UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
Departamento de Ingeniería Informática

MODELO PREDICTIVO DEL DESEMPEÑO DE BÚSQUEDA DE INFORMACIÓN EN LÍNEA EN ESTUDIANTES DE EDUCACIÓN BÁSICA

Propuesta de Tesis

Nombre: Gonzalo Javier Martinez Ramirez
R.U.T.: 18.045.598-1
Año Ingreso: 2010
Teléfono: (+56) 9 96112973
E-mail: gonzalo.martinez@usach.cl
Profesor patrocinador: Roberto Ignacio González Ibañez

Santiago – Chile
2017

RESUMEN

El presente documento corresponde a la propuesta de tesis para la carrera de Ingeniería Civil en Informática y Magister en Ingeniería Informática, cuyo título es “Modelo predictivo del desempeño de búsqueda de información en línea en estudiantes de educación básica”. A continuación se introduce el problema a resolver a lo largo del trabajo, las herramientas y metodologías a emplear para abordar el problema.

Durante la última década, debido a los rápidos avances de las tecnologías de la información y comunicación ha aumentado la cantidad de recursos digitales en Internet, la diversidad de fuentes de información, y además, se ha facilitado el acceso a estos. Asimismo, las búsquedas *web* han pasado a ser parte de las tareas comunes que realizan los estudiantes de los planteles educativos. Considerando la diversidad de fuentes de información y tipos de recursos en línea, resulta necesario desarrollar competencias informacionales durante el proceso de formación en los distintos niveles educativos (básica, media y universitaria).

En el marco del proyecto iFuCo (*Enhancing learning and teaching future competences of online inquiry in multiple domains*), formado por investigadores de Chile y Finlandia, el cual desea investigar y modelar los comportamientos y competencias de investigación en línea de estudiantes de enseñanza básica, se propone la construcción de un modelo de predicción del desempeño de búsqueda de información en línea en estudiantes de educación básica el cual se vaya perfeccionando a través del registro de datos históricos y ofrezca retroalimentación tanto al estudiante como al docente.

La investigación será guiada por la metodología Descubrimiento de Conocimiento en Base de Datos (conocido como KDD, las iniciales de *Knowledge Discovery in Databases*) con el fin de descubrir patrones en los datos que permitan la creación de un modelo de predicción del desempeño de búsqueda de información. Además, para apoyar el proceso de investigación, se desarrollará un módulo de la plataforma NEURONE (*oNlinE inqUiry expeRimentatiON systEm*). El módulo propuesto alimentará y perfeccionará el modelo de predicción y entregará predicciones de forma continua. El desarrollo de esta plataforma se guiará bajo la metodología Desarrollo de Rápido de Aplicaciones (conocido como RAD, las iniciales de *Rapid Application Development*) la cual se orienta a un desarrollo iterativo e incremental para la rápida construcción de prototipos de *software*.

Palabras Claves: Alfabetización informacional, Competencias informacionales, Estrategias de intervención.

TABLA DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. OBJETIVOS Y ALCANCES DE LA SOLUCIÓN

1.1. OBJETIVO GENERAL

Diseñar y evaluar un modelo predictivo del desempeño de búsqueda de información en línea de estudiantes de enseñanza básica.

1.2. OBJETIVOS ESPECÍFICOS

1. Realizar una revisión bibliográfica sobre trabajos recientes relacionados con minería de datos en el contexto educacional.
2. Realizar una exploración, limpieza, pre-procesamiento y transformación de los datos recopilados por la plataforma NEURONE (acrónimo de *oNlinE inqUiry expeRimentatiON systEm*).
3. Seleccionar características de comportamiento de búsqueda de los estudiantes para la construcción de modelos predictivos.
4. Construir modelos para la predicción del desempeño de búsqueda en línea de estudiantes de educación básica.
5. Evaluar los modelos predictivos del desempeño de búsqueda en linea de estudiantes de educación básica.
6. Implementar los modelos predictivos en la plataforma NEURONE.

2. DESCRIPCIÓN DEL PROBLEMA

2.1. MOTIVACIÓN

La alfabetización informacional (conocida en inglés como *information literacy*) es definida como "el grupo de habilidades en las que se requiere reconocer cuándo la información es necesaria y tener la habilidad de encontrar, evaluar y usar efectivamente dicha información necesaria"[1] (Association *et al.*, 2000, p. 2). Durante la última década, debido a los rápidos avances de las tecnologías de la información y comunicación (TICs) ha aumentado la cantidad de recursos digitales en Internet y además se ha facilitado el acceso a ellos. Estos avances han provocado una brecha entre el ser humano y la habilidad de reconocer cuando la información es necesaria para satisfacer su necesidad de búsqueda, la cual se puede asociar principalmente a dos razones: En primer lugar, las competencias de alfabetización informacional no son enseñadas ni reforzadas a temprana edad. Segundo, las búsquedas *web* han pasado a ser parte de las tareas comunes que realizan los estudiantes, disminuyendo las visitas a bibliotecas y el uso de fuentes revisadas.

