UNIVERSIDAD DEL PACÍFICO

Captación de clientes en la industria de telecomunicaciones peruana utilizando detección de comunidades

Diseño de soluciones de negocio

Autor: Victoria Zevallos Munguia

Profesor: Walter Aliaga

Lima, 2018

1 Tabla de contenido

2 Introducción

Actualmente el mundo de los negocios se encuentra frente a un gran cambio, producto del desarrollo e interacción de factores claves para su desenvolvimiento. Uno de ellos es el apogeo de la economía a nivel mundial, como indica el estudio realizado por el Banco Mundial a inicios del 2018, donde se menciona que el crecimiento económico se acelerará al 3.1% en el 2018 (World Bank Group, 2018). Además, el avance tecnológico de los últimos años se ha desarrollado de manera excepcional en diversos campos, permitiendo el desarrollo de nuevas soluciones y productos en prácticamente cualquier campo. La suma de ambos factores ha ayudado a la redefinición de una gran cantidad de industrias y al surgimiento de nuevos competidores y modelos de negocio, cuyas propuestas de valor se basan en innovación y están enfocadas en satisfacer las necesidades de los clientes (EY, 2017). En consecuencia, se genera un escenario en el cual la forma de competir cambia y la necesidad de emplear nuevas tecnologías e innovaciones se vuelve imprescindible para sobrevivir en la era del conocimiento y la disrupción.

En este contexto, una de las industrias más importantes y que más cambios ha experimentado a nivel mundial es la de Telecomunicaciones. La fusión y adquisición de varias empresas y la desregularización de los mercados y han sido responsables de estos cambios (PWC, 2016). Como resultado, la intensa competencia entre las empresas de este sector se basa en la captación de nuevos clientes y la mejora constante de la oferta brindada. Es así como en los últimos 20 años las operadoras han gastado montos significativos de dinero con el fin de incrementar su cuota de mercado, sin reparar en la ganancia neta de la compañía (PWC, 2016). Con este fin, ahora la inversión se dirige a encontrar y desarrollar nuevas estrategias que permitan a las compañías posicionarse en el mercado mediante la explotación de diversos recursos, uno de ellos es la información.

En cuanto a las empresas de telecomunicaciones, dada la gran cantidad de competencia en la industria y el surgimiento de competidores fuera de la industria tradicional, como las aplicaciones móviles que facilitan la comunicación (PWC, 2016), el uso de estas nuevas técnicas brinda nuevas oportunidades para competir. Como se explico previamente, la industria ha identificado que el foco debe estar en la captación y retención de clientes. Por esta razón diversas compañías del sector a nivel mundial están usando técnicas de inteligencia artificial, minería de datos y *big data*; para optimizar las decisiones de diversos procesos en la empresa, siendo uno de los mas importantes la cadena de valor del consumidor (Linoff & Berry, 2011) (Ortiz, 2017). Mientras que, en esta nueva realidad, los

clientes son cada vez mas exigentes y reconocen su valor comercial (EY, 2017), por lo que las empresas en general deben generar estrategias inteligentes para captarlos.

En el contexto Latinoamericano, la industria de telefonía móvil es la que genera el 5% del PBI de América Latina, siendo uno de los principales motores del crecimiento económico y desarrollo social de la región, según informe de GSMA (GSMA Intelligence, 2017). A pesar de esto, muchos países de la región aun permanecerán con una baja penetración de 80% para fines del 2020 (GSMA, 2016). En consecuencia, a partir de esta fecha, cuando la mayoría de los mercados hayan alcanzado madurez (PWC, 2016) la competencia dejará de estar enfocada en la atracción de nuevos clientes de la industria y pasara a centrase en captación de clientes de la competencia con mayor fuerza. Por consiguiente, la evolución de la industria de las telecomunicaciones en Latinoamérica sigue los pasos de las industrias en países altamente desarrollados. Bajo esa premisa se puede intuir cuales serán los nuevos retos y las herramientas mas adecuadas para enfrentarlos.

En el Perú, el sector de las telecomunicaciones se encuentra en una situación similar al del resto de los países sudamericanos. Según el presidente de OSIPTEL, el dinamismo este sector ha sido calve para el crecimiento de la economía de la nación (OSIPTEL, 2017). Además, un estudio muestra que en el 2016 la tasa de penetración móvil en el país fue de 66% con una proyección de aproximadamente 80% al 2020 (GSMA, 2016). Es decir, se encuentra en una etapa de expansión hacia zonas rurales y con cierto grado de competencia para atraer clientes que ya se encuentran dentro de la industria. Se espera que una vez la penetración haya alcanzado niveles altos, los operadores móviles competirán por la cuota de mercado.

Esta competencia se esta viendo altamente favorecida por tres factores principales. El primero es el mecanismo de portabilidad instaurado por OSIPTEL años atrás; el segundo es la apertura del mercado y la llegada de nuevos competidores (OSIPTEL, 2017); y el tercero es el empoderamiento de los usuarios. Esta circunstancia permite que los clientes, individuos y empresas (EY, 2017), estén dispuestos y puedan elegir a la compañía con la cual se sientan mas satisfechos, aquella que los entienda y cuya propuesta de valor satisfaga mejor las necesidades particulares de cada uno (EY, 2017). En ese escenario, la mejora de la estrategia de captación de clientes es claramente fundamental para superar a la competencia y utilizar la situación como una oportunidad para aumentar la cuota de mercado.

Se han realizado varios trabajos de investigación, usando métodos de minería de datos, enfocados en la relación con los clientes, específicamente en la predicción de deserción para

la retención de estos mediante el análisis de la red social de telecomunicaciones a la que pertenecen (Columelli, Nuñez del Prado, & Zarate Gamarra, 2016) (Pushpa & Shobha, 2013) (Wei & Chiu, 2002). De igual forma se han investigado nuevas técnicas de detección de comunidades que puedan trabajar con redes reales, teniendo en cuenta su alta dimensión (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) (Varun & Pushpa, 2016) (Seungyo & Dongseung, 2016) (Upadhyay & Singh, 2017). Además, hay autores que han basado sus trabajos en realizar un análisis comparativo entre diferentes técnicas de detección de comunidades para determinar cuales son mas eficientes dependiendo de las características de la red (Garg & Rani, 2017) (Mothe, Mkhitaryan, & Mariam, 2017) (Chejara & Godfrey, 2017). Sin embargo, ninguno de estos utiliza el análisis de redes y la detección de comunidades para mejorar la estrategia de captación de clientes.

En consecuencia, en este trabajo se propone analizar la estructura de la red de telecomunicaciones (TSNA) de los usuarios de telefonía móvil en el Perú para captar nuevos clientes. Este análisis consistirá en la identificación de comunidades del grafo que representa la red social móvil peruana, donde los usuarios son los vértices y las relaciones de comunicación entre ellos las aristas. Como plantean algunos autores, las técnicas de detección de comunidades facilitan el entendimiento de redes basadas en data real, cuya principal característica es la alta complejidad de estas por la cantidad de usuarios e interacciones. (Chejara & Godfrey, 2017). Posteriormente se identificarán a los usuarios mas influyentes de las comunidades pues son los que es mas importante captar por la posición que ocupan en la estructura de la red. Finalmente se presentará a los usuarios identificados como trascendentales para el proceso de captación.

Finalmente, el presente trabajo de investigación se organiza en ocho apartados: (1) marco problemático, (2) objetivos, (3) viabilidad de la investigación, (4) estado del arte, (5) bases teóricas, (6) metodología (7) resultados (8) conclusiones y recomendaciones.

3 Marco Problemático

En los últimos años, el desarrollo y evolución de la tecnología ha sido el principal motor de la economía mundial (The Boston Consulting Group, 2018). Dado que, el surgimiento de nuevas tecnologías ha tenido como consecuencia la disrupción y transformación de las industrias, gobiernos y personas. Esta nueva sociedad supone nuevos retos y brinda herramientas, antes inimaginables para las empresas (EY, 2017). Afortunadamente, el progreso tecnológico trajo consigo el desarrollo de computadores capaces almacenar grandes volúmenes de información y la creación de metodologías basadas en analítica avanzada que permiten la extracción de conocimiento valioso de la data. Estas nuevas herramientas aplicadas en un contexto empresarial abren la posibilidad a un sinfín de usos; permiten solucionar problemas de negocio de manera innovadora y adelantarse a la competencia (Capgemini Consulting, 2013).

