

**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**

**Gonzalo Javier Martinez Ramirez**

Profesor guía: Roberto Ignacio González Ibáñez

Tesis de grado presentado en conformidad a  
los requisitos para obtener el grado de Magíster  
en Ingeniería Informática.

Santiago – Chile

2017

## RESUMEN

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 (primaria, secundaria 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 comportamiento de búsqueda de información en línea en estudiantes de educación básica el cual se vaya perfeccionando a través del registro de datos históricos y que de un feedback en tiempo real.

La investigación será guiada por la metodología KDD con el fin de descubrir patrones en los datos que permitan la creación de un modelo de predicción del comportamiento de búsqueda. Además, para apoyar el proceso de investigación, se desarrollará una plataforma que funcione como extensión de la plataforma NEURONE (*oNlinE inqUiry expeRimentatiON systEm*). La plataforma propuesta alimentará y perfeccionará el modelo de predicción y entregará predicciones en tiempo real. Esta plataforma se guiará bajo la metodología RAD (*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 de investigación en línea, comportamiento de estudiantes, minería de datos, modelos de clasificación.

## ABSTRACT

Today

**Keywords:**

# TABLA DE CONTENIDOS

<b>Capítulo 1. Introducción</b>	<b>1</b>
1.1 Antecedentes y motivación . . . . .	1
1.2 Descripción del problema . . . . .	2
1.3 Solución propuesta . . . . .	3
1.3.1 Características de la solución . . . . .	3
1.3.2 Propósito de la solución . . . . .	4
1.4 Objetivos y alcances de la solución . . . . .	4
1.4.1 Objetivo general . . . . .	4
1.4.2 Objetivos específicos . . . . .	4
1.4.3 Alcances . . . . .	5
1.5 Metodología y herramientas utilizadas . . . . .	5
1.5.1 Metodología a usar . . . . .	5
1.5.2 Herramientas de desarrollo . . . . .	7
1.6 Organización del documento . . . . .	8
<b>Capítulo 2. Marco teórico</b>	<b>9</b>
2.1 Marco conceptual . . . . .	9
2.1.1 Recuperación de Información Humano Computador . . . . .	9
2.1.2 Rendimiento . . . . .	10
2.1.3 Alfabetización informacional . . . . .	11
2.1.4 Competencias de investigación en línea . . . . .	11
2.1.5 Aprendizaje automático . . . . .	11
2.1.6 Técnicas de aprendizaje automático supervisado . . . . .	12
2.1.7 Técnicas de reforzamiento . . . . .	15
2.1.8 Minería de datos educacional . . . . .	16
2.2 Estado del arte . . . . .	16
2.2.1 Alfabetización informacional . . . . .	16
2.2.2 Competencias de investigación en línea . . . . .	17
2.2.3 Comportamiento de búsqueda de información de estudiantes . . . . .	17
2.2.4 Usos de la minería de datos educacional . . . . .	18
2.2.5 Técnicas utilizadas en la minería de datos educacional . . . . .	20
2.3 Marco de investigación . . . . .	22
2.4 Resumen . . . . .	22

## Tabla de Contenidos

---

<b>Referencias bibliográficas</b>	<b>28</b>
<b>Apéndice A. Capítulo Apéndice</b>	<b>29</b>
A.1 Sección del apéndice . . . . .	29
A.1.1 Subseccion del apéndice . . . . .	30
<b>Apéndice B. Another Appendix Chapter</b>	<b>31</b>

## ÍNDICE DE TABLAS

2.1. Clasificación del estado del arte . . . . .	22
B.1. Ejemplo de una tabla . . . . .	31

## ÍNDICE DE FIGURAS

1.1. Ciclo de construcción y perfeccionamiento del modelo . . . . .	3
1.2. Proceso de búsqueda de información de un estudiante . . . . .	4
2.1. Árbol de decisión . . . . .	13
2.2. SVM . . . . .	14
2.3. Perceptrón . . . . .	14
2.4. Perceptrón multicapa . . . . .	15
2.5. Aprendizaje por reforzamiento . . . . .	15
A.1. A scientific diagram using the pgfplots package by Christian Feuersaenger using the same colors which are also used for the layout . . . . .	29
B.1. Flowchart of fundamental disease transmission mechanisms . . . . .	31

# CAPÍTULO 1. INTRODUCCIÓN

A través del presente capítulo se introduce el problema a resolver a lo largo del trabajo. En primer lugar, se presentan los antecedentes que motivan su realización. A continuación, se describen las características del problema, la propuesta de solución, sus alcances y objetivos. Finalmente, se describe la metodología utilizada y la organización del presente documento.

