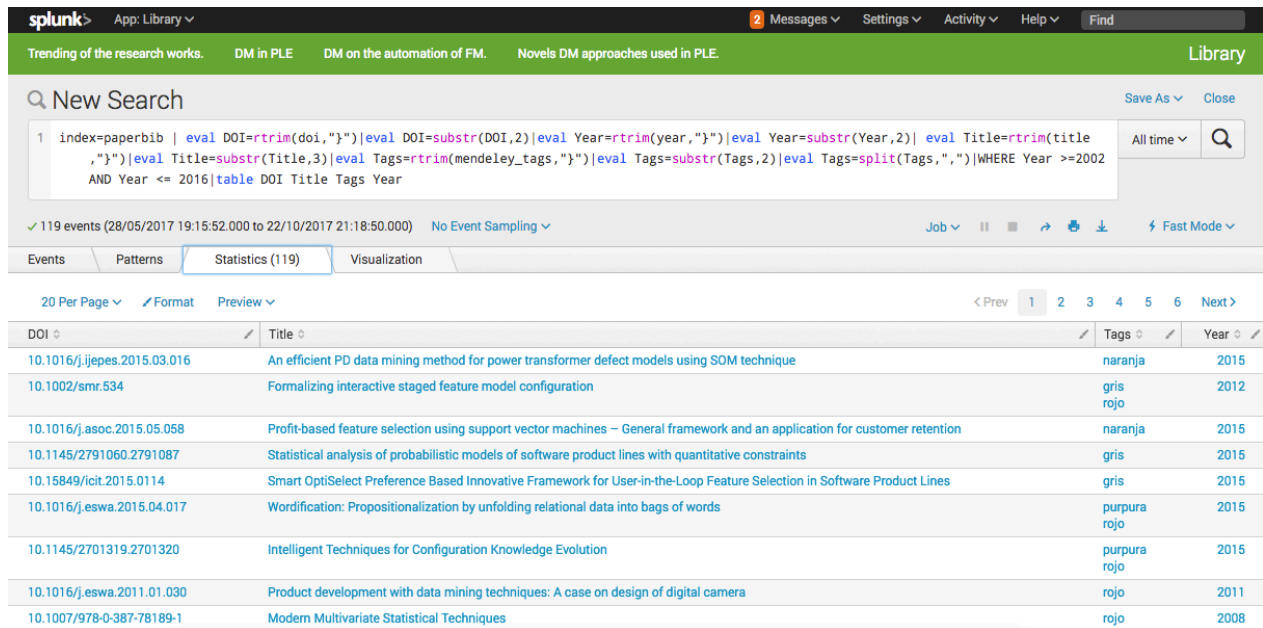


Figura 9.6: Captura tomada de la aplicación *library*, en donde se presentan las formas de filtrar los datos y presentar los reportes personalizados.

clave: Software product lines estos resultados indicaron que en este trabajo de investigación las palabras claves de nuestro estudio estaban muy orientadas por el titulo, y al cambiar la tecnología de indexación era necesario definir palabras adicionales que propongan un contenido que no se encuentre ni en el titulo, ni en el *abstract*, por consiguiente se definio que las palabras claves que se tendrán en cuenta para productos derivados de esta investigación serán: *Product line engineering*, *PMML*, *RIPS*, *Java*, *Phyton scikit-learn*. Considerando estas palabras se pueden proponer la siguiente generación de cadenas de busqueda en los trabajos futuros.

9.1.5. Proceso de extracción

El proceso de extracción de los datos esta dado por la aplicación *library*, en este proceso se usaron expresiones regulares.

Listing 9.2: Ejemplo de regex en Splunk

```

1  rex field=_raw "A1\s*\-\s*(?P<autor>.*?)\n" max_match=10|
2  rex field=_raw "KW\s*\-\s*(?P<palabra_clave>.*?)\n" max_match=20 |
3  rex field=_raw "JF\s*\-\s*(?P<journal>.*?)\n" |
4  rex field=_raw "TY\s*\-\s*(?P<tipo_documento>.*?)\n" |
5  rex field=_raw "Y1\s*\-\s*(?P<fecha>.*?)\n/" |
6  rex field=_raw "T1\s*\-\s*(?P<titulo>.*?)\n" |
7  top palabra_clave

```

En el codigo 9.2 tomamos en cuenta unas tuberias (|) para recoger todos los campos de nuestro archivo *sms.ris* y luego hacer el *podium* de las palabras claves.

9.2. Diseño del prototipo

En el diseño de este trabajo de investigación se debía considerar usar una base de datos que no tuviera tantos registros pero a su vez fuese popular y muy interpretable, se decidió usar IRIS, el *dataset* de iris está disponible en la web del repositorio de aprendizaje de maquinas para todo el publico y los conjuntos de datos que se encuentran en esta página son usados frecuentemente en la literatura[2, 10, 29, 105, 106]. Los datos en IRIS contiene tres clases de flores que necesitan ser clasificadas en *Iris Setosa*, *Iris Versicolour* & *Iris Virginica*, cada clase tiene 50 registros, en donde los atributos de clasificación están dados por *sepal length in cm*, *sepal width in cm*, *petal length in cm* & *petal width in cm*.

Para el pre procesamiento y el entendimiento de los datos no se realizo un esfuerzo significativo, los datos vienen de tal manera que es muy cómodo para un científico de datos el trabajo con ellos. por está razón, se ubica en el protagonista del desarrollo, el modelo generado en el estándar PMML, el cual el grupo de minería de datos (DMG) está orgulloso de presentar como un lenguaje de marcas de modelos predictivos (*Predictive Model Markup Language*), PMML usa XML, el lenguaje de marcas extensible, para representar la estructura de los modelos en la MD, la estructura general de este estándar puede consultarse en la página WEB del *data mining group*. Con este estándar tan bien definido e inexplorado por las aplicaciones en el ecosistema de los grandes datos, se recomienda que era la piedra angular y la API para el desarrollo de un código en Java, que fuese capaz de traducir las restricciones en el formato simple para modelos de características, *Simple XML Feature Model format*(SXFm)[26], y luego en trabajos futuros proponer una arquitectura mucho mas completa, con mas sabores de formatos. La mayor ventaja al trabajar con SXFM es que su estructura es muy simplificada, de esta manera se implemento el flujo de trabajo presente en el modelo de la Figura 9.7. donde se parte de CRISP[13, 14, 107] y se termina con la creación de el modelo de MD en XMML.