Una de las industrias con mayor crecimiento y en constante expansión en las ultimas décadas ha sido la de telecomunicaciones (PWC, 2016), debido al avance tecnológico y a la necesidad de las personas de comunicarse entre ellas sin que la distancia sea un problema. Años atrás, una de las principales características de la industria eran las altas barreras de entrada a nuevos competidores básicamente por el costo de infraestructura y las normativas legales. Dicha situación permitió la existencia de monopolios en diversos países donde las operadoras al no tener competencia descuidaron el servicio brindado, principalmente en Latinoamérica (GSMA, 2016). Sin embargo, con el paso del tiempo estas barreras empezaron a caer y la industria a transformarse. Actualmente la llegada de nuevos competidores aumenta año a año en varios países; los clientes están mas empoderados, saben que son cruciales para las empresas; y las instituciones y políticas publicas velan por proteger al usuario (OSIPTEL, 2017). En este nuevo escenario, el comprender a los usuarios de telefonía y poder satisfacerlos es vital para las empresas que compiten en este sector.

Sin embargo, a pesar de la gran importancia que tienen los consumidores por las razones expuestas, las empresas peruanas de telecomunicaciones no han sabido satisfacer los deseos de los consumidores. En consecuencia, solo el 34% de los clientes recomiendan su operador telefónico (Everis, 2017), dicha cifra deja entrever que mas del 50% no esta contento con el servicio recibido. Así, se explica porque desde julio del 2014 hasta el cierre del 2017, se registraron un total de 6,156,258 líneas móviles portadas (OSIPTEL, 2017). Mientras que solo en enero del 2018 hubo otras 473,655 portaciones (OSIPTEL, 2018). En términos de

liderazgo de mercado, la empresa Entel gano 2,739,203 y la empresa Claro 2,146,489 hasta enero del 2018 (OSIPTEL, 2018).

Esta situación, mas allá de ser un problema supone una excelente oportunidad para las empresas de telefonía. Existente dos posibles estrategias que estas pueden usar, la primera consiste en retener a los usuarios y así evitar la deserción, para esto se deben aplicar técnicas que permitan entender muy bien a sus propios clientes; mientras que la segunda consiste en aprovechar la insatisfacción de los clientes de otros operadores y diseñar tácticas que permitan identificarlos, entenderlos, ofrecerles soluciones y finalmente convertirlos en clientes de la empresa. La segunda opción supone retos importantes donde no solo importa el usuario como individuo sino la relación que tiene con otros usuarios.

En conclusión, las empresas que pertenecen al rubro de las telecomunicaciones se encuentran en un mercado altamente competitivo cuya principal característica son los clientes empoderados, donde innovar y diseñar nuevas estrategias de captación de clientes es muy importante. Las formas de competencia tradicional pierden efectividad con el paso del tiempo creando la necesidad de nuevos métodos que permitan generar estrategias basadas en técnicas que permitan explotar grandes volúmenes de data y extraer información y conocimiento valioso de la misma. En el Perú, no se ha hecho un estudio acerca de la aplicación de técnicas de análisis de redes sociales basadas en las interacciones de los usuarios para mejorar la captación de clientes, en el contexto de las empresas de telecomunicaciones y sus características particulares. Con esta información las Telcos podrán realizar una mejor captación de clientes, adelantándose a la competencia mediante propuestas de valor enfocadas en satisfacer mejor las necesidades de los usuarios de la competencia.

4 Objetivos

4.1 Objetivo general

En vista de la importancia de la captación de nuevos clientes en el escenario de la industria de telecomunicaciones peruana, se propone seguir una metodología que permita a las *Telcos* alcanzar este fin. Esta metodología se basa en un entendimiento del problema de negocio y la explotación de datos sobre las interacciones entre usuarios de la red de telefonía móvil para determinar que usuarios de la competencia se debería captar. Dichos individuos se caracterizan por estar altamente relacionados con los clientes de la empresa, por lo que la captación es mas probable. De igual modo se busca distinguir a los usuarios mas influyentes de la red por la repercusión futura que tendrá la captación de estos. En ese sentido, el principal objetivo de este trabajo de investigación es identificar potenciales clientes para la captación en base a las interacciones entre usuarios de la red de telefonía móvil en el Perú.

4.2 Objetivos específicos

En aras de alcanzar el objetivo principal del trabajo de investigación, se plantean cinco objetivos específicos que ayudarán a conseguir dicho propósito:

El primer objetivo especifico consiste en la construcción de un grafo basado en los datos sobre interacciones entre los usuarios de telefonía móvil. Donde los nodos del grafo representan a los usuarios y las aristas simbolizan la relación de comunicación que existe entre ellos, ya sean llamadas telefónicas o mensajes de texto. Este grafo contará con una gran cantidad de nodos y relaciones pues solo así podrá representar apropiadamente el fenómeno de comunicación móvil real en el Perú. De manera que la relación de comunicación móvil existente entre personas será simbolizada como una red social y por consiguiente se podrán aplicar técnicas de análisis de redes para entender el comportamiento los clientes.

El segundo objetivo especifico busca identificar la mejor técnica de ponderación de aristas para un grafo basado en una red social de telecomunicaciones. Dado que la comunicación móvil entre personas se puede dar a través de diferentes canales y medirse mediante diferentes variables, se pretende encontrar que método de ponderación de variables es el mas adecuado. Actualmente existen diversas técnicas de ponderación por lo que se debe probar con algunas de ellas para encontrar la mas indicada. Dicho técnica permitirá construir un grafo representativo que facilite y mejore el análisis de la estructura de la red.

El tercer objetivo especifico planteado es evaluar que algoritmo de detección de comunidades en grafos construidos a partir de datos de telecomunicaciones funciona mejor para la captación de nuevos clientes. En base a las interrelaciones dentro del grafo se busca grupos afines, también conocidos como comunidades. La cuales se caracterizan por presentar una fuerte interacción entre los miembros de la comunidad, mientras mantienen una menor interacción con los miembros de las otras comunidades. Es importante que la técnica de detección de comunidades tenga en cuenta las características propias de una red de comunicaciones real, como son la complejidad y el dinamismo; y pueda desempeñarse bien con ellas. Es decir, que sea adaptable a un entorno de procesamiento en paralelo o cuyo costo computacional no se eleve exponencialmente con relación a la complejidad. De esta manera será viable su implementación y uso en la industria de telecomunicaciones para obtener conocimientos importantes y desarrollar estrategias.

El cuarto objetivo especifico consiste en identificar a los usuarios mas influyentes de red social de telecomunicaciones. Estas personas se caracterizan por el alto grado de conexión que tienen con otros miembros de la red. En consecuencia, la captación puede ser mas eficiente si se concentran los esfuerzos en atraer a estos clientes clave, pues tendrá repercusión en otros usuarios que posteriormente serán mas fáciles de captar. Dada la complejidad de la red, primero se debe extraer las comunidades y posteriormente identificar a los nodos centrales de las mismas, ya que, estos son los principales candidatos para ser captados.

El quinto objetivo especifico pretende realizar una representación grafica del grafo y de las comunidades encontradas. Aprovechando que la representación de una red de telecomunicaciones mediante grafos mas allá de ser útil para el análisis estructural, brinda una representación grafica e interactiva del comportamiento y de las relaciones dentro de la red. De igual forma, una vez encontradas las comunidades es importante que los tomadores de decisiones puedan verlas representadas gráficamente con diferentes colores y las relaciones dentro y entre ellas. Así se facilitará la caracterización de las comunidades obtenidas y el uso del conocimiento extraído para mejorar efectivamente la capación de clientes de la competencia.

5 Viabilidad

Para desarrollar el proyecto de investigación se cuenta con un conjunto de datos de las interacciones móviles de los clientes de una de las empresas mas importantes del sector de telecomunicaciones del país. El cual cuenta con información sobre llamadas realizadas y mensajes de texto enviados entre los usuarios de telefonía móvil. Mediante la ponderación de estos datos se podrá generar una matriz de conectividad. Posteriormente, en base a esta se formará un grafo donde los números de destino y origen personificarán a los usuarios, por ende, serán los nodos y las aristas estarán conformadas por una ponderación de las variables que caracterizan la interacción en la red de telefonía móvil.

Además, se cuenta con información relevante para entender adecuadamente el problema de negocio, asimismo, el conocimiento de la realidad nacional del país en el cual estará enfocada la investigación resulta ser importante. De igual forma, el conocimiento sobre el uso de la analítica de datos y su aplicación para la resolución de problemas empresariales será fundamental para la ejecución de la investigación.