Considerando la diversidad de fuentes de información y tipos de recursos en línea, resulta necesario desarrollar competencias informacionales durante el proceso de formación en los distintos niveles educativos (básica, media y universitaria). La enseñanza de la alfabetización informacional se imparte principalmente por bibliotecas universitarias, y en menor medida en la etapa escolar obligatoria (Weiner, 2014). En Chile, la enseñanza de competencias informacionales es cubierta en bibliotecas universitarias y cursos introductorios de mallas universitarias (Marzal & Saurina, 2015). De acuerdo con Urra y Castro (2016), los estudiantes universitarios de Chile presentan problemas con las competencias informacionales, ya que no aplican la búsqueda de información de forma crítica ni sistemática. Una de las posibles causas de por qué los estudiantes tienen dificultades con estas competencias es el hecho de que en los colegios y en el inicio de su educación se prioriza la reiteración de la información. Las consecuencias de no considerar cuándo y por qué se necesita la información, dónde encontrarla y cómo evaluarla, se ven reflejadas en la evaluación crítica de la información, y en el desempeño de los estudiantes (Urra & Castro, 2016).

Head (2013, p. 475) a través de encuestas a estudiantes universitarios, establece que al momento de realizar investigaciones el 84 % de los estudiantes universitarios utiliza como fuente primaria de búsqueda Wikipedia[2] y un 87 % consulta a sus amigos, sin verificar la veracidad de la información

[1] Traducción libre.
[2] https://es.wikipedia.org/

que obtienen. Como consecuencia, los estudiantes al no ser instruidos en parafrasear, resumir o citar fuentes revisadas, caen al plagio de forma premeditada o no intencionada.

A partir de los argumentos anteriormente expuestos, respecto a la enseñanza de competencias de alfabetización informacional se puede observar que no ha sido completamente satisfecha y la brecha entre los usuarios e alfabetización informacional permanece abierta.

Esta propuesta de tesis se enmarca en el contexto del proyecto de investigación "*Enhancing Learning and Teaching Future Competences of Online Inquiry in Multiple Domains*"[3] (iFuCo, desde ahora en adelante), el cual pretende abordar la temática de la alfabetización informacional en estudiantes de enseñanza básica (5to y 6to básico) con el objetivo de estudiar sus patrones de comportamiento y ofrecer mallas curriculares y asignaturas adecuadas respecto al tema (Sormunen *et al.*, 2017).

2.2. REVISIÓN DE LA LITERATURA

Esta sección tiene como objetivo entregar las bases teóricas, conceptuales y empíricas que soportan el desarrollo de esta investigación. En primer lugar, se presenta el marco conceptual donde se entregan las definiciones y conceptos necesarios para abordar esta investigación. Finalmente, se presenta el estado del arte relacionado con el tema.

2.2.1. Marco conceptual

Alfabetización informacional

La alfabetización informacional (conocida en ingles *information literacy*) es definida como "el grupo de habilidades en las que se requiere reconocer cuándo la información es necesaria y tener la habilidad de encontrar, evaluar y usar efectivamente dicha información necesaria" [4] (Association *et al.*, 2000, p. 2). Es un campo que cubre varias áreas, entre las que se destaca la alfabetización digital, las habilidades de uso de bibliotecas, la ética informacional, la lectura crítica, el pensamiento crítico, los derechos de autor, la seguridad y privacidad, entre otras. A través del estudio de estas áreas como factores que influyen a la alfabetización informacional se puede obtener una visión clara de cómo los estudiantes llevan a cabo sus tareas de obtención y selección de información.

[3] https://www.researchgate.net/project/Enhancing-learning-and-teaching-for-future-competences-of-online-inquiry-in-multiple-domains-iFuCo
[4] Traducción libre.

Competencias de investigación en línea

Se definen las competencias de investigación (*inquiry skills* en inglés) como "las habilidades para explorar preguntas, para poder reunir, interpretar y sintetizar diferentes tipos de información y datos, además de desarrollar y compartir una explicación para responder preguntas dadas" [5] (Council *et al.*, 2000, p. 13). En base a este concepto nacen las competencias de investigación en línea (conocidas en inglés como *online inquiry skills*), que son una instancia específica de las competencias de investigación, pero aplicada sobre información disponible en línea (Quintana, Zhang & Krajcik, 2005).

Las competencias de investigación en línea involucran una serie de actividades cognitivas, como generar una pregunta de investigación, buscar información relevante en colecciones digitales, evaluar y seleccionar la información encontrada, e integrar coherentemente la información seleccionada para responder la pregunta original (Eisenberg & Berkowitz, 1990).