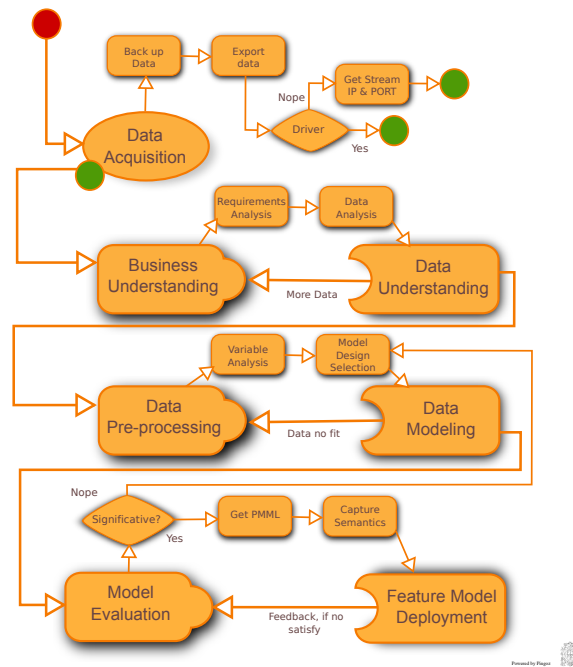


Figura 9.7: Captura tomada de la aplicación *Mendely*, en donde se presenta la selección de un documento con sus etiquetas.

Con en el modelo de la Figura 9.7 se tuvieron en cuenta los archivos⁴, para el desarrollo del prototipo, se implemento una prueba orientada a la creación de un archivo, luego este archivo se construyo en formato SXFM incorporando la información del modelo de MD.

9.3. Implementación del prototipo(Caso de aplicación)

El detalle del proceso del desarrollo de modelos de características se expresa entonces en la Figura 9.7, este proceso iniciará desde el punto rojo, con la adquisición de los datos, para nuestro caso particular se obtuvo una base de datos de los registros de atención individual de prestaciones en salud (RIPS), se toma solo el 0,01 % de la base de datos de RIPS para desarrollar el ejemplo. La entrega fue hecha mediante medios magnéticos y esta base de datos no estará disponible sin el consentimiento de la facultad de salud pública.

Se realizaron varias copias de la información, se uso *google drive* para realizar un respaldo en la nube y luego aprovisionar de esa copia a las maquinas en las cuales se estaba trabajando, la base de datos completa pesa aproximadamente 10 GB, y se dividió en archivos de 100 MB para su manipulación. En muchas ocasiones se puede presentar el escenario en donde los datos se encuentren en protocolos o infraestructuras diferentes a un simple archivo de texto como fue el caso que se desarrollo en este trabajo de investigación, sin embargo la comunidad ha creado *frameworks* poderosos para conectarse por medio de *drivers* o por un *socket* o protocolo.

El entendimiento del negocio de estos datos almacenados en columnas separadas por barras verticales o el símbolo de tubería (`|`), se realizo considerando que los datos eran suficientes para proponer una línea de productos de tratamientos médicos. Para completar el proceso de entendimiento de los datos se consulto en página WEB del ministerio de salud y en diferentes fuentes publicas de carácter oficial. Con *Phyton*⁵ se desarrollaron los *scripts* del pre procesamiento de los datos.

Listing 9.3: El Snipet de código “canopy.py” se usa para llevar los datos a una estructura de datos llamada “dataframe”, luego enmascarar los datos y filtrar los dato irrelevantes.



```
1 import numpy as np
2 import pandas as pd
3 from sklearn import preprocessing
4 infile = "/Users/josepplloo/Documents/scripts/xaa.csv"
5 infile2 = "/Users/josepplloo/Documents/scripts/xab.csv"
6 dictfile = "/Users/josepplloo/Documents/scripts/dictresult.csv"
7 df = pd.read_csv(infile, header = 0)
```

En el código 9.3, se importan las librerías para el pre procesamiento y se leen dos archivos `infile` e `infile2` que luego se concatenan, adicional se creó `dictfile` para la protección de los datos, este archivo guardó los valores originales, que por seguridad, funcionalidad y buenas practicas es muy común renombrar las categorías en este caso por números. Por ejemplo para el campo `SexoDesc` se tiene los valores: FEMENINO, MASCULINO, NR-NO REPORTADO, NO DEFINIDO y se sustituyeron por los valores 0,1,3,2 respectivamente.

Listing 9.4: El Snipet de código “canopy.py” se usa para llevar los datos a una estructura de datos llamada “dataframe”, luego enmascarar los datos y filtrar los dato irrelevantes.

```
1 df = pd.read_csv(infile, header = 0)
```

⁴  iris.pmm  iris.sxfrm

⁵  underbrush.py  canopy.py

```

2 df2 = pd.read_csv(infile2, header = 0)
3 a=[]
4 frames = [df, df2]
5 result = pd.concat(frames)
6 #cargo el archivo de datos
7 l1 = preprocessing.LabelEncoder()
8 l2 = preprocessing.LabelEncoder()
9 l3 = preprocessing.LabelEncoder()
10 l4 = preprocessing.LabelEncoder()
11 l5 = preprocessing.LabelEncoder()
12 l6 = preprocessing.LabelEncoder()
13
14
15 #Guardo los encabezados de la tabla y los transformo
16 a.append("PersonaID: id de la persona")
17 a.append(result.PersonaID.unique())
18 result=result.drop('PersonaID',1)

```

Continuamos con la segunda parte del mismo archivo canopy.py 9.4 en donde se desarrollo la codificación de las etiquetas o columnas. luego en la línea 18, a pesar de guardar la información sobre los usuarios en el archivo de diccionario, se decide borrar estos datos por no ser relevantes, este mismo ejercicio con la columna PersonaID se realizo para las columnas, TipoEventoRIPSDesc, Codigo, DxEgreso, FinalidadConsultaCD, CausaExternaCD, Prestador, MunicipioCD, EstadoSalidaDesc, CostoConsulta, CostoProcedimiento, NetoAPagarConsulta, NumeroDiasEstancia, fechaaid, dado que estos datos se encontraban constantes o no ofrecían información relevante para el prototipo o la prueba de concepto. En la porción de código 9.5, se presenta la forma en la que se considero y se filtraron los datos.