El presente proyecto, tiene como tiempo estimado de ejecución 6 meses, tiempo necesario para alcanzar los objetivos de este. A parte, se dispone de una estructura computacional proporcionada por la universidad que permite trabajar con grandes volúmenes de datos y realizar las pruebas necesarias. En conclusión, la ejecución del proyecto cumple con los requisitos necesarios para ser viable, como son; existencia de datos, conocimientos sobre el tema y enfoque de la investigación, tiempo de ejecución y acceso a recursos computacionales necesarios.

6 Estado del arte

Diversos trabajos se han realizado respecto a la detección de comunidades en grafos, por su utilidad para entender la estructura interna de las redes y los subgrupos que la componen. Esta información puede ser aplicable en diversos campos, uno de ellos es el contexto empresarial, donde la aplicación de esta técnica se utiliza para analizar mejor a los clientes y desarrollar mejores estrategias para la atracción, predicción de *churn* y fidelización de estos (Indrawati & Alamsyah, 2017). Específicamente para el sector de telecomunicaciones, este análisis supone retos importantes, entre los cuales se encuentran la complejidad, dimensión, heterogeneidad, direccionalidad y dinamismo de la red (Pushpa & Shobha, 2013). En ese sentido la revisión de la literatura no se ha centrado únicamente en estudios de detección de comunidades o predicción de desertores aplicadas a redes de usuarios de telefonía móvil, sino también en investigaciones que desarrollan y comparan algoritmos que permitan trabajar con redes complejas de manera eficiente.

6.1 Detección de comunidades

Uno de los trabajos mas relevantes, respecto a la detección de comunidades en redes de gran dimensión es la publicación "*Fast unfolding of communities in large networks*", desarrollada por cuatro autores en la Universidad Católica de *Louva*in (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). El algoritmo propuesto en este trabajo marca un punto de partida en el manejo y descubrimiento de comunidades en grafos complejos de alta dimensión, representación de redes reales. El método es heurístico, basado en la optimización del modularidad. El cual está compuesto por dos fases que iteran, la primera consiste en realojar los nodos en el vecindario que maximice la medida mencionada y la segunda, en construir una nueva red usando las comunidades encontradas en la fase previa como nodos. En el momento en el que fue propuesto se demostró que superaba a otros métodos de detección de comunidades en términos de eficiencia computacional (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Las pruebas fueron hechas con conjuntos de datos de características diversas, uno de estos fue un grafo construido a partir de el inmenso data set de una compañía telefónica belga. Además, se realizo un análisis en cuanto a la precisión para encontrar comunidades mediante un grafo con comunidades previamente descubiertas. En todas estas pruebas el performance del método propuesto alcanzo valores satisfactorios.

Otra investigación realizada con el fin de detectar comunidades en las redes de telecomunicaciones es propuesta por E. Varun y Pushpa Ravikumar en 2016. En este trabajo,

ellos proponen basar la detección de comunidades en la descomposición de la red en n-cliques. La principal causa de su razonamiento radica en que, al trabajar con redes de telecomunicaciones, caracterizadas por su heterogeneidad y su gran cantidad de datos, es necesario entender la estructura de estas para hacer un buen análisis. En consecuencia, analiza diferentes variedades de subestructuras como son cliques, *N-cliques y K-cores* para extraer importantes características de la red que permitan un mayor conocimiento sobre los subgrupos y sus interrelaciones (Varun & Pushpa, 2016). Además, proponen que esta técnica permite a las compañías móviles utilizar la actividad social de sus clientes para mejorar su oferta e incentivos (Varun & Pushpa, 2016), centrándose en los atributos que cada comunidad valora mas, es decir, personalización del servicio.

El siguiente estudio es relevante en términos de eficiencia computacional y resultados obtenidos al momento de detectar comunidades en redes complejas. Fue propuesto por Seungyo Ryu & Dongseung Kim, el planteamiento propone dos modificaciones al algoritmo de *Louvain* en su primera fase, presentado previamente. La primera, conocida como limitación de la búsqueda de comunidades, consiste en restringir el numero de iteraciones en la búsqueda de la comunidad adecuada. La segunda, se basa en restringir las reglas de búsqueda interna, mediante la depuración de candidatos poco relevantes y cuyo impacto en la maximización de modularidad sea baja, llamada temprana asignación de comunidad (Seungyo & Dongseung, 2016). Dichos algoritmos son evaluados en tres grafos pequeños, dos grafos autogenerados y tres conjuntos de datos reales de gran complejidad, ambos muestran un mejor desenvolvimiento computacional que el algoritmo original solo en los grafos de alta dimensión. Por tanto, es una técnica que permite detectar comunidades de manera rápida, ha demostrado ser mas veloz en 40.5x que el *Louvain*, sin comprometer la calidad de la solución (Seungyo & Dongseung, 2016).

Otra metodología propuesta recientemente para detectar comunidades en grafos es desarrollada por A. Upadhyay & M. Singh. Esta iniciativa se basa en una aproximación para la detección de comunidades en grafos con peso, siendo una variante del algoritmo "*Attractiveness-Based Community Detection*" (ABCD). El cual agrupa los nodos en comunidades según el orden de visita a los componentes y sus conexiones (Ruifang, Shan, Ruisheng, & Wenbin, 2014). Entonces el nuevo planteamiento busca transformar ABCD en un algoritmo de orden independiente. Removiendo esta dependencia, el algoritmo primero agrupa los componentes altamente conectados sin tomar en cuenta el orden de estos (Upadhyay & Singh, 2017). El algoritmo es probado con dos conjuntos de datos de fenómenos reales, uno mediano y otro grande. Los resultados demuestran que es mas efectivo para trabajar con grafos multirelacionados con peso que el ABCD (Upadhyay & Singh, 2017).

Investigación	Autores	Algoritmo	Objetivo
Fast unfolding of communities in large networks	(Blondel, Guillaume, Lambiotte, & Lefebvre, 2008)	• *Louvain*	Propuesta del algoritmo de Louvain para trabajar con redes complejas.
Telecommunication community detection by decomposing network into n-cliques	(Varun & Pushpa, 2016)	• *N-cliques* • *K-cores*	Extraer características importantes de la estructura y subestructura de la red de telecomunicaciones.
Quick Community Detection of Big Graph Data Using Modified Louvain Algorithm	(Seungyo & Dongseung, 2016)	• *Louvain - Community Search Limitation* • *Louvain - Early Community Assignment*	Mejorar la detección de comunidades de Louvain en grafos complejos mediante dos modificaciones de este.
Weighted graph clustering for community detection of large social networks	(Ruifang, Shan, Ruisheng, & Wenbin, 2014)	• *ABCD*	Mejora el algoritmo ABCD para mejorar el performance al ser aplicado a grafos con peso.

Tabla 6.1: Resumen de investigaciones sobre detección de comunidades

6.2 Prevención e influencia de desertores

Al respecto, la investigación realizada por Columelli, Nuñez del Prado, & Zarate consiste en el análisis de la red social de telefonía de una operadora de telecomunicaciones africana para clasificar a los posibles desertores y la influencia de estos sobre la red. En primer lugar, hacen un comparativo entre varios clasificadores, donde el algoritmo Extremely Random Forest obtiene un mejor performance según la medida lift (Columelli, Nuñez del Prado, & Zarate Gamarra, 2016). Luego usan medidas como grado de centralidad y page rank para medir el grado de influencia emitida y recibida de cada posible desertor. Finalmente, utilizan

un sistema de lógica difusa para obtener una métrica unificada del riesgo de deserción en base a las medidas de probabilidad de deserción, influencia emitida (grado de centralidad) e influencia recibida (*page rank*) (Columelli, Nuñez del Prado, & Zarate Gamarra, 2016). Este estudio es valioso pues realiza un análisis de la red de telecomunicaciones, propone una metodología interesante para medir el riesgo de *churn*, medida que puede ser usada para encontrar a los usuarios mas importantes.

Un estudio importante, basado en la detección de comunidades e identificación de la estructura y elementos de una red sobre datos de telecomunicaciones para su posterior clasificación, fue desarrollado en India por Pushpa Ravikumar & G. Shobha. Esta investigación propone analizar la estructura y comportamiento de una red multirelacional y la ubicación de elementos importantes para predecir el *churn* de clientes de una operadora móvil. Se evalúa la posición social de los usuarios representados por nodos en base a las múltiples conexiones que tienen con otros usuarios de la red, esta medida se conoce como centralidad. Luego este valor es usado para caracterizar los grados de influencia e importancia de ciertos miembros (Pushpa & Shobha, 2013). Finalmente se utiliza el algoritmo iterativo REGE basado en la equivalencia regular, es decir la similitud entre relaciones; de esta forma se clasifica a los usuarios en desertores o no desertores (Pushpa & Shobha, 2013). Este análisis es importante porque analiza la estructura de la red y la importancia social de los nodos, en este caso dicho conocimiento se utiliza para predecir el *churn* y evitarlo; sin embargo, también puede utilizarse para la atracción de los clientes mas propensos a abandonar a la competencia.