## 1.1. ANTECEDENTES Y 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”<sup>1</sup> (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 Urrea 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. 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 (Urrea & Castro, 2016).

---

<sup>1</sup> Traducción libre.



A través de encuestas a estudiantes universitarios, Head (2013, p. 475) establece que al momento de realizar investigaciones el 84 % de los estudiantes universitarios utiliza como fuente primaria de búsqueda Wikipedia<sup>2</sup> y un 87 % consulta a sus amigos, sin verificar la veracidad de la información 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 ver 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*”<sup>3</sup> (iFuCo, desde ahora en adelante), el cual pretende abordar la temática de la alfabetización informacional en estudiantes de enseñanza básica con el objetivo de estudiar sus patrones de comportamiento y ofrecer modelos curriculares adecuados respecto al tema (Sormunen *et al.*, 2017).

## 1.2. DESCRIPCIÓN 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<sup>4</sup> 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 diversos factores 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?

<sup>2</sup><https://es.wikipedia.org/>

<sup>3</sup><https://www.researchgate.net/project/Enhancing-learning-and-teaching-for-future-competences-of-online-inquiry-in-multiple-domains-iFuCo>

<sup>4</sup>En otros países es conocido como enseñanza primaria.

**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?

### 1.3. SOLUCIÓN PROPUESTA

#### 1.3.1. Características de la solución

La solución consiste en incorporar un módulo en NEURONE (R. 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<sup>5</sup> (*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<sup>6</sup>, 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 1.1, 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 del desempeño de búsqueda y prediciendo de forma continua a partir del comportamiento actual de búsqueda de información del estudiante.

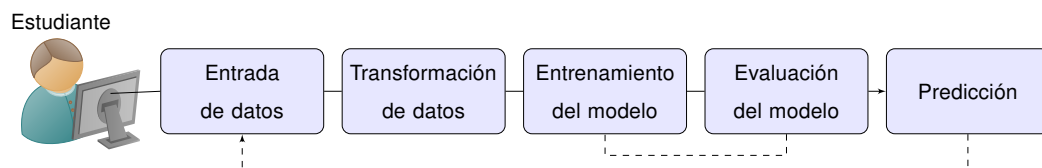


Figura 1.1: Ciclo de construcción y perfeccionamiento del modelo

Fuente: Elaboración propia, (2017)

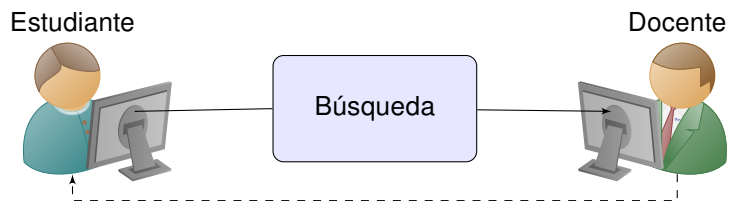
<sup>5</sup>Traducción libre.

<sup>6</sup><https://spark.apache.org/>

### 1.3.2. 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 1.2, 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 1.2: Proceso de búsqueda de información de un estudiante*

Fuente: Elaboración propia, (2017)

## 1.4. OBJETIVOS Y ALCANCES DE LA SOLUCIÓN

### 1.4.1. Objetivo general

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

### 1.4.2. 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 línea de estudiantes de educación básica.
6. Implementar los modelos predictivos en la plataforma NEURONE.

#### 1.4.3. Alcances

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.

### 1.5. METODOLOGÍA Y HERRAMIENTAS UTILIZADAS

#### 1.5.1. 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.

### *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.

### *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).

### 1.5.2. Herramientas de desarrollo

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

#### *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<sup>7</sup>.

#### *Software*

En cuanto herramientas *software*, el desarrollo se llevará a cabo en la distribución GNU/Linux Debian<sup>8</sup> en su versión 9.0. El modelo se llevará a cabo en Spark ML<sup>9</sup>. Para el análisis estadístico se hará uso de R. Además, cada módulo desarrollado estará contenido en contenedores de Docker

---

<sup>7</sup><https://www.digitalocean.com/>

<sup>8</sup><https://www.debian.org/>

<sup>9</sup><https://spark.apache.org/>

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  $\text{\LaTeX}$  para el documento escrito.