Listing 9.5: El Snipet de código “canopy.py” se usa para llevar los datos a una estructura de datos llamada “dataframe”, luego enmascarar los datos y filtrar los dato irrelevantes.

```

1 a.append("RegimenAdministradoraDesc: id del regimen:
2 Contributivo-Subsidiado-Vinculado-Particular-Otro-Desplazado")
3 a.append(result.RegimenAdministradoraDesc.unique())
4 l2.fit(result.RegimenAdministradoraDesc.unique())
5 a.append(l2.transform(result.RegimenAdministradoraDesc.unique()))
6 result.RegimenAdministradoraDesc=
7 l2.transform(result.RegimenAdministradoraDesc)
8
9 a.append("DxPrincipal: diagnostico principal")
10 a.append(result.DxPrincipal.unique())
11 l3.fit(result.DxPrincipal.unique())
12 a.append(l3.transform(result.DxPrincipal.unique()))
13 result.DxPrincipal=l3.transform(result.DxPrincipal)
14 a.append("tome el diagnostico G800")
15 a.append(l3.transform(["G800"]))

```

En la línea 3 se establecen cuales son los elementos únicos para luego generar tantas mascaras para las etiquetas como sean necesarias. En la línea 14 se toman solo los datos con el diagnostico G800 tomado de forma arbitraria. De esta manera se pudo controlar y filtrar la información para

que el software estadístico la pudiera consumir. En la elección del diagnostico se tuvo la necesidad constante de estar encontrando un diagnóstico diferente, para hacer pruebas en el futuro, por esta razón se desarrollo el código 9.6.

Listing 9.6: El Snipet de código “underbrush.py” se usa para filtrar un campo en una estructura de datos llamada “dataframe”.

```
1 import numpy as np
2 import pandas as pd
3 infile = "/Users/josepplloo/Documents/scripts/xab_clean.csv"
4 df = pd.read_csv(infile, header = 0)
5 for aux in df.DxPrincipal.unique():
6     print aux
```

Con este código se aprovechó que se tenía un conjunto de archivos que representa cada uno el 0,01 % de los datos, y se usó para buscar las mejores muestras de los diagnósticos, esto significó que se tuvieron en cuenta los diagnósticos con una población de datos de no mas de 24 códigos de procedimientos médicos, debido a que el software **R** no puede en este caso, realizar una clasificación de esta dimensión (mayor a 24 dimensiones).

Presentamos la Tabla 9.4 en donde mostramos las variables o las columnas de la base de datos.

Tabla 9.4: Resultados del análisis de las variables.

Nombre de la variable	Descripción	Usada
PersonaID	ID de la persona.	NO
TipoEventoRIPSDesc	Evento en la Factura o lá orden médica	NO
Codigo:	id de la eps.	
RegimenAdministradoraDesc	ID del régimen puede ser Contributivo, Subsidiado, Vinculado, Particular, Otro ó Desplazado.	SI
DxPrincipal*	Diagnóstico principal.	SI
DxEgreso	Diagnóstico de salida.	NO
FinalidadProcedimientosCD	Si la finalidad es diagnostica ó terapéutica.	SI
FinalidadConsultaCD	Descripción de la consulta que se realiza	NO
TipoUsuarioCD	Si el usuario es de tipo: Contributivo, Subsidiado, Vinculado-Particular, Otro ó Desplazado.	SI
CausaExternaCD	Si es victima de maltrato o violencia.	NO
Prestador	ISP	NO
AmbitosProcedimientoCD	Si es ambulatorio, hospitalario ó urgencias.	SI
CodigoProcedimiento	Procedimiento médico.	SI
MunicipioCD	ID municipio.	NO
EstadoSalidaDesc	Si sale vivo o muerto.	NO
CostoConsulta	El costo de la consulta.	NO
CostoProcedimiento	El costo del procedimiento.	NO
NetoAPagarConsulta	Total a pagar.	NO
NumeroDiasEstancia	Número de días en el servicio.	NO
fechaid	Fecha.	NO
Edad	Edad.	SI
SexoDesc	Genero del usuario.	SI

En la Tabla 9.4, se elijen las variables dada la necesidad del problema de clasificación: Dado un diagnóstico principal, se desean predecir los códigos de procedimiento, según el género, la edad, el tipo de usuario, la finalidad del procedimiento, el diagnostico principal y el régimen de salud, en este problema los datos que no fueron usados presentaban un comportamiento constante en la muestra de datos de entrenamiento y pruebas. Siguiendo las recomendaciones en scikit-learn.org se tenían mas de 50 datos, se necesitaba predecir una categoría y teníamos datos etiquetados, lo que indicó un problema de clasificación. Esta discusión sobre el mejor clasificador, como orientar la solución de un problema de clasificación o que algoritmo usar[2, 10, 39] es interesante, más no un punto importante en los resultados y el análisis de este trabajo de investigación, no significa que no se deba tener en cuenta. Por otra parte, cualquier problema de la minería de datos en cualquier dominio y contexto se puede llevar a cabo y no es ajeno a este trabajo de investigación, en donde el verdadero protagonista es la generación del archivo PMML, la cual es independiente de cada modelo y para el cual se propusieron dos ejemplos desarrollados en **R**⁶, uno para árboles de clasificación y otro para bosques aleatorios.

La codificación de los algoritmos en **R** se realizo muy orientada a la creación del archivo PMML como se muestra en el código 9.7, en donde se presentan la implementación de árboles y bosques. Finalmente se crea un PMML con los resultados de bosques aleatorios.