Investigación	Autores	Medida de centralidad	Objetivo
Measuring Churner Influence on Pre-paid Subscribers Using Fuzzy Logic	(Columelli, Nuñez del Prado, & Zarate Gamarra, 2016)	• Grado de centralidad • *Page rank*	Proponer metodología para medir riesgo de deserción unificando medidas.
Social network classifier for churn prediction in telecom data	(Pushpa & Shobha, 2013)	• Equivalencia multirrelacional • Grado de centralidad	Proponer metodología para clasificar usuarios desertores y su influencia en la red.

Tabla 6.2: Resumen de investigaciones sobre prevención e influencia de desertores

6.3 Comparación entre algoritmos de detección de comunidades

En cuanto a la efectividad y ventajas de los diversos métodos planteados y desarrollados a través del tiempo para detectar comunidades en grafos, algunos autores han desarrollado trabajos basados en la comparación de técnicas disponibles en la literatura. Estos análisis son valiosos para poder elegir la metodología a emplear en determinado caso y según las características del grafo. En ese sentido a continuación se presentan tres investigaciones que han realizado un análisis comparativo de diferentes métodos de detección de comunidades.

En primer lugar, el trabajo presentado por Garg y Rani, compara en base a la modularidad, diferentes aproximaciones de la detección de comunidades en redes del mundo real. Estas son: *edge betweenness, fast greedy, random walk, spin glass, label propagation, leading eigenvector y louvain algorithm*. Como resultado después de probar estos métodos en diferentes conjuntos de datos, los autores concluyen que, en redes con gran cantidad de nodos, el algoritmo *louvain, random walk y spin glass* son los mejores en cuanto a modularidad para la detección de comunidades (Garg & Rani, 2017).

En esta línea se encuentra también el estudio llevado a cabo por Mothe, Mkhitaryan, & Mariam midiendo la efectividad de diversas técnicas de detección de comunidades aplicadas a múltiples redes artificiales generadas con el modelo estocástico de bloques. La idea del análisis es medir modularidad y tiempo de procesamiento de diferentes métodos sobre redes cuya estructura es conocida por los investigadores a-priori, de esta manera se puede evaluar mejor el resultado de los algoritmos. Las técnicas aplicadas son: *louvain, fast greedy, leading eigenvector, random walk, infomap, label propagation*. Los resultados muestran que los algoritmos de *louvain y leading eigenvector* obtienen mejores resultados, alto grado de modularidad. En cuanto al tiempo de *ejecucuion louvain y label propagation* se ejecutan mas rápido que los otros algoritmos conforme el numero de vértices aumenta (Mothe, Mkhitaryan, & Mariam, 2017).

Un trabajo similar es realizado por Chejara & Godfrey aplicado en dos fases, la primera a redes pequeñas, medianas y la segunda a redes grandes. En este caso ellos centran gran parte de su análisis en el estudio de grandes redes por la importancia y complejidad de sistemas de la vida real. Los métodos elegidos para hacer la comparación son: *edge betweenness, infomap, louvain algorithm, fast greedy, spin glass, random walk y label propagation*. Según este análisis en base a la modularidad y tiempo de ejecución, los algoritmos con mejor performance en redes grandes son *louvain, fast greedy y edge betweenness* (Chejara & Godfrey, 2017).

Investigación	Autores	Algoritmos	Conclusiones
A comparative study of community detection algorithms using graphs and R	(Garg & Rani, 2017)	• *Edge betweenness* • *Fast greedy* • *Random walk* • *Spin glass* • *Label propagation* • *Leading eigenvector* • *Louvain*	Los algoritmos con mejor modularidad son: *Louvain, Random walk y Spin glass.*
Community detection: Comparison of state of the art algorithms	(Mothe, Mkhitaryan, & Mariam, 2017)	• *Louvain* • *Fast greedy* • *Leading eigenvector* • *Random walk* • *Infomap* • *Label propagation*	Los algoritmos con mejor modularidad son: *Louvain, Leading eigenvector.* Los algoritmos con menor tiempo de ejecución son: *Louvain, Label propagation.*
Comparative analysis of community detection algorithms	(Chejara & Godfrey, 2017)	• *Edge betweenness* • *Infomap* • *Louvain* • *Fast greedy* • *Spin glass* • *Random walk* • *Label propagation*	Los algoritmos con mejor modularidad son: *Louvain, Fast greedy y Edge betweenness*. Los algoritmos con menor tiempo de ejecución son: *Louvain y Edge betweenness.*

Tabla 6.3: Resumen de investigaciones comparativas

7 Bases teóricas

Los conceptos, teorías y metodologías clave usados en el presenté trabajo de investigación se detallan a continuación.

7.1 Teoría de grafos

Es una rama de estudio para la exploración y aplicación de técnicas usada principalmente por las áreas de matemática discreta y ciencias de la computación (West, 2001). Su objetivo es estudiar y explicar las propiedades de los grafos. Es importante pues muchos problemas de la vida real pueden ser representados mediante grafos y por ende aprovechar las propiedades matemáticas de estos para analizarlos y resolverlos (Wilson, 1979).

7.1.1 Grafo

Formalmente, un grafo se define como $G = (V, E)$, un par ordenado. Donde $V(G)$ es un conjunto finito no vacío de vértices o nodos y $E(G)$ es un conjunto de aristas que conectan uno o dos elementos, subconjunto de V (Wilson, 1979). Eso significa que cada arista se relaciona con uno o dos elementos del subconjunto de V (Trudeau, 1993). Una arista $\{v, w\}$ se dice que conecta los vértices v y w, y es usualmente abreviada como vw (Wilson, 1979). Además, el grado de un vértice v se define como el numero de aristas que lo tienen como extremo (Wilson, 1979).

7.1.2 Propiedades de los grafos

Las principales propiedades de los grafos son las siguientes:

Incidencia: Se dice que una arista e es incidente a un vértice v de un grafo G si esta lo une a otro vértice w (Wilson, 1979).

Adyacencia: Se dice que dos vértices v y w de un grafo G son adyacentes si hay una arista vw que los une. De manera similar, dos aristas diferentes e y f son adyacentes si tienen un vértice v en común (Wilson, 1979).

Ponderación: Corresponde a una función que asigna un valor o peso a una arista e de un grafo G para aumentar la representatividad de la relación (Trudeau, 1993).

7.1.3 Vecindario

Dado el grafo $G = (V, E)$, donde $v \in V$ los vecinos de v son el conjunto de vértices adyacentes a v. Formalmente se define como: $N(v) = \{u \in V : \exists e \in E(e = \{u, v\} \cup u = v \cap e = \{v\})\}$ (Trudeau, 1993).

7.1.4 Grafo dirigido

Un grafo dirigido se define como $G = (V, E)$, un par ordenado. Donde $V(G)$ es un conjunto finito no vacío de vértices o nodos y $E(G)$ es una colección de elementos contenidos en VxV. Eso significa que E es una colección de pares de vértices ordenados. Las aristas $e \in E$ se llaman aristas dirigidas. (Trudeau, 1993).

7.1.5 Representación matricial

Una de las estructuras mas usadas para almacenar y representar grafos son las matrices. Una de ellas es la matriz de incidencia, donde el grafo G esta representado por una matriz de orden $n\,x\,m$, dado que n es el numero de vértices y m el numero de aristas; donde cada entrada ij es 1 si el vértice i es incidente a la arista j, y 0 de otra forma (Wilson, 1979). La otra es la matriz de adyacencia de orden $n\,x\,n$, siendo n es el numero de vértices, donde cada entrada ij es el numero de aristas que unen el vértice i y el vértice j $(\text{Wilson}, 1979)$.

7.2 Red social

Una red social se define como la colección de una cadena de individuos y las conexiones personales existentes entre ellos (Bonchi, Castillo, Gionis, & Jaimes, 2011). Brinda una aproximación donde la atención no se centra solo en un individuo sino en las relaciones entre varios individuos de la red y su interacción, dado que para muchas aplicaciones esta información es mas valiosa (Dasgupta, et al., 2008). Las características de esta red son la multi-relacionalidad, heterogeneidad y naturaleza dinámica, pues cambia a través del tiempo (Varun & Pushpa, 2016). Con el paso del tiempo y producto del avance tecnológico cada vez se puede recolectar mayor cantidad de datos para mejorar el modelamiento de redes sociales mas complejas, en especial sobre las conexiones (Bonchi, Castillo, Gionis, & Jaimes, 2011).