Minería de datos educacional

La minería de datos utiliza una combinación de bases de conocimientos explícita, conocimientos analíticos complejos y conocimiento de campo para descubrir las tendencias y los patrones ocultos, estas tendencias y patrones forman la base de los modelos predictivos que permiten a los analistas realizar nuevas observaciones de los datos existentes (Luan, 2002). La gran cantidad de información generada hoy en día por los estudiantes permite que la minería de datos obtenga datos relevantes y, a través de métodos estadísticos y otras herramientas, relacione la información para conocer si el proceso de enseñanza aprendizaje ha dado resultados positivos.

Mining (2012, p. 9) define la minería de datos educacional (MDE, desde ahora en adelante) como "la teoría que desarrolla métodos, aplica técnicas estadísticas y de aprendizaje automático para analizar los datos recogidos durante el proceso de la enseñanza y aprendizaje" [6]. Actualmente, los usos más generales que se le están dando a la MDE básicamente se enfocan en mejorar la estructura del conocimiento y determinar el apoyo pedagógico al estudiante.

[5] Traducción libre.
[6] Traducción libre.

2.2.2. Estado del arte

Usos de la minería de datos educacional

Actualmente, la aplicación de la MDE radica en universidades, tales como Paul Smith's College, la cual utiliza sus datos históricos para mejorar las tasas de retención de alumnos (Bichsel, 2012). En este contexto, University of Georgia desarrolló un modelo para predecir la tasa de graduación y abandono estudiantil, el cual se alimenta en base la información recopilada (Morris, Wu & Finnegan, 2005). Finalmente, la Purdue University han usado MDE para determinar que la evaluación en etapas tempranas y de forma frecuente permite cambiar los hábitos de los estudiantes con calificaciones bajo la media en cursos introductorios, en base a este trabajo, el mismo equipo de investigación desarrollo un sistema de alerta académica temprana para saber el desempeño de los estudiantes (Baepler & Murdoch, 2010).

Merceron y Yacef (2005) establece cómo los algoritmos de minería de datos pueden escoger información pedagógica importante. El conocimiento obtenido ayuda a mejorar el cómo administrar la clase, como el alumno aprende, y cómo proporcionar un feedback a los alumnos. Basado en este trabajo, Abdullah, Malibari y Alkhozae (2014) realiza un sistema de predicción del rendimiento de los estudiantes basado en la actividad actual y mediciones anteriores clasificando cuales estudiantes rendirán bien y los que no.

Henrie, Halverson y Graham (2015) clasifica los datos generados por los estudiantes en un sistema computacional en tres categorías: comportamiento, cognitivas y emocionales. El comportamiento de los estudiantes es una de las más estudiadas, en esta categoría se estudian las variables cuantitativas: resultados de consultas realizadas en un motor de búsqueda, teclas presionadas, rastreo ocular, tareas de búsqueda, esfuerzo (intentos por finalizar tareas asignadas), participación, tiempos de permanencia o de respuestas y uso de sitios *web*, entre otros. Tomando tiempos de permanencia y el uso de sitios, Shah, Hendahewa y González-Ibáñez (2016) presenta diversas métricas para evaluar el rendimiento de la búsqueda de información, en base a las distintas acciones hechas por usuarios. Tales métricas se usan con el objetivo de pronosticar la probabilidad de un usuario de tener éxito en el futuro, en base a su desempeño actual.

Técnicas utilizadas en la minería de datos educacional

Chen y Liu (2008) evalúa el rendimiento académico de estudiantes de pregrado estudiando datos académicos del Departamento de Ciencias de la Computación de National Defence University of Malaysia (NUDM) utilizando una combinación de técnicas de minería de datos, como ANN

(*Artificial Neural Network*) y árboles de decisión como un método de clasificación con el que se producen ocho reglas para la identificación automática de los estilos cognitivos de los estudiantes basados en sus patrones de aprendizaje. Los hallazgos obtenidos se aplicaron para desarrollar un modelo que pueda apoyar el desarrollo de programas educativos *web*.

Moreno-Clari, Arevalillo-Herraez y Cerveron-Lleo (2009) predice la probabilidad de que los estudiantes de acuerdo a sus registros académicos históricos fallen en un curso *online* en Moodle[7] haciendo uso de las técnicas de maximización de la entropía, el método de agrupamiento K-means y X-means, usando el *software* WEKA.