Listing 9.7: El Snipet de código “RandomForest.R”se usa para desarrollar el problema de clasificación en **R**.

```

1 library(readr)
2 xaa_0 <- read_csv(
3   "~/Documents/scripts/RIPS_2013_1/splitDx/1499.csv",
4   col_types = cols(
5     DxPrincipal=col_skip(),
6     RegimenAdministradoraDesc=col_factor(c("3", "2")),
7     FinalidadProcedimientosCD=col_factor(c("1","2")),
8     TipoUsuarioCD=col_factor(c("2","1")),
9     AmbitosProcedimientoCD=col_factor(c("2","0","1")),
10   CodigoProcedimiento=col_factor(c(3345, 3302, 3335,
11      3329, 2086, 2340, 3337, 3330, 3307, 2112, 3331,
12      2806, 2362, 2588, 2201, 1847, 2093, 3125, 155,
13      1820, 1819, 3332, 3358, 3305)),
14    Edad=col_integer(),
15    SexoDesc=col_factor(c("0","1"))))
16 xaa<-na.omit(xaa_0)
17 xaa<-data.frame(xaa)
18 set.seed(2)
19 traintest <-sample (1: nrow(xaa), 0.7*nrow(xaa))
20 xaa.train <-xaa[traintest,]
21 xaa.test <- xaa[-traintest,]

```

Con la libreria **readr** se pudieron leer los datos pre procesados por **canopy.py** separados por comas (csv) en la línea 1, en las siguientes líneas se definieron los factores, se omitieron los datos nulos, al igual que el **DxPrincipal** y se almaceno en un **dataframe**, en la línea 18 se establece la semilla de generación para garantizar que las construcciones en el algoritmo no muten con cada ejecución, sin esta semilla no se podría entrenar correctamente; en las líneas siguientes se dividen

⁶  ClasificationTree.py  RandomForest.R

los datos para entrenamiento y pruebas.

En el código 9.8 la elección de la librería para implementar los árboles de clasificación fue **rpart**.

Listing 9.8: El Snipet de código “RandomForest.R” se usa para desarrollar el problema de clasificación en **R**.

```
1 library(rpart)
2 tree.model <-rpart(CodigoProcedimiento ~., data=xaa.train)
3 print("###Tree###")
4 plot(tree.model)
5 text(tree.model)
6 title("Training_Set's_Classification_Tree")
7 print(tree.model)
8 print("###Summary###")
9 print(summary(tree.model))
10 print("###prediction###")
11 tree.pred<-predict(tree.model,xaa.test,type="class")
12 print(table(xaa.test$CodigoProcedimiento,tree.pred))
13 print("###prune_tree###")
14 tree.prune<-prune(tree.model, cp=0.02)
15 print(summary(tree.prune))
16 plot(tree.prune)
17 text(tree.prune)
18 title("Training_Set's_Classification_Prune_Tree")
19 library(partykit)
20 tree.rparty<-as.party(tree.prune)
21 print("###prune_tree_partykit###")
22 print(summary(tree.rparty))
23 plot(tree.rparty)
```

Este *snipet* de código tomó los datos de entrenamiento en la línea 2, luego imprime los resultados y entrena en la línea 11, finalmente se poda. En este mismo código se desarrollo bosques aleatorios, un mejor enfoque para afrontar los problemas de clasificación[10, 39], como se muestra en el *snipet* de código 9.9.

Listing 9.9: El Snipet de código “RandomForest.R” se usa para desarrollar el problema de clasificación en **R**.

```
1 library("randomForest")
2 rf<-randomForest(CodigoProcedimiento~.,data=xaa,
3                 do.trace=100, ntree=100)
4 print(rf)
5 library(pmml)
6 tree.xml<-pmml(rf)
7 saveXML(tree.xml,file=
8         "~/Documents/scripts/RIPStreeDx1219.xml")
```

En el código 9.9 se presentan de árboles aleatorios, con la librería **randomForest** en la línea 1, desde la línea 4 se usó la librería que permite convertir los modelos de **R** en XML y luego en formato PMML. Por otra parte recomendamos generar los árboles obtenidos de los modelos en **R** como se muestra en la figura 9.8.

Es de notar que como árboles aleatorios no es interpretable, se elije un árbol (el que mas se ajuste a los datos) en el bosque para realizar una gráfica.

RIPS Classification Tree

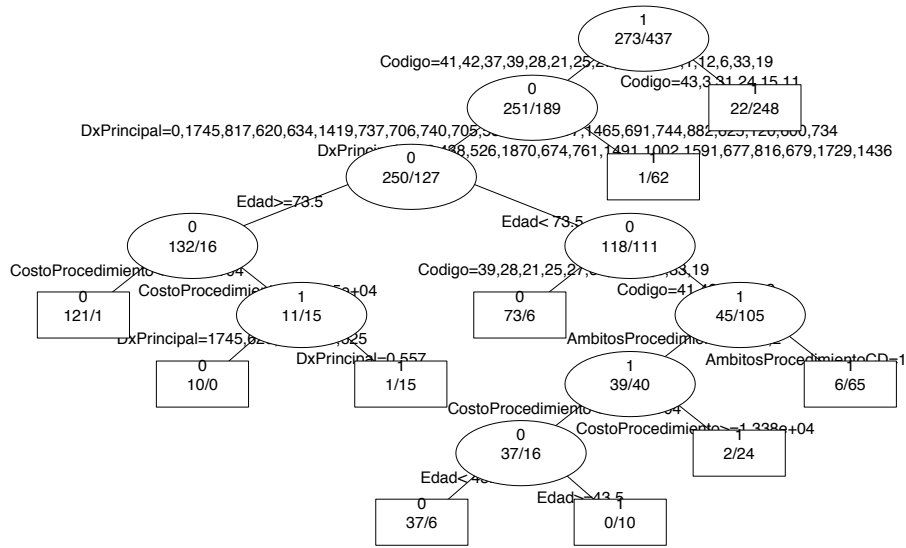


Figura 9.8: Modelos generados desde R en el código RandomForest.R.

Con el insumo para el paso final en el modelo 9.7, se construyó el código 9.10, en donde se toma como parámetro el PMML para generar el SXFM.