En consecuencia, una gran variedad de disciplinas ha mostrado interés en el campo y su análisis, una de ellas son los negocios (Bonchi, Castillo, Gionis, & Jaimes, 2011).

7.3 Análisis de una red social (SNA)

El SNA se basa en la abstracción de las interacciones sociales en el mundo real para el análisis y entendimiento de una red social (Indrawati & Alamsyah, 2017). Con este propósito, utiliza elementos de la teoría de grafos, así, representa a las personas como y a las relaciones entre ellas como las aristas (Newman, 2011). El objetivo es encontrar la estructura que define las interacciones entre usuarios. Es decir, el numero de usuarios, las conexiones existentes y a las personas mas importantes o influyentes de la red (Pushpa & Shobha, 2013). Explorar la naturaleza y fuerza de las interconexiones entre usuarios puede ayudar a comprender la estructura y dinamismo de una red social (Dasgupta, et al., 2008). En este análisis se tienen en cuenta los atributos multi-relacionales y heterogéneos que caracterizan a las redes sociales reales (Varun & Pushpa, 2016). El uso del SNA es muy utilizado para la identificación de segmentos y captación de clientes (Pandapotan, Alamsyah, & Paryasto, 2015).

7.3.1 Análisis de una red social de telecomunicaciones (TSNA)

Una red social de telecomunicaciones consiste en un conjunto de clientes, también llamados abonados, que mantienen una o mas tipos de relaciones de comunicación entre ellos (Pushpa & Shobha, 2013). Con el objetivo de analizar dicha red, esta puede ser representada mediante un grafo. Donde los nodos o vértices del grafo representan el id del numero telefónico, el cual define a cada usuario como único, mientras que la relación entre dos usuarios se simboliza con una arista que es adyacente a los vértices de dichos usuarios (Pushpa & Shobha, 2013). Esta arista es dirigida y su valor va a depender de la función ponderación que se elija en base a las variables de interconexión entre usuarios (Pushpa & Shobha, 2013).

Estas redes se pueden clasificar en dos tipos, homogéneas y heterogéneas. Las primeras son aquellas con un solo tipo de relación entre los abonados, mientras que las segundas presentan varios tipos de relación entre los usuarios y pueden ser llamadas redes multi-relacionales (Wei & Chiu, 2002). Finalmente, es importante resaltar que, al analizar una red compleja con muchas personas, es de esperarse la existencia de subgrupos, los cuales son importantes para la comprensión de la estructura y comportamiento de los usuarios de telefonía (Pushpa & Shobha, 2013). En ese sentido parte del análisis de la red consiste en utilizar algoritmos de detección de comunidades para identificar dichos grupos internos y su estructura (Bonchi, Castillo, Gionis, & Jaimes, 2011).

7.4 Detección de comunidades

Una comunidad se define como un subgrupo de la red social (Pushpa & Shobha, 2013); por tanto, también puede ser representada por un grafo. El cual consta de un conjunto de vértices o nodos unidos entre si mediante aristas, donde la densidad de dichas conexiones es alta dentro de la comunidad; mientras, los vínculos entre los nodos de diferentes comunidades tienden a ser de baja densidad (Hendrikx & Nuñez del Prado, 2016). Entonces, la detección de comunidades es el termino usado para referirse a clusterizacion en grafos. La extracción de *clusters* o comunidades relevantes en grandes redes sociales es un verdadero reto; sin embargo, provee información importante para la toma de decisiones sobre la estructura y el comportamiento de los individuos dentro de las redes y los subgrupos (Wu & Liu, 2008). Los hallazgos encontrados no son siempre los esperados (Wu & Liu, 2008), y es justamente este resultado el que hace que la detección de comunidades sea tan valiosa, dado que revela interacciones y estructuras que no se ven a simple vista.

7.5 Medidas de calidad de comunidades

7.5.1 Modularidad

Modularidad (Q) se define como la medida de calidad de la división de una red en comunidades (Garg & Rani, 2017). La red alcanza un alto valor de modularidad si existe una fuerte conexión entre los nodos dentro de la comunidad y una conexión débil entre los nodos de diferentes comunidades (Clauset, Newman, & Moore, 2004). La formula propuesta por Clauset, es la siguiente:

$$Q(P) = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{d_i d_j}{2m} \right] \sigma(C_i, C_j)$$

En donde, A_{ij}, es la matriz de adyacencia del grafo; d_i, es el grado del nodo i y d_j, es el grado del nodo j. Siendo m el numero de aristas y la expresión $\frac{d_i.d_j}{2m}$. corresponde al numero de aristas esperado entre nos nodos i y j. Mientras la función $\sigma(c_i, c_j)$ es una función binaria que considera solo a las aristas cuyos vértices pertenecen a la misma comunidad.

7.5.2 Grado de pertenecía

Es definido como el ratio de la suma de pesos o numero de aristas que conectan a un nodo con otros nodos de la misma comunidad entre el peso total o numero de aristas conectadas a ese nodo en toda la red (Chen, Shang, Lv, & Fu, 2010).

$$B(i, C) = \frac{\sum_{i=C} w_{i,j}}{k_i}$$

Donde w_{ij}, es el peso de la arista entre el nodo i y el nodo j y el nodo i pertenece a la comunidad C. Además k_i,es la fuerza del nodo i.

7.6 Medidas de influencia

7.6.1 Grado de centralidad

Se define como el numero de aristas adyacentes que tiene un nodo i (Freeman, 1978). Esta métrica cuantifica la importancia estructural de un nodo, es decir, que tan importante e influyente es una persona dentro de la red a la que pertenece (Pushpa & Shobha, 2013). Por tanto, en esta aproximación se asume que el valor de un individuo depende de la cantidad de conexiones que posee (Pushpa & Shobha, 2013).

$$d(i) = \sum_j a_{ij}$$

Donde a_{ij} es la entrada de la matriz de adyacencia del grafo $G = (V, E)$ (Freeman, 1978) . Esta medida puede ser normalizada mediante la división entre el numero total de nodos de la red (Columelli, Nuñez del Prado, & Zarate Gamarra, 2016).

7.7 Análisis de los principales componentes (PCA)

Es matemáticamente definido como una transformación ortogonal lineal que permite reducir la cantidad de dimensiones, sin perder mucha información relevante (Jolliffe, 2002), es decir transforma un grupo de variables correlacionadas en un grupo mas pequeño de variables no correlacionadas llamado principales componentes (Einasto, et al., 2011). Con este objetivo usa un sistema de coordenadas de manera que la mayor varianza cae en la primera coordenada, ósea el primer componente y así sucesivamente va disminuyendo (Einasto, et

al., 2011). Posteriormente, mapea los datos originales en el nuevo espacio de componentes principales (Jolliffe, 2002).

7.8 Sistema de lógica difusa

Un sistema de lógica difusa permite manejar data numérica y lingüística de manera simultanea (Mendel, 1995). Tiene una aproximación bastante cercana a la forma de funcionar del cerebro humano para tomar decisiones. Es un mapeo no lineal de un vector de data de entrada a un valor escalar de salida en base a reglas heurísticas (Mendel, 1995); donde el antecedente y consecuente son conjuntos difusos. Un conjunto difuso, se caracteriza por contener elementos de forma parcial, es decir que la pertenecía de un elemento a un conjunto es verdadera en cierto grado (Zadeh, 1965). Las reglas de inferencia son determinadas por expertos en el campo o aprendidas por el sistema mediante un algoritmo de predicción (Mendel, 1995).

7.9 Distancia Euclidiana

La distancia euclidiana entre dos puntos en el espacio euclidiano n-dimensional (Bourbaki, 2003), se define de la siguiente manera:

$$d_E(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Donde, $P = (p_1, p_2, \dots, p_n)$ y $Q = (q_1, q_2, \dots, q_n)$ son dos puntos en el espacio euclidiano (Bourbaki, 2003).