Lahtinen, Ala-Mutka y Järvinen (2005) estudia las dificultades de aprender programación con el objetivo de crear material adecuado para introducir el curso a los estudiantes utilizando el método de agrupamiento K-means y *Hierarchical clustering* (más conocido como Ward's *clustering*, a través de este estudio se obtuvo las dificultades que sufren los estudiantes al momento de enfrentar tareas de programación. Basado en este trabajo Akinola, Akinkunmi y Alo (2012) aplica ANN para predecir el resultado de los cursos de programación en estudiantes de pregrado basados en su historial académico, los resultados de este estudio muestran que los estudiantes con un conocimiento a priori de física y matemática tienen mejor desempeño en los cursos que el resto.

Borkar y Rajeswari (2014) evalúa el rendimiento de los estudiantes, donde selecciona algunos atributos mediante minería de datos, haciendo uso de una red neuronal multicapa perceptrón y usando una validación cruzada selecciona las características más influyentes, estableciendo las reglas necesarias para poder detectar las características necesarias para poder predecir el rendimiento de los estudiantes. Basado en los métodos propuestos y el mismo conjunto de datos de este trabajo, Jayakameswaraiah y Ramakrishna (2014) compara los métodos de perceptrón multicapa, Naive Bayes, SMO y J48 con el objetivo de obtener el mejor algoritmo de clasificación y predicción entre todos ellos. De todos los métodos comparados, el método de perceptrón multicapa obtuvo un *accuracy* de 75 %.

Identificación de factores

Borkar y Rajeswari (2013) sugiere un método de evaluación del rendimiento de los estudiantes usando reglas asociativas de minería de datos, estimando el resultado de los estudiantes basado en la asistencia a sus cursos y su avance académico. Basado en este trabajo, Shazmeen, Baig y Pawar (2013) evalúa el rendimiento de diferentes algoritmos de clasificación y análisis predictivo, proponiendo técnicas de preprocesamiento de datos para lograr mejores resultados.

[7] https://moodle.org/

Oskouei y Askari (2014) identifica que los factores que afectan el rendimiento de los estudiantes de primer semestre de la carrera de Ingeniería de Software de Irán e India, aplicando técnicas de clasificación y predicción para mejorar la precisión de las predicciones de los resultados de los estudiantes. Los resultados muestran que los factores de género, entorno familiar, nivel de educación de los padres, y el estilo de vida afectan el rendimiento académico de los estudiantes independiente del país.

Tal como se muestra en los antecedentes anteriores, las investigaciones en MDE se realizan mayoritariamente en aprendizaje *online* y en casos puntuales en educación superior, por lo que es limitada la información respecto a educación básica o media, específicamente en la predicción de errores y fracaso escolar. Para mayor información de trabajos relacionados con la MDE, consultar los siguientes *reviews* (Anoopkumar & Rahman, 2016; Dutt, Ismail & Herawan, 2017; Shahiri, Husain *et al*., 2015; Sukhija, Jindal & Aggarwal, 2015).

2.3. DEFINICIÓN DEL PROBLEMA

En el contexto de la enseñanza de la alfabetización informacional, las evaluaciones de los cursos se centran principalmente en los resultados de los estudiantes sin tomar en cuenta el proceso formativo y factores asociados que podrían influir directa o indirectamente sobre los resultados finales y el desempeño de búsqueda de los alumnos.

En el contexto del proyecto de investigación iFuCo, el cual pretende realizar un análisis cuantitativo y cualitativo de la alfabetización informacional y las competencias de búsqueda en línea en estudiantes de enseñanza básica[8] en los países de Chile y Finlandia, surgen las siguientes interrogantes (*research questions*, RQ desde ahora en adelante):

RQ 1 ¿De qué manera se puede estimar durante el proceso de aprendizaje de competencias informacionales la influencia de factores cognitivos (por ejemplo, los *keystrokes*) y emocionales (por ejemplo, valencia) en el desempeño de búsqueda de la información de los estudiantes?

RQ 2 ¿En qué medida es posible detectar situaciones anormales de conducta, y determinar las causas que llevan a un estudiante a fallar durante el proceso de búsqueda de información?

RQ 3 ¿De qué manera se puede implementar un módulo de clasificación y predicción del desempeño de los estudiantes en la búsqueda de información en herramientas de apoyo de la alfabetización informacional para proporcionar una retro evaluación oportuna a estudiantes y docentes?

[8] En otros países es conocido como enseñanza primaria.

3. DESCRIPCIÓN DE LA SOLUCIÓN PROPUESTA

3.1. CARACTERÍSTICAS DE LA SOLUCIÓN

La solución consiste en incorporar un módulo en NEURONE (González-Ibáñez, Gacitua, Sormunen & Kiili, 2017) que clasifique y prediga de forma continua el desempeño de búsqueda de los estudiantes de enseñanza básica en un curso de alfabetización informacional, específicamente en el tema de investigaciones en línea[1] (*online inquiry*).