Listing 9.10: El Snipet de código “ReadPMMLtoSXFm.java”se usa para desarrollar el problema de clasificación en **R**.

```
1 public class ReadPMMLtoSXFM {
2
3     static Path pathy = FileSystems.getDefault().getPath("");
4     static String SXFMPATH = pathy.toAbsolutePath().toString()
5 + "/src/main/resources/sxfm/";
6     static String nombre_archivo = "RIPSMModel";
7     static ArrayList<String> key;
8     static ArrayList<String> value;
9     static String nombre_modelo = "";
10    static ArrayList<String> restricciones_modelo =
11    new ArrayList<String>();
12    static int cont = 0;
13    static boolean bandera = false;
```

En este código se almaceno en GIT por tratarse de un

Capítulo 10

Conclusiones y Trabajos Futuros

10.1. Conclusiones

En estos momentos, una gran cantidad de datos es almacenada en bases de datos y almacenes de datos. Los datos crecen rápidamente porque la información se guarda usando periféricos de computadora, códigos de barras, sensores y sistemas biométricos. También una gran variedad de datos que se consideraban difíciles de manejar o que estaban aislados ahora tienen fines prácticos, estos datos son grandes y complejos, con millones de registros y muchas variables. Además, diferentes agencias gubernamentales, instituciones educativas e industrias han acumulado estas grandes cantidades de datos, como ejemplo para el desarrollo de los objetivos del proyecto de investigación la Facultad Nacional de Salud Pública pone a disposición del Grupo de Investigación de Ingeniería y Tecnologías de las Organizaciones y de la Sociedad (ITOS) los Registros Individuales de Prestación de Servicios de Salud (RIPS) que se definen como el conjunto de datos mínimos y básicos que el Sistema General de Salud Social requiere para sus procesos cuya denominación, estructura y características se ha unificado y estandarizado para todas las entidades a que hace referencia el artículo segundo de la resolución 3374 de 2000 (las instituciones prestadoras de servicios de salud (IPS), de los profesionales independientes, o de los grupos de práctica profesional, las entidades administradoras de planes de beneficios y los organismos de dirección, vigilancia y control del SGSSS.) Al usar estudios sistemáticos se redujo el sesgo en la investigación[104], el resultado fue un estudio completo y replicable. Con los hallazgos realizados, se pudo seleccionar más fácilmente las técnicas de minería de datos que se aplicaron en la elaboración de los modelos de características. El SMS presentó un conjunto de documentos caracterizados y categorizados en una gran cantidad de dimensiones y características. El entendimiento del negocio fue muy humilde en el sentido de que eran datos muy interpretables

10.2. Trabajos futuros

Bibliografía

- [1] He Zhang, Muhammad Ali Babar, and Paolo Tell. Identifying relevant studies in software engineering. *Information and Software Technology*, 53(6):625–637, jun 2011.
- [2] Alan J. Izenman. *Modern Multivariate Statistical Techniques*, volume 64 of *Springer Texts in Statistics*. Springer New York, New York, NY, oct 2008.
- [3] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, 104(1):11–33, jan 2016.
- [4] Fatemeh Nemati Koutanaei, Hedieh Sajedi, and Mohammad Khanbabaei. A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, 27:11–23, nov 2015.
- [5] Chun-Yu Lin and Gül E. Okudan. Planning for multiple-generation product lines using dynamic variable state models with data input from similar products. *Expert Systems with Applications*, 40(6):2013–2022, may 2013.
- [6] Jae Kwon Bae and Jinhwa Kim. Product development with data mining techniques: A case on design of digital camera. *Expert Systems with Applications*, 38(8):9274–9280, aug 2011.
- [7] Ministerio de Tecnologías de la Información 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. page 177.
- [8] Diego Silva Ardila. Indicadores Básicos de Tenencia y Uso de Tecnologías de la Información y Comunicación en empresas 2013 Cifras Definitivas. Technical report, DANE, Bogotá, 2015.
- [9] Yuval Elovici and Dan Braha. A decision-theoretic approach to data mining. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 33(1):42–51, jan 2003.
- [10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 1 of *Springer Series in Statistics*. Springer New York, New York, NY, 2009.
- [11] Rafael Capilla, J. Bosch, and Kyo-Chul Kang. *Systems and Software Variability Management*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
- [12] Mohsen Asadi, Samaneh Soltani, Dragan Gasevic, Marek Hatala, and Ebrahim Bagheri. Toward automated feature model configuration with optimizing non-functional requirements. *Information and Software Technology*, 56(9):1144–1165, sep 2014.

- [13] Mark F. Hornick, Erik Marcadé, and Sunil Venkayala. Chapter 1. *Java Data Mining*, 1:1–4, 2009.
- [14] J. a. Harding, M. Shahbaz, Srinivas, and a. Kusiak. Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering*, 128(4):969, 2006.
- [15] David Benavides, Sergio Segura, and Antonio Ruiz-Cortés. Automated analysis of feature models 20 years later: A literature review. *Information Systems*, 35(6):615–636, sep 2010.
- [16] Timothy W Simpson and Jianxin Roger Jiao. *Advances in Product Family and Product Platform Design*. Springer New York, New York, NY, 2014.
- [17] Bo Zhang and Martin Becker. RECoVar: A solution framework towards reverse engineering variability. In *2013 4th International Workshop on Product Line Approaches in Software Engineering (PLEASE)*, pages 45–48. IEEE, may 2013.
- [18] F. Zhou and R. J. Jiao. Feature model augmentation with sentiment analysis for product line planning. In *2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1689–1693. IEEE, dec 2015.
- [19] Xiaoli Lian and Li 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)*, pages 191–200. IEEE, mar 2015.
- [20] Alcemir Rodrigues Santos, Raphael Pereira de Oliveira, and Eduardo Santana de Almeida. Strategies for consistency checking on software product lines. In *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering - EASE '15*, number Section 3, pages 1–14, New York, New York, USA, 2015. ACM Press.
- [21] Alexander Felfernig, Stefan Reiterer, Martin Stettinger, and Juha Tiihonen. Intelligent Techniques for Configuration Knowledge Evolution. *Proceedings of the Ninth International Workshop on Variability Modelling of Software-intensive Systems - VaMoS '15*, pages 51–58, 2015.
- [22] M. H. ter Beek, A. Legay, A. Lluch Lafuente, and A. Vandin. Statistical analysis of probabilistic models of software product lines with quantitative constraints. In *Proceedings of the 19th International Conference on Software Product Line - SPLC '15*, pages 11–15, New York, New York, USA, 2015. ACM Press.
- [23] R. Al-msie'deen, A.-D. Seriai, Marianne Huchard, Christelle Urtado, and Sylvain Vauttier. Mining features from the object-oriented source code of software variants by combining lexical and structural similarity. In *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, pages 586–593. IEEE, aug 2013.
- [24] Carla I M Bezerra, José Maria Monteiro, Rossana M C Andrade, and Lincoln S Rocha. Analyzing the Feature Models Maintainability over their Evolution Process. In *Proceedings of the Tenth International Workshop on Variability Modelling of Software-intensive Systems - VaMoS '16*, pages 17–24, New York, New York, USA, 2016. ACM Press.
- [25] Raúl Mazo, Juan C Muñoz-Fernández, Luisa Rincón, Camille Salinesi, and Gabriel Tamura. VariaMos. In *Proceedings of the 19th International Conference on Software Product Line - SPLC '15*, pages 374–379, New York, New York, USA, 2015. ACM Press.