7.10 *Cross-industry standard process for data mining* (CRISP-DM)

CRISP-DM es un modelo de minería de datos estándar y libre. Fue desarrollado por lideres de diferentes industrias con distintos puntos de vista, lo que permitió desarrollar una herramienta y modelo de aplicación neutral que incentiva las mejores practicas y la estructura necesaria para obtener buenos resultados en el proceso de minería de datos (Shearer, 2000). Según la revista Forbes es el proceso estándar de analítica mas usado en las empresas (Brown, 2015). El modelo organiza el proceso de minería de datos en seis fases, las cuales ayudan a las organizaciones a entender el proceso de minería de datos y brindan una guía

para la ejecución del proyecto (Shearer, 2000). Dichas fases se muestran en la Figura 7.10.1 y se explican a continuación:

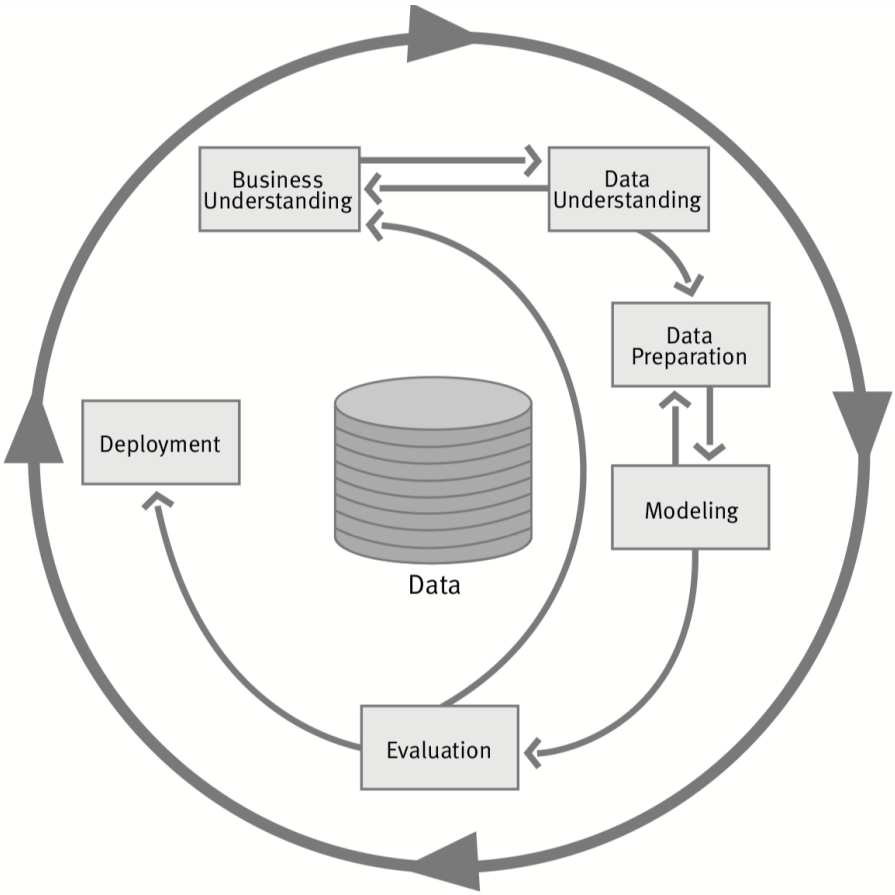

Figura 7.10.1: Fases del modelo de referencia CRISP-DM (Chapman, et al., 1999)

7.10.1 Fase 1: Comprensión del negocio

Es posiblemente la parte mas importante de cualquier proyecto de minería de datos. Consiste en entender los objetivos y requerimientos del proyecto desde la perspectiva del negocio para desarrollar un plan que permita alcanzar dichos objetivos (Chapman, et al., 1999). Solo así se llega a comprender el problema y los requerimientos necesarios como son la data por extraer, los algoritmos a utilizar, las métricas a considerar, etc. Esta fase es vital para que los practicantes de esta ciencia entienden el negocio para el cual están proponiendo una solución (Shearer, 2000).

Esta fase consiste en desarrollar cuatro actividades clave, las cuales son: determinar los objetivos de negocio, evaluar la situación actual, determinar las metas del proceso de minería de datos, realizar un plan de trabajo (Chapman, et al., 1999). Estas tareas ayudan a comprender el negocio y alinear los objetivos con el plan de acción.

7.10.2 Fase 2: Comprensión de los datos

Esta etapa inicia claramente con la recolección de datos, luego se procede con el análisis para la extracción de las principales características y problemas encontrados en los mismos (Shearer, 2000). Las actividades que conforman esta etapa pueden ser desarrollados en indistinto orden dependiendo de la necesidad de los analistas (Chapman, et al., 1999). Dichas actividades incluyen: la recolección, la descripción, la exploración de los datos y por ultimo la verificación de la calidad de estos (Chapman, et al., 1999).

7.10.3 Fase 3: Preparación de datos

La siguiente etapa, constituye todas las actividades necesarias para construir el conjunto de datos final que recibirá el modelo (Chapman, et al., 1999). Por tanto, no hay un orden preestablecido para la ejecución de estas cinco actividades, las cuales se mencionan a continuación: selección de datos, limpieza de datos, construcción de datos, integración de datos, y formateado de datos (Chapman, et al., 1999).

7.10.4 Fase 4: Modelado

En esta fase, se seleccionan varias técnicas y modelos, y se regulan los parámetros de estos, para posteriormente ser aplicados a los datos (Shearer, 2000). Usualmente hay varias técnicas para resolver el mismo tipo de problema; sin embargo, los requerimientos de cada modelo pueden variar por lo que usualmente es necesario regresar a la etapa anterior (Chapman, et al., 1999). Las actividades cruciales de este periodo son: seleccionar la técnica de modelamiento, generar diseño de evaluación, construir el modelo y finalmente, evaluar el modelo (Chapman, et al., 1999).

7.10.5 Fase 5: Evaluación

Hasta esta parte se ha construido un modelo o modelos que parecen tener un buen performance en base a la perspectiva de analítica de datos (Chapman, et al., 1999). Sin embargo, es importante evaluar si el modelo seleccionado logra alcanzar el objetivo propuesto. Una recomendación es repasar todas las etapas anteriores y analizar si alguna cuestión de negocio no esta siendo suficientemente considerada (Chapman, et al., 1999). Por consiguiente, las actividades de esta fase son: evaluación de resultados, revisión del proceso y determinación de siguientes pasos (Chapman, et al., 1999).

7.10.6 Fase 6: Despliegue

La creación del modelo no es generalmente el fin del proyecto (Shearer, 2000). El conocimiento ganado debe ser organizado y presentado de manera tal que la empresa pueda usarlo e introducirlo en el modelo de toma de decisiones (Chapman, et al., 1999). La complejidad de esta fase depende del problema y de los requerimientos de negocio. Los pasos clave de esta fase final son: despliegue del plan, plan de monitoreo y mantenimiento, generar reporte final y revisión del proyecto (Chapman, et al., 1999).

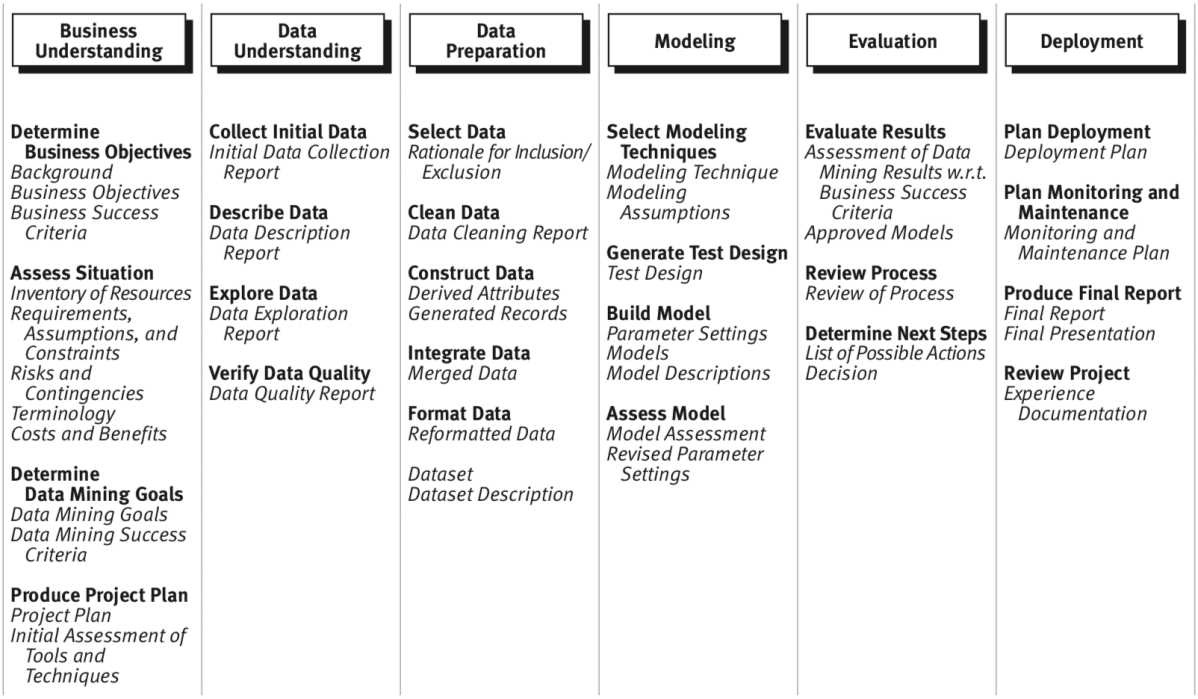

Figura 7.10.2: Tareas y resultados del modelo de referencia CRISP-DM por cada fase (Chapman, et al., 1999)