Los datos son recopilados y almacenados por NEURONE, estos datos provienen de registros del proceso de búsqueda de información en línea en un sistema cerrado, los cuales son: historial de navegación, consultas realizadas, movimientos del *mouse*, escritura por teclado, número de *clicks* y tiempos de permanencia en páginas *web*. Además, se conoce con anticipación los documentos y párrafos ideales a seleccionar por parte de los estudiantes.

El módulo propuesto hará uso de Apache Spark[2], el cual es un *framework* de código abierto para el procesamiento de datos masivos, el cual incluye librerías de minería de datos y aprendizaje automático. Este módulo se conectará con el sistema NEURONE, funcionando como una extensión del mismo, consultando su base de datos, alimentando y perfeccionando el modelo.

El ciclo de construcción, evaluación y optimización del modelo se ilustra en la Figura 3.1, donde a través de los datos históricos obtenidos de NEURONE construye el modelo, lo evalúa y lo optimiza en un proceso continuo, entregando como resultado la clasificación y predicción del desempeño de búsqueda de información actual del estudiante. Dentro de este proceso cíclico, el modelo será optimizado de forma continua a través de reentrenamiento.

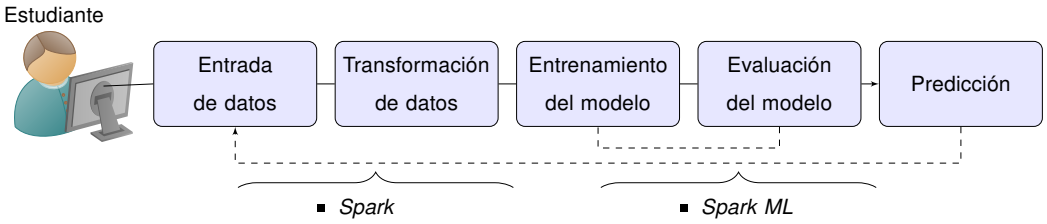

Figura 3.1: Ciclo de construcción y perfeccionamiento del modelo a través de reentrenamiento

Fuente: Elaboración propia, (2017)

[1] Traducción libre.
[2] https://spark.apache.org/

3.2. PROPÓSITO DE LA SOLUCIÓN

El propósito de la solución consiste en proveer evaluaciones de desempeño de búsqueda oportunas que permitan a los docentes aplicar acciones correctivas durante el proceso de formación y desarrollo de competencias informacionales en cursos de alfabetización informacional.

Con el módulo propuesto en este trabajo, el docente obtiene una estimación temprana del desempeño del estudiante en el proceso de búsqueda de información, de tal forma que él pueda guiar al estudiante en el proceso. Tal como ilustra la Figura 3.2, el estudiante interactúa con el sistema educacional, en este caso NEURONE y la plataforma propuesta a través de técnicas de minería de datos informa al docente de los patrones y predicciones del desempeño de búsqueda del estudiante con el objetivo de ayudar en la toma de decisiones al docente correspondiente, para diseñar y planificar de mejor forma la entrega de contenidos hacia el estudiante.

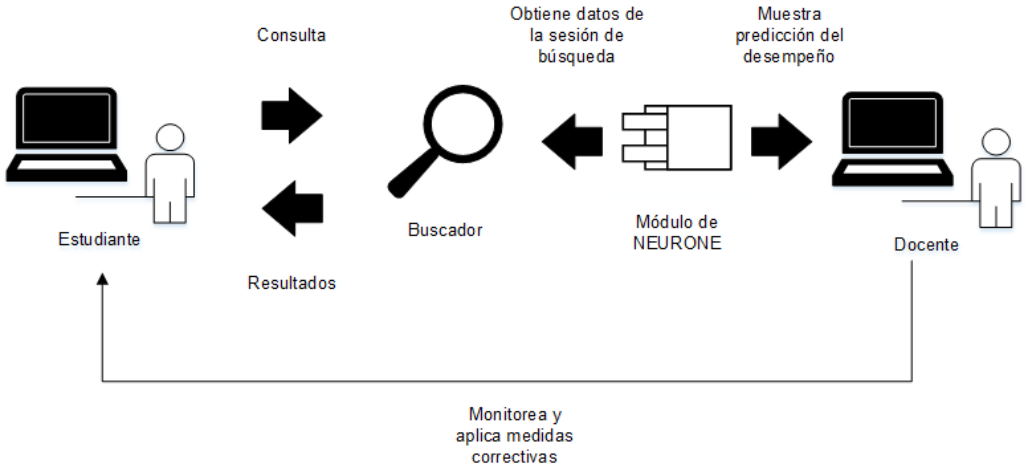

Figura 3.2: Proceso de búsqueda de información de un estudiante

Fuente: Elaboración propia, (2017)

3.3. ALCANCES Y LIMITACIONES DE LA SOLUCIÓN

Los modelos se construyen a partir de un conjunto de datos específicos, estos datos tienen su propio contexto y origen que limitan la generalización de los modelos a construir. A continuación, se describen las principales limitaciones y alcances de la solución.