- [26] Marcilio Mendonca, Moises Branco, and Donald Cowan. S.p.l.o.t.: Software product lines online tools. In *Proceedings of the 24th ACM SIGPLAN Conference Companion on Object Oriented Programming Systems Languages and Applications, OOPSLA '09*, pages 761–762, New York, NY, USA, 2009. ACM.
- [27] Thomas Thüm, Sven Apel, Christian Kästner, Ina Schaefer, and Gunter Saake. A Classification and Survey of Analysis Strategies for Software Product Lines. *ACM Computing Surveys*, 47(1):1–45, jun 2014.
- [28] Jean-Marc Davril, Maxime Cordy, Patrick Heymans, and Mathieu Acher. Using fuzzy modeling for consistent definitions of product qualities in requirements. In *2015 IEEE Second International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*, pages 1–8. IEEE, aug 2015.
- [29] Dewan Md Farid, Li Zhang, Chowdhury Mofizur Rahman, M.A. Hossain, and Rebecca Strachan. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, 41(4):1937–1946, mar 2014.
- [30] Hemanta Kumar Bhuyan and Narendra Kumar Kamila. Privacy preserving sub-feature selection in distributed data mining. *Applied Soft Computing*, 36:552–569, nov 2015.
- [31] Jianxin Jiao and Yiyang Zhang. Product portfolio identification based on association rule mining. *Computer-Aided Design*, 37(2):149–172, feb 2005.
- [32] Qian Chen, Wenhao Zhu, Chaoyou Ju, and Wu Zhang. Cross domain web information extraction with multi-level feature model. In *2014 10th International Conference on Natural Computation (ICNC)*, pages 780–784. IEEE, aug 2014.
- [33] Cheng Fan, Fu Xiao, and Shengwei Wang. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy*, 127:1–10, aug 2014.
- [34] Vijay Vaishnavi and Bill Kuechler. Design Science Research in Information Systems, 2013.
- [35] Diego Martín Alaimo. *Proyectos ágiles con #Scrum : flexibilidad, aprendizaje, innovación y colaboración en contextos complejos*. Kleer, BA, 1 edition, 2013.
- [36] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18, aug 2015.
- [37] C. Rolland, C. Ben Achour, C. Cauvet, J. Ralyté, a. Sutcliffe, N. Maiden, M. Jarke, P. Haumer, K. Pohl, E. Dubois, and P. Heymans. A proposal for a scenario classification framework. *Requirements Engineering*, 3(1):23–47, 1998.
- [38] Roel Wieringa, Neil Maiden, Nancy Mead, and Colette Rolland. Requirements engineering paper classification and evaluation criteria: A proposal and a discussion. *Requirements Engineering*, 11(1):102–107, 2006.
- [39] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, volume C, pages 161–168, New York, New York, USA, 2006. ACM Press.

- [40] Frank van der Linden, Klaus Schmid, and Eelco Rommes. *Software Product Lines in Action*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.
- [41] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14*, pages 1–10, New York, New York, USA, 2014. ACM Press.
- [42] Shuai Wang, Shaukat Ali, Arnaud Gotlieb, and Marius Liaaen. A systematic test case selection methodology for product lines: results and insights from an industrial case study. *Empirical Software Engineering*, 21(4):1586–1622, aug 2016.
- [43] Camille Salinesi, Colette Rolland, Daniel Diaz, and Raúl Mazo. Looking for Product Line Feature Models Defects: Towards a Systematic Classification of Verification Criteria. In *2009 17th IEEE International Requirements Engineering Conference*, pages 385–386. IEEE, aug 2009.
- [44] Amina Souag, Raúl Mazo, Camille Salinesi, and Isabelle Comyn-Wattiau. Reusable knowledge in security requirements engineering: a systematic mapping study. *Requirements Engineering*, 21(2):251–283, jun 2016.
- [45] Mário André de F. Farias, Renato Novais, Methanias Colaço Júnior, Luís Paulo da Silva Carvalho, Manoel Mendonça, and Rodrigo Oliveira Spínola. A systematic mapping study on mining software repositories. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16*, pages 1472–1479, New York, New York, USA, 2016. ACM Press.
- [46] Larissa Rocha Soares, Pasqualina Potena, Ivan Do Carmo Machado, Ivica Crnkovic, and Eduardo Santana D De Almeida. Analysis of non-functional properties in software product lines: A systematic review. *Proceedings - 40th Euromicro Conference Series on Software Engineering and Advanced Applications, SEAA 2014*, pages 328–335, 2014.
- [47] Samuel Sepúlveda, Ania Cravero, and Cristina Cachero. Requirements modeling languages for software product lines: A systematic literature review. *Information and Software Technology*, 69:16–36, jan 2016.
- [48] Jane D A S Eleutério, Felipe N Gaia, Genáina N Rodrigues, and Cecília M F Rubira. Dependable Dynamic Software Product Line—a Systematic Mapping Study. (April), 2015.
- [49] Noor Hasrina Bakar, Zarinah M. Kasirun, and Norsaremah Salleh. Feature extraction approaches from natural language requirements for reuse in software product lines: A systematic literature review. *Journal of Systems and Software*, 106:132–149, aug 2015.
- [50] Mathieu Acher, Philippe Collet, Philippe Lahire, and Robert France. Comparing Approaches to Implement Feature Model Composition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6138 LNCS, pages 3–19. 2010.
- [51] Jean Marc Davril, Patrick Heymans, Guillaume Bécan, and Mathieu Acher. On breaking the curse of dimensionality in reverse engineering feature models. *CEUR Workshop Proceedings*, 1453:19–22, 2015.
- [52] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, apr 2005.