8 Metodología

En esta investigación, se utilizará la metodología CRISP-DM, pues brinda una estructura importante para el desarrollo de proyectos de minería de datos enfocados en facilitar soluciones a problemas empresariales reales (Chapman, et al., 1999). Debido a que el principal objetivo de este trabajo esta enfocado en proponer una metodología que permita mejorar la captación de clientes en la industria de telecomunicaciones del Perú, mediante el uso de técnicas de minería de datos aplicadas a grafos, como son la detección de comunidades y la identificación de usuarios mas influyentes. En ese sentido, la metodología CRISP-DM brinda el marco y los pasos necesarios para alcanzar este fin. A continuación, se desarrollan las 6 fases propuestas por la metodología.

8.1 Comprensión del negocio

En esta fase, se busca alinear los objetivos del negocio, los objetivos de la propuesta y proponer el plan para lograrlo. En ese sentido, la primera tarea es determinar los objetivos del negocio. Entonces, dado el alto nivel de competencia dentro de industria de telecomunicaciones en el Perú, las empresas buscan desarrollar estrategias efectivas y mejor enfocadas para captar clientes de la competencia. Ese viene a ser el principal objetivo del negocio. La segunda tarea consiste en evaluar la situación actual, es decir que recursos están disponibles para hacer el análisis. En consecuencia, se solicito a una empresa importante del sector datos sobre la interacción entre los abonados, llamadas y mensajes de texto, por un periodo de dos meses. La tercera tarea es determinar las metas del proceso de minería de datos en términos del negocio. En ese sentido, la meta del presente trabajo es identificar potenciales clientes para la captación en base a las interacciones entre usuarios de la red de telefonía móvil en el Perú. La tarea final de esta etapa es realizar un plan de trabajo. Este plan consiste en aplicar los pasos de la metodología CRISP-DM para alcanzar los objetivos planteado. En cuanto a las técnicas y herramientas, se propone representar la red social de telecomunicaciones mediante un grafo, por las propiedades de este, aplicar detección de comunidades y realizar un análisis de los usuarios mas influyentes en cada comunidad.

8.2 Comprensión de los datos

Esta fase propone distintas tareas que pueden ser realizadas en el orden que mejor que adapte para comprender el conjunto de datos que se tiene (Chapman, et al., 1999). Una de las tareas es la recolección de data, para el caso la empresa brindara el conjunto de datos

requerido sobre comunicación entre abonados. El cual consta de los siguientes campos: numero de origen, numero de destino, tipo de llamada, duración de la llamada, numero de llamadas y numero de mensajes de texto. Una muestra de la data se puede apreciar en la tabla 8.2.1. Otra de las tareas es la descripción de los datos, donde se busca extraer las principales características del conjunto de datos, numero de campos, distribución de variables, etc. Con estos hallazgos se creará un reporte resumen. Así se podrá determinar si los datos se prestan para el análisis pensado y permiten alcanzar los objetivos de proyecto.

Numero de origen	Numero de destino	Tipo de llamada	Duración de llamada (seg)	Numero de llamadas	Numero de sms
dad878015fc	261d09ee269	MOBILE	23	2	0
3b3af9f22ec	f228da1389ba	MOBILE	57	1	0
e5f9dfeb9b9	7ed5dc16f1b7	MOBILE	379	1	0
ecaf1439ea6	65f09eaaafd0	MOBILE	228	2	0
c7b6021317	333	SERVICE	30	2	0
e9427168e1	b22d1587830	MOBILE	456	6	0

Tabla 8.2.1: Conjunto de datos inicial

La siguiente tarea es explorar los datos, es decir hacer reportes sobre cantidad de usuarios únicos, distribuciones de las variables, frecuencia de llamadas, entre otros. El objetivo es obtener los hallazgos mas importantes sobre la data para asegurar que cumpla con los requisitos del modelo para sustentar su elección. La ultima tarea mencionada es verificar la calidad de los datos, se revisará la data en busca de valores extremos, valores perdidos, inconsistencia de los campos, entre otros.

8.3 Preparación de datos

La siguiente fase consiste en todas las actividades necesarias para construir el conjunto de datos final que será la entrada del modelo (Chapman, et al., 1999). La primera tarea aquí es la selección de datos que se usara para el análisis, en este caso serán los campos de números de origen y destino, duración de llamada, numero de llamadas y numero de mensajes. Además, dependiendo de la importancia del tipo de llamada se decidirá si trabajar con todos los tipos o solo con los catalogados como móviles. La segunda tarea consiste en limpiar la data, es decir usar un conjunto de técnicas para lidiar con los problemas identificados en la verificación de la calidad de datos.

La tercera tarea consiste en construir datos, en este caso, el foco se centra en obtener un único valor que resuma las interacciones entre los usuarios de telefonía (Chapman, et al., 1999). Con ese objetivo se probarán tres técnicas que reciben las variables iniciales de interacción y las transforman en un valor que representa la relación entre dos usuarios. La primera técnica es la de PCA, que al ser una transformación ortogonal lineal permite reducir las variables a sus principales componentes y encontrar una función lineal en base a estos en el espacio transformado. Entonces el resultado de la función será el valor resumen. La segunda técnica es la de distancia euclidiana entre usuarios, es usada en la investigación "*Social Network Classifier for Churn Prediction in Telecom Data*" con el mismo objetivo (Pushpa & Shobha, 2013). Se inicia construyendo un vector con los tres valores de interacción entre dos usuarios tanto de ida como de vuelta. Luego se calcula la distancia entre estos dos vectores y ese es el valor asignado a la relación entre usuarios. La tercera técnica consiste en utilizar un sistema de lógica difusa, que es un mapeo no lineal de la data de la data de entrada y de salida. Las variables de entradas pertenecen a conjuntos difusos y mediante reglas se obtiene el valor de la variable de salida que también pertenece un conjunto difuso. El valor resultante será asignado a la relación entre usuarios.

La cuarta tarea consiste en la integración de los datos, para el caso, se construirán tres modelos de grafos que representen la relación entre los usuarios de telefonía móviles, donde los usuarios se identifican con el numero de teléfono y son personificados mediante los vértices del grafo, mientras las aristas son las relaciones entre ellos y la ponderación de estas será el valor obtenido por cada uno de los métodos del paso anterior. La quinta y ultima tarea es el formateo de la data, para el caso no es necesario pues el formato es el indicado.

8.4 Modelado

En esta fase, se aplican varios modelos y técnicas, y se calibran los parámetros para obtener un buen resultado (Chapman, et al., 1999). La primera tarea es seleccionar el modelo a utilizar, para detectar comunidades existen una diversidad de algoritmos propuestos; sin embargo, en este trabajo se probarán los siguientes algoritmos: *Louvain, Spin glass, Leading eigen vector y Label propagation*. La principal razón radica en que en las investigaciones que hicieron un análisis comparativo entre diversos métodos, estos fueron los que obtuvieron mejores resultados en cuanto a modularidad y tiempo de procesamiento al trabajar con grafos complejos. Un resumen de estos valores se encuentra en la tabla 8.4.1. Después se utilizará la medida de grado de centralidad para determinar cuales son los nodos mas influyentes de las comunidades.

La segunda tarea consiste en diseñar un modelo de evaluación (Chapman, et al., 1999), el cual usara la medida de modularidad y grado de pertenencia para evaluar cual de los algoritmos detecto mejor las comunidades. También se tendrá en cuenta el tiempo de procesamiento pues dada la gran cantidad de datos el algoritmo debe ser eficaz. La tercera tarea consiste en crear los modelos, estos serán creados en Python, de igual manera la evaluación. La cuarta tarea consiste en evaluar los resultados en base al modelo de evaluación propuesto y encontrar el mejor resultado.