1. El curso de alfabetización informacional y sus respectivos registros de datos, pertenecen al proyecto iFuCo, el cual es un trabajo colaborativo entre universidades de Finlandia (University

of Tampere, University of Jyväskylä y University of Turku) y de Chile (Universidad de Santiago de Chile y Pontificia Universidad Católica de Chile).

2. Los registros de datos provienen de un estudio enmarcado en un curso de alfabetización en información, aplicado al área de Ciencia y Ciencias Sociales, en ambos países.

3. Los datos son recolectados y almacenados por un sistema externo llamado NEURONE (*oNlinE inqUiry expeRimentatiON systEm*), trabajo de memoria de un estudiante de la carrera de Ingeniería de Ejecución en Computación e Informática de la Universidad de Santiago de Chile.

4. La solución funciona como un sistema predictor del desempeño del estudiante en la búsqueda de información, sin ofrecer acciones correctivas en caso de bajo desempeño.

4. METODOLOGÍA, HERRAMIENTAS Y AMBIENTE DE DESARROLLO

4.1. METODOLOGÍA A USAR

El presente proyecto presenta una componente de investigación y desarrollo de *software* (I+D), esto debido a la relación que existe entre ambas componentes, la investigación necesita una herramienta de *software* de apoyo que permita recibir los datos de NEURONE, alimentar el modelo de predicción y que permita al usuario interactuar con resultados de la predicción realizada.

La componente de investigación del proyecto será guiada por la metodología Descubrimiento de Conocimiento en Base de Datos (conocido como KDD, las iniciales de *Knowledge Discovery in Databases*) (Fayyad, Piatetsky-Shapiro & Smyth, 1996), mientras que la componente de desarrollo será guiada por la metodología de desarrollo de *software* Desarrollo de Rápido de Aplicaciones (conocido como RAD, las iniciales de *Rapid Application Development*) (Martin, 1991). A continuación, se explica el uso de ambas metodologías en el trabajo propuesto.

4.1.1. Metodología usada en la investigación

Respecto a la componente de investigación, esta será guiada bajo la metodología KDD, la cual se define como “un proceso no trivial de identificar patrones en los datos que sean válidos, novedosos, potencialmente útiles y finalmente comprensibles” (Fayyad *et al.*, 1996, p. 5). En primer lugar, se seleccionan y limpian los datos que se deben extraer para poder realizar el modelado del comportamiento de búsqueda. Luego, se transforman los datos y se realiza minería de datos sobre ellos para buscar los patrones de interés que pueden expresarse como un modelo o que expresen dependencia de los datos. Finalmente, se identifican los patrones realmente interesantes que representan el conocimiento, usando diferentes técnicas, incluyendo análisis estadísticos para posteriormente interpretar los datos obtenidos.

4.1.2. Metodología usada para el desarrollo

Respecto a la componente de desarrollo de *software*, se toma en cuenta las condiciones bajo las cuales se desarrolla el proyecto, las cuales se expresan a continuación:

- El sistema es de rápido desarrollo.
- El sistema es de tamaño pequeño.
- Es un proyecto cuyos requerimientos están sujetos a cambios.
- Inicialmente no existe un número total de requerimientos especificado. Estos se irán desarrollando de forma creciente durante el avance del proyecto.
- El desarrollador no cuenta con un conocimiento profundo de la arquitectura y todas las herramientas de desarrollo, por lo tanto, se requiere un tiempo de investigación y aprendizaje.
- Se requiere documentar los aspectos fundamentales de la arquitectura, una vez que se tenga un producto estable. Esta documentación permitirá la continuidad del proyecto.
- Se requiere de varias entregas funcionales, para medir el progreso del proyecto y verificar que se cumplan los objetivos propuestos.

Dado los antecedentes mencionados anteriormente, se determina que el proyecto presenta características que se ajustan bien a un modelo de desarrollo evolutivo enfocado a la generación de prototipos. A partir de esto, se recurre a un enfoque de desarrollo inspirado en la metodología RAD, metodología de desarrollo rápido que minimiza la planificación en favor de la creación rápida de prototipos. La planificación se realiza en cada iteración, permitiendo que el *software* se desarrolle más rápido y se tenga una mayor flexibilidad con los requisitos (McConnell, 1996).

4.2. HERRAMIENTAS DE DESARROLLO

Las herramientas a utilizar en el trabajo de tesis, se dividen tanto en *hardware* como en *software*.