- [53] Guanzhong Yang and Yaru Zhang. A feature-oriented modeling approach for embedded product line engineering. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 1607–1612. IEEE, aug 2015.
- [54] Wolfram Fenske and Sandro Schulze. Code Smells Revisited: A Variability Perspective. *Proceedings of the Ninth International Workshop on Variability Modelling of Software-intensive Systems*, pages 3:3—3:10, 2015.
- [55] Goetz Botterweck and Andreas Pleuss. Evolution of Software Product Lines. In *Evolving Software Systems*, volume 234, pages 265–295. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.
- [56] Valentin Rothberg, Nicolas Dintzner, Andreas Ziegler, and Daniel Lohmann. Feature Models in Linux. In *Proceedings of the Tenth International Workshop on Variability Modelling of Software-intensive Systems - VaMoS '16*, pages 65–72, New York, New York, USA, 2016. ACM Press.
- [57] Sebastián Maldonado, Richard Weber, and Fazel Famili. Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. *Information Sciences*, 286:228–246, dec 2014.
- [58] Shu-Hsien Liao, Chyuan-Meei Chen, and Chung-Hsin Wu. Mining customer knowledge for product line and brand extension in retailing. *Expert Systems with Applications*, 34(3):1763–1776, apr 2008.
- [59] Negar Hariri, Carlos Castro-Herrera, Mehdi Mirakhorli, Jane Cleland-Huang, and Bamshad Mobasher. Supporting Domain Analysis through Mining and Recommending Features from Online Product Listings. *IEEE Transactions on Software Engineering*, 39(12):1736–1752, dec 2013.
- [60] Anil Kumar Thurimella and Bernd Bruegge. Issue-based variability management. *Information and Software Technology*, 54(9):933–950, sep 2012.
- [61] Thorsten Berger, Ralf Rublack, Divya Nair, Joanne M. Atlee, Martin Becker, Krzysztof Czarnecki, and Andrzej Wąsowski. A survey of variability modeling in industrial practice. In *Proceedings of the Seventh International Workshop on Variability Modelling of Software-intensive Systems - VaMoS '13*, page 1, New York, New York, USA, 2013. ACM Press.
- [62] Jaejoon Lee, Kyo C. Kang, Pete Sawyer, and Hyesun Lee. A holistic approach to feature modeling for product line requirements engineering. *Requirements Engineering*, 19(4):377–395, nov 2014.
- [63] Minjung Kwak and Harrison Kim. *Advances in Product Family and Product Platform Design*, volume 43. Springer New York, New York, NY, 2014.
- [64] Mikoláš Janota, Goetz Botterweck, and Joao Marques-Silva. On lazy and eager interactive reconfiguration. In *Proceedings of the Eighth International Workshop on Variability Modelling of Software-Intensive Systems - VaMoS '14*, pages 1–8, New York, New York, USA, 2013. ACM Press.

- [65] S Pablo, Diego Garc, Marta Zorrilla, and Dpto Matem. Software Product Line Engineering for e-Learning Applications : A Case Study. *Proceedings of the 2012 International Symposium on Computers in Education (SIIE)*, pages 1 – 6, 2012.
- [66] Krzysztof Czarnecki, Steven She, and Andrzej Wasowski. Sample Spaces and Feature Models: There and Back Again. In *2008 12th International Software Product Line Conference*, pages 22–31. IEEE, sep 2008.
- [67] Adam Woznica, Phong Nguyen, and Alexandros Kalousis. Model mining for robust feature selection. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, page 913, New York, New York, USA, 2012. ACM Press.
- [68] Nan Niu and Steve Easterbrook. On-Demand Cluster Analysis for Product Line Functional Requirements. In *2008 12th International Software Product Line Conference*, pages 87–96. IEEE, sep 2008.
- [69] Pavel Valov, Jianmei Guo, and Krzysztof Czarnecki. Empirical comparison of regression methods for variability-aware performance prediction. In *Proceedings of the 19th International Conference on Software Product Line - SPLC '15*, pages 186–190, New York, New York, USA, 2015. ACM Press.
- [70] Uzma Afzal, Tariq Mahmood, and Zubair Shaikh. Intelligent software product line configurations: A literature review. *Computer Standards & Interfaces*, 48:30–48, nov 2016.
- [71] Gangarn Sompras and Pattarachai Lalitrojwong. Extracting product features and opinions from product reviews using dependency analysis. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 5, pages 2358–2362. IEEE, aug 2010.
- [72] Conrad S. Tucker and Harrison M. Kim. Optimal Product Portfolio Formulation by Merging Predictive Data Mining With Multilevel Optimization. *Journal of Mechanical Design*, 130(4):041103, 2008.
- [73] Suppawong Tuarob and Conrad S. Tucker. Automated Discovery of Lead Users and Latent Product Features by Mining Large Scale Social Media Networks. *Journal of Mechanical Design*, 137(7):071402, jul 2015.
- [74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [75] Jaroslav Pokorný. Database technologies in the world of big data. *Proceedings of the 16th International Conference on Computer Systems and Technologies - CompSysTech '15*, pages 1–12, 2015.
- [76] Klaus Pohl, Günter Böckle, and Frank van der Linden. *Software Product Line Engineering: Foundations, Principles and Techniques*. 2005.
- [77] Daniela Castelluccia and Nicola Boffoli. Service-oriented product lines. *ACM SIGSOFT Software Engineering Notes*, 39(2):1–6, 2014.