Conjunto de datos	Numero de nodos	Numero de aristas	Algoritmos							
			Louvain		Fast greedy		Leading eigen vector		Label propagation	
			Q	T	Q	T	Q	T	Q	T
Facebook	4039	88234	0.834	0.1	0.774	1.53	0.799	-	0.814	0.081
Cond-2003	31163	15751	0.76	0.323	0.678	30.15	-	-	0.659	23.35
Youtube	1134890	2987624	0.685	14.14	-	-	-	-	-	-
Amazon	334863	925872	0.809	4.38	0.735	2596	-	-	-	-
DBLP	317080	1049866	0.925	4.68	0.876	736	-	-	-	-

Tabla 8.4.1: Comparación de modelos en base a modularidad y tiempo (Chejara & Godfrey, 2017) (Garg & Rani, 2017)

8.5 Evaluación

Esta etapa consiste en repasar la correcta ejecución de las etapas anteriores y asegurar que el modelo propuesto cumpla con los objetivos de proyecto (Chapman, et al., 1999). Con este objetivo se siguen tres actividades. La primera actividad consiste en evaluar los resultados del modelo, es decir que tan bien logra resolver el problema de negocio y determina si algunas razones de negocio el modelo es deficiente. La segunda actividad consiste en revisar el proceso de minería de datos usado para ver si hay algún factor que se haya omitido. La tercera actividad consiste en decidir los pasos futuros en cuanto al proyecto.

8.6 Despliegue

Esta ultima etapa se presentará una representación grafica de grafo y un informe con las conclusiones obtenidas a la empresa de telefonía móvil para la captación de clientes.

9 Bibliografía

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*.

Bonchi, F., Castillo, C., Gionis, A., & Jaimes, A. (2011). Social Network Analysis and Mining for Business Applications. *ACM TIST*, 22:1-22:37.

Bourbaki, N. (2003). *Topological Vector Spaces.* Springer.

Brown, S. M. (2015). What IT Needs To Know About The Data Mining Process. *Forbes*.

Capgemini Consulting. (2013). *The Digital Advantage: How digital leaders outperform their peers in every insudtry.* Boston: mitsloan.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & (DaimlerChrysler), R. W. (1999). *CRISP-DM 1.0.* SPSS.

Chejara, P., & Godfrey, W. W. (2017). Comparative analysis of community detection algorithms. *Information and Communication Technology (CICT), 2017 Conference on.* Gwalior, India, India: IEEE.

Chen, D., Shang, M., Lv, Z., & Fu, Y. (2010). Detecting overlapping communities of weighted networks via a local algorithm. *Physica A: Statistical Mechanics and its Applications, vol. 389*, 4177–4187.

Cheng-Shang, C., Duan-Shin, L., Li-Heng, L., & Sheng-Min, L. (2017). A Probabilistic Framework for Structural Analysis and Community Detection in Directed Networks. *IEEE/ACM Transactions on Networking*, 31-46.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E, vol 70*.

Columelli, L., Nuñez del Prado, M., & Zarate Gamarra, L. (2016). Measuring Churner Influence on Pre-paid Subscribers Using Fuzzy Logic. *2016 XLII Latin American Computing Conference (CLEI)* (págs. 1-10). IEEE.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A., & Joshi, A. (2008). Social ties and their relevance to churn in mobile telecom networks. *EDBT.* EDBT.

Einasto, M., Liivamagi, L., Saar, E., Einasto, J., Tempel, E., Tago1, E., & Martinez, V. (2011). Principal component analysis. *Astronomy & Astrophysics manuscript*.

Everis. (2017). *Connected Telco LATAM: la perspectiva del cliente para la transformación digital.* Mexico: Everis.

EY. (2017). *Las ventajas de la disrupcion.* Lima: EY.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 215-239.

Garg, N., & Rani, R. (2017). A comparative study of community detection algorithms using graphs and R . *Computing, Communication and Automation (ICCCA), 2017 International Conference on.* Greater Noida, India: IEEE.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences, vol. 99*, 7821-7826.

GSMA. (2016). *La Economía Móvil: América Latina 2016.* London: GSM Association.

GSMA Intelligence. (2017). *Economía Movil 2017: América Latina y Caribe.* London: GSM Association.

Hendrikx, H., & Nuñez del Prado, M. (2016). Toward a Route Detection Method base on Detail Call Records. *2016 IEEE Latin American Conference on Computational Intelligence (LA-CCI).* Cartagena, Colombia: IEEE.

Indrawati, & Alamsyah, A. (2017). Social network data analytics for market segmentation in Indonesian telecommunications industry . *Information and Communication*

Technology (ICoIC7), 2017 5th International Conference on. Malacca City, Malaysia: IEEE.

Jaradat, A., & Al-Zoubi, R. M. (2017). Community detection using network structure. *Information Technology (ICIT), 2017 8th International Conference on.* Amman, Jordan: IEEE.

Jolliffe, I. (2002). *Principal Component Analysis.* New York: Springer-Verlag.

Linoff, G., & Berry, M. (2011). *Data Mining Techniques.* Indianapolis: Wiley Publishing.

Mendel, J. (1995). Fuzzy logic systems for engineering: a tutorial. *Proceedings of the IEEE* , 345 - 377.

Mothe, J., Mkhitaryan, K., & Mariam, H. (2017). Community detection: Comparison of state of the art algorithms. *Computer Science and Information Technologies (CSIT), 2017.* Yerevan, Armenia: IEEE.

Newman, M. (2011). Network: An Introduction. *Oxford University Press*.

Ortiz, J. (19 de Julio de 2017). ¿Cómo debe ser el nuevo modelo de operaciones B2B en la industria de telecomunicaciones? *Computerworld*.

OSIPTEL. (2009). *Portabilidad numerica: OSIPTEL*. Obtenido de OSIPTEL: https://www.osiptel.gob.pe/documentos/33079-portabilidad-numerica

OSIPTEL. (2017). *Reporte Estadístico .* Lima: OSIPTEL.

OSIPTEL. (9 de Febrero de 2018). *Noticias.* Obtenido de OSIPTEL: https://www.osiptel.gob.pe/noticia/portalbilidad-movil-473mil-usuarios-operadora

Pandapotan, I., Alamsyah, A., & Paryasto, M. (2015). Indonesian Music Fans Group Identification using Social Network Analysis. *3rd International Conference on Information and Communication Technology.* Indonesia: Kaskus Forum.

Pushpa, R., & Shobha, G. (2013). Social network classifier for churn prediction in telecom data. *Advanced Computing and Communication Systems (ICACCS), 2013 International Conference on.* Coimbatore, India: IEEE.

PWC. (2016). *Análisis Industrial: Sector Telco.* PWC.

Ruifang, L., Shan, F., Ruisheng, S., & Wenbin, G. (2014). Weighted graph clustering for community detection of large social networks. *Procedia Computer Science, Vol. 31*, 85-94.

Schwab, K. (12 de Diciembre de 2015). The Fourth Industrial Revolution. *Snapshot*.

Seungyo, R., & Dongseung, K. (2016). Quick Community Detection of Big Graph Data Using Modified Louvain Algorithm . *High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2016 IEEE 18th International Conference on.* Sydney, NSW, Australia: IEEE.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 13-31.

The Boston Consulting Group. (2018). *The most innovative companies 2018.* Boston.

The McKinsey Quartely. (2007). *How companies approach innovation: A McKinsey Global Survey.*

Trudeau, R. J. (1993). *Introduction to Graph Theory.* Dover Pub.

Upadhyay, A., & Singh, M. (2017). Community detection based on graph models of data. *Communication and Signal Processing (ICCSP), 2017 International Conference on.* Chennai, India: IEEE.

Varun, E., & Pushpa, R. (2016). Telecommunication community detection by decomposing network into n-cliques. *Emerging Technological Trends (ICETT), International Conference on.* Kollam, India: IEEE.

Wei, C., & Chiu, I. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications, 23*, 103-112.

West, D. (2001). *Introduction to Graph Theory .* Illinois: PRENTICE HAL.

Wilson, R. J. (1979). *Introduction to Graph Theory.* Essex: Logman Group.

World Bank Group. (2018). *Global Economic Prospects.* Washington: World Bank Group.

Wu, X., & Liu, Z. (2008). How community structure influences epidemic spread in social networks. *Physica A: Statistical Mechanics and its Applications*, 623-630.

Zadeh, L. (1965). Fuzzy sets. *Information and Control Volume 8, Issue 3*, 338-353.