4.2.1. Herramientas de *hardware*

El desarrollo se llevará a cabo con procesador Intel Core i7 7ma Generación *Kabylake* de 3.6 Ghz, con memoria Ram de 16 GB y 2 TB de disco duro. Además, los despliegues de prueba se realizan sobre un servidor privado virtual (VPS, por sus siglas en inglés) con el sistema operativo GNU/Linux Ubuntu Server alojado en el proveedor DigitalOcean[1].

[1] https://www.digitalocean.com/

4.2.2. Herramientas de *software*

En cuanto herramientas *software*, el desarrollo se llevará a cabo en la distribución GNU/Linux Debian en su versión 9.0. El modelo se llevará a cabo en Spark ML[2]. Para el análisis estadístico se hará uso de R. Además, cada módulo desarrollado estará contenido en contenedores de Docker para facilitar el despliegue en producción del modelo desarrollado. Todo el trabajo realizado, tanto código como documento escrito estará bajo el sistema de control de versiones Git. Finalmente, se hará uso de LaTeX para el documento escrito.

4.3. AMBIENTE

El ambiente de desarrollo del presente proyecto será tanto en el domicilio particular del candidato a tesista, como también en el Departamento de Ingeniería Informática de la Universidad de Santiago de Chile, específicamente, en el laboratorio de sistemas colaborativos.

Después de finalizar cada iteración, la retroalimentación a este proyecto es ofrecida por miembros del equipo de investigación del proyecto iFuCo y el profesor guía, quien además, brinda apoyo en aspectos tecnológicos y metodológicos. Finalmente, el equipo de desarrollo de este trabajo de tesis es unipersonal, con colaboración en fundamentos teóricos de otros tesistas y memoristas involucrados en el proyecto.

[2]https://spark.apache.org/

5. PLAN DE TRABAJO

El presente proyecto contempla 612 horas de trabajo efectivas y se realizará en el transcurso del segundo semestre del año 2017, el cual se inicia el 7 de agosto y termina el 7 de diciembre del presente año contemplando 16 semanas de trabajo. Se dispone como día de trabajo todos los días hábiles de la semana, y un horario de trabajo desde las 9:00 hasta las 18:00 hrs considerando una hora de descanso.

El plan de trabajo propuesto se muestra en la Tabla 5.1, dada las metodologías empleadas las actividades se realizan de forma secuencial. Cabe destacar que a la fecha de entrega de este informe, el alumno candidato a tesista ya ha avanzado el estado del arte e investigación de tecnologías.

Tabla 5.1: Plan de trabajo propuesto

Elaboración propia, (2017)

Actividad	Duración (HH)
Actualización del estado del arte centrado en trabajos recientes	1
Exploración, preprocesamiento y transformación de los registros del proceso de búsqueda de información obtenidos por NEURONE	3
Definición de las características para la construcción de modelos de predicción del desempeño de búsqueda de estudiantes	1
Comparación y selección de los algoritmos y/o técnicas de minería de datos para la construcción de modelos	1
Validación de los algoritmos y/o técnicas con datos conocidos	5
Evaluación de los modelos construidos utilizando métricas de desempeño	2

REFERENCIAS BIBLIOGRÁFICAS

Abdullah, A., Malibari, A. & Alkhozae, M. (2014). Student's performance prediction system using multi agent data mining technique. *International Journal of Data Mining & Knowledge Management Process*, *4*(5), 1. (citado en página 5).

Akinola, O., Akinkunmi, B. & Alo, T. (2012). A data mining model for predicting computer programming proficiency of computer science undergraduate students. (citado en página 6).

Anoopkumar, M. & Rahman, A. M. Z. (2016). A review on data mining techniques and factors used in educational data mining to predict student amelioration. En *Data Mining and Advanced Computing (SAPIENCE), International Conference on* (pp. 122-133). IEEE. (citado en página 7).

Association, A. L. *et al.* (2000). Information literacy competency standards for higher education. (citado en páginas 2, 3).

Baepler, P. & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, *4*(2), 17. (citado en página 5).

Bichsel, J. (2012). *Analytics in higher education: Benefits, barriers, progress, and recommendations*. EDUCAUSE Center for Applied Research. (citado en página 5).

Borkar, S. & Rajeswari, K. (2013). Predicting student's academic performance using education data mining. *International Journal of Computer Science and Mobile Computing (IJCSMC)*, *2*(7), 273-279. (citado en página 6).

Borkar, S. & Rajeswari, K. (2014). Attributes selection for predicting student's academic performance using education data mining and artificial neural network. *International Journal of Computer Applications*, *86*(10). (citado en página 6).

Chen, S. Y. & Liu, X. (2008). An integrated approach for modeling learning patterns of students in web-based instruction: A cognitive style perspective. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *15*(1), 1. (citado en página 5).

Council, N. R. *et al.* (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. National Academies Press. (citado en página 4).