- [78] Bahram Hamraz, Nicholas H. M. Caldwell, and P. John Clarkson. A Multidomain Engineering Change Propagation Model to Support Uncertainty Reduction and Risk Management in Design. *Journal of Mechanical Design*, 134(10):100905, 2012.
- [79] Guillaume Bécan, Mathieu Acher, Benoit Baudry, and Sana Ben Nasr. Breathing ontological knowledge into feature model synthesis: an empirical study. *Empirical Software Engineering*, 21(4):1794–1841, aug 2016.
- [80] Mathieu Acher, Philippe Collet, Philippe Lahire, and Robert B. France. Familiar: A domain-specific language for large scale management of feature models. *Science of Computer Programming (SCP)*, 78(6):657–681, 2013.
- [81] Guillaume Bécan, Razieh Behjati, Arnaud Gotlieb, and Mathieu Acher. Synthesis of attributed feature models from product descriptions. In *Proceedings of the 19th International Conference on Software Product Line - SPLC '15*, pages 1–10, New York, New York, USA, 2015. ACM Press.
- [82] András London, Áron Pelyhe, Csaba Holló, and Tamás Németh. Applying graph-based data mining concepts to the educational sphere. In *Proceedings of the 16th International Conference on Computer Systems and Technologies - CompSysTech '15*, pages 358–365, New York, New York, USA, 2015. ACM Press.
- [83] Xiangzhou Zhang, Yong Hu, Kang Xie, Shouyang Wang, E.W.T. Ngai, and Mei Liu. A causal feature selection algorithm for stock prediction modeling. *Neurocomputing*, 142:48–59, oct 2014.
- [84] Mingxian Wang and Wei Chen. A Data-Driven Network Analysis Approach to Predicting Customer Choice Sets for Choice Modeling in Engineering Design. *Journal of Mechanical Design*, 137(7):071409, jul 2015.
- [85] Giuseppe Manco, Pasquale Rullo, Lorenzo Gallucci, and Mirko Paturzo. Rialto: A Knowledge Discovery suite for data analysis. *Expert Systems with Applications*, 59:145–164, oct 2016.
- [86] Felipe F. Bocca and Luiz Henrique Antunes Rodrigues. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Computers and Electronics in Agriculture*, 128:67–76, 2016.
- [87] E. Mangalova and E. Agafonov. Wind power forecasting using the $\langle \text{mml:math altimg="si14.gif" display="inline. overflow="scroll" xmlns:xocs="http://www.elsevier.com/xml/xocs/dtd" xml} \rangle$. *International Journal of Forecasting*, 30(2):402–406, apr 2014.
- [88] Behrouz Zamani Dadaneh, Hossein Yeganeh Markid, and Ali Zakerolhosseini. Unsupervised probabilistic feature selection using ant colony optimization. *Expert Systems with Applications*, 53:27–42, jul 2016.
- [89] Narges Sajadfar and Yongsheng Ma. A hybrid cost estimation framework based on feature-oriented data mining approach. *Advanced Engineering Informatics*, 29(3):633–647, aug 2015.
- [90] Christian Kastner, Alexander Dreiling, and Klaus Ostermann. Variability Mining: Consistent Semi-automatic Detection of Product-Line Features. *IEEE Transactions on Software Engineering*, 40(1):67–82, jan 2014.

- [91] Changyun Huang, Kazuhiro Yamashita, Yasutaka Kamei, Kenji Hisazumi, and Naoyasu Ubayashi. Domain analysis for mining software repositories: Towards feature-based DSL construction. In *2013 4th International Workshop on Product Line Approaches in Software Engineering (PLEASE)*, pages 41–44. IEEE, may 2013.
- [92] Jayant Kalagnanam, Moninder Singh, Sudhir Verma, Michael Patek, and Yuk Wah Wong. A system for automated mapping of bill-of-materials part numbers. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, page 805, New York, New York, USA, 2004. ACM Press.
- [93] X Z Wang and C McGreavy. Automatic Classification for Mining Process Operational Data. *Industrial & Engineering Chemistry Research*, 37(6):2215–2222, jun 1998.
- [94] Rashid A Rayson P Edwards M. A systematic survey of online data mining technology intended for law enforcement. *ACM Computing Surveys*, 48,1, 15,,(<http://www.scopus.com/inward/record.url?eid=2-s2.0-84945398245&partnerID=40&md5=328a99120df>) 2015.
- [95] Fabian Rojas Blum, Jocelyn Simmonds, and María Cecilia Bastarrica. Software process line discovery. In *Proceedings of the 2015 International Conference on Software and System Process - ICSSP 2015*, number August 2015, pages 127–136, New York, New York, USA, 2015. ACM Press.
- [96] A Rashid, J.-C. Royer, and A Rummler. *Aspect-oriented, model-driven software product lines: The AMPLE way*. 2011.
- [97] Saba Pedram, Mehran Mohsenzadeh, and Amir Azimi Alasti Ahrabi. A Review of Feature Model Position in the Software Product Line and Its Extraction Methods. *International Journal of Computer Science and Security (ICSS)*, 9(5):274–279, 2015.
- [98] Kc Kang, V Sugumaran, and S Park. *Applied software product line engineering*. 2010.
- [99] Jinghua Wang, Jane You, Qin Li, and Yong Xu. Extract minimum positive and maximum negative features for imbalanced binary classification. *Pattern Recognition*, 45(3):1136–1145, mar 2012.
- [100] Martin Atzmueller, Stephan Doerfel, and Folke Mitzlaff. Description-oriented community detection using exhaustive subgroup discovery. *Information Sciences*, 329:965–984, feb 2016.
- [101] Matic Perovšek, Anže Vavpetič, Janez Kranjc, Bojan Cestnik, and Nada Lavrač. Wordification: Propositionalization by unfolding relational data into bags of words. *Expert Systems with Applications*, 42(17-18):6442–6456, oct 2015.
- [102] Sarah Nadi and Stefan Krüger. Variability Modeling of Cryptographic Components: Clafer Experience Report. *Proceedings of the Tenth International Workshop on Variability Modelling of Software-intensive Systems*, pages 105–112, 2015.
- [103] Weitao Chen, Xianju Li, Yanxin Wang, Gang Chen, and Shengwei Liu. Forested landslide detection using LiDAR data and the random forest algorithm: A case study of the Three Gorges, China. *Remote Sensing of Environment*, 152:291–301, sep 2014.

- [104] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18, aug 2015.
- [105] Sebastián Maldonado, Álvaro Flores, Thomas Verbraken, Bart Baesens, and Richard Weber. Profit-based feature selection using support vector machines – General framework and an application for customer retention. *Applied Soft Computing*, 35:740–748, oct 2015.
- [106] Wenbin Qian and Wenhao Shu. Mutual information criterion for feature selection from incomplete data. *Neurocomputing*, 168:210–220, nov 2015.
- [107] Claudia V. Isaza, Henry O. Sarmiento, Tatiana Kempowsky-Hamon, and Marie-Veronique LeLann. Situation prediction based on fuzzy clustering for industrial complex processes. *Information Sciences*, 279(7):785–804, sep 2014.