Dutt, A., Ismail, M. A. & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*. (citado en página 7).

Eisenberg, M. B. & Berkowitz, R. E. (1990). *Information problem solving: The big six skills approach to library & information skills instruction.* ERIC. (citado en página 4).

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37. (citado en página 11).

González-Ibáñez, R. [R.], Gacitua, D., Sormunen, E. & Kiili, C. (2017). NEURONE: oNlinE inqUiRy experimentatiON systEm. En T. be included in Proceedings of the 80th Annual Meeting of the Association for Information Science & T. (2017) (Eds.). (citado en página 8).

Head, A. J. (2013). Project Information Literacy: What can be learned about the information-seeking behavior of today's college students? (citado en página 2).

Henrie, C. R., Halverson, L. R. & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, *90*, 36-53. (citado en página 5).

Jayakameswaraiah, M. & Ramakrishna, S. (2014). A study on prediction performance of some data mining algorithms. *International Journal*, *2*(10). (citado en página 6).

Lahtinen, E., Ala-Mutka, K. & Järvinen, H.-M. (2005). A study of the difficulties of novice programmers. En *ACM Sigcse Bulletin* (Vol. 37, *3*, pp. 14-18). ACM. (citado en página 6).

Luan, J. (2002). Data mining and its applications in higher education. *New directions for institutional research*, *2002*(113), 17-36. (citado en página 4).

Martin, J. (1991). *Rapid application development*. Macmillan Publishing Co., Inc. (citado en página 11).

Marzal, M. Á. & Saurina, E. (2015). Diagnóstico del estado de la alfabetización en información (ALFIN) en las universidades chilenas. *Perspectivas em Ciência da Informação*, *20*(2), 58-78. (citado en página 2).

McConnell, S. (1996). *Rapid development: Taming wild software schedules*. Pearson Education. (citado en página 12).

Merceron, A. & Yacef, K. (2005). Educational data mining: A case study. En *AIED* (pp. 467-474). (citado en página 5).

Mining, T. E. D. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. En *Proceedings of conference on advanced technology for education*. (citado en página 4).

Moreno-Clari, P., Arevalillo-Herraez, M. & Cerveron-Lleo, V. (2009). Data analysis as a tool for optimizing learning management systems. En *Advanced Learning Technologies, 2009. ICALT 2009. Ninth IEEE International Conference on* (pp. 242-246). IEEE. (citado en página 6).

Morris, L. V., Wu, S.-S. & Finnegan, C. L. (2005). Predicting retention in online general education courses. *The American Journal of Distance Education*, *19*(1), 23-36. (citado en página 5).

Oskouei, R. J. & Askari, M. (2014). Predicting academic performance with applying data mining techniques (generalizing the results of two different case studies). *Computer Engineering and Applications Journal*, *3*(2), 79-88. (citado en página 7).

Quintana, C., Zhang, M. & Krajcik, J. (2005). A framework for supporting metacognitive aspects of online inquiry through software-based scaffolding. *Educational Psychologist*, *40*(4), 235-244. (citado en página 4).

Shah, C., Hendahewa, C. & González-Ibáñez, R. [Roberto]. (2016). Rain or shine? Forecasting search process performance in exploratory search tasks. *Journal of the Association for Information Science and Technology*, *67*(7), 1607-1623. (citado en página 5).

Shahiri, A. M., Husain, W. *et al.* (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422. (citado en página 7).

Shazmeen, S. F., Baig, M. M. A. & Pawar, M. R. (2013). Performance evaluation of different data mining classification algorithm and predictive analysis. *IOSR Journal of Computer Engineering*, *10*(6), 01-06. (citado en página 6).

Sormunen, E., González-Ibáñez, R., Kiili, C., Leppänen, P., Mikkilä-Erdmann, M., Erdmann, N. & Escobar-Macaya, M. (2017). A Performance-based Test for Assessing Students' Online Inquiry Competences in Schools. En E. C. in Information Literacy (ECIL) (Ed.). (citado en página 3).

Sukhija, K., Jindal, M. & Aggarwal, N. (2015). The recent state of educational data mining: A survey and future visions. En *MOOCs, Innovation and Technology in Education (MITE), 2015 IEEE 3rd International Conference on* (pp. 354-359). IEEE. (citado en página 7).

Urra, M. C. V. & Castro, S. O. (2016). Alfabetización en información: Estudio de su impacto en estudiantes de último año del pregrado de las facultades de educación y ciencias naturales y exactas en la Universidad de Playa Ancha de Ciencias de la Educación, 20-40. (citado en página 2).

Weiner, S. A. (2014). Who teaches information literacy competencies? Report of a study of faculty. *College Teaching*, *62*(1), 5-12. (citado en página 2).