- Anaconda Python 3.5
- Apache Kafka 0.10.2.0
- Apache Spark 2.1.0
- Flask 0.12.2 <sup>10</sup>
- MongoDB
- Python 3.5

## 1.6. ORGANIZACIÓN DEL DOCUMENTO

El resto del documento se estructura de la siguiente forma.

**Capítulo 2** Se estipulan los conceptos teóricos que se deben definir para tener una base consensuada respecto de los distintos conceptos que se tratan en este documento. En el mismo capítulo se aborda el estado del arte donde se hace una revisión bibliográfica de los últimos avances en el área.

---

<sup>10</sup><http://flask.pocoo.org/>

## CAPÍTULO 2. MARCO TEÓRICO

A través del presente capítulo se entregan las bases teóricas, conceptuales y empíricas que soportan cada desarrollo de esta investigación. En primer lugar, se presenta el marco conceptual donde se entregan las definiciones y conceptos necesarios para abordar esta investigación. En segundo lugar, se presenta el estado del arte relacionado con el tema.

### 2.1. MARCO CONCEPTUAL

En esta sección se presentan conceptos y bases teóricas respecto a la temática que conduce el desarrollo de este trabajo, el cual tiene relación con el uso de interfaces no tradicionales, específicamente con una interfaz operada con el cuerpo. Además, se indaga sobre ciertas definiciones para establecer lo que se pretende medir en este estudio, lo que involucra la experiencia de usuario y métricas de rendimiento en la realización de tareas. Finalmente, se proponen ciertas características fundamentales respecto de este tipo de proyectos relacionados con diseños experimentales con usuarios.

#### 2.1.1. Recuperación de Información Humano Computador

La Recuperación de Información Humano Computador<sup>1</sup> (HCIR, por sus iniciales en inglés de *Human Computer Information Retrieval*) es el estudio de los métodos que integran la inteligencia humana y la búsqueda algorítmica para ayudar a la gente a mejorar la búsqueda, exploración y aprendizaje de información (Marchionini, 2006). Dentro de esta área interactúan otras disciplinas, como la recuperación de información, llamada en inglés *Information Retrieval* (IR), la que está enfocada principalmente en proveer a los usuarios de fácil acceso a la información de su interés trabajando con la representación, almacenamiento, organización y acceso a objetos de información como documentos, páginas *web*, catálogos en línea y objetos multimedia (Ricardo & Berthier, 2011) y la búsqueda de información, llamada en inglés *Information Seeking* (IS) que se entiende como un proceso más orientado al usuario y abierto que IR. En IS, no se sabe si existe una respuesta a la consulta del usuario, por lo que el proceso de búsqueda puede proporcionar el aprendizaje necesario para satisfacer su necesidad de información (Ricardo & Berthier, 2011).

La búsqueda de información es un campo de la investigación que relaciona el desarrollo del área de las tecnologías y ciencias de la computación con la psicología y ciencias sociales en el procesamiento de la información, en donde el usuario toma un papel activo por medio de interacciones

---

<sup>1</sup>Traducción libre.



explicitas e implícitas con la información (Carroll, 1997). Es una disciplina que contempla tanto al sistema como al usuario, así como la relación que se establece a través del comportamiento del usuario y sus experiencias (Kelly *et al.*, 2009). Se ubica en un punto medio entre las disciplinas de HCI e IR.

### 2.1.2. Rendimiento

En el contexto de la recuperación de información se definen *precision* y *recall* en función de un conjunto de documentos relevantes y un conjunto de documentos recuperados (Powers, 2011, p. 38), las cuales se definen a continuación.

**Precision** Métrica que mide la razón de documentos relevantes recuperados con respecto al total de documentos recuperados. Representado en la Ecuación (2.1), el resultado de esta métrica es un valor continuo entre 0 y 1, mientras más cercano a 1, mayor fue su precisión al encontrar los documentos relevantes.

$$Precision = \frac{[\{\text{documentos relevantes}\} \cap \{\text{documentos recuperados}\}]}{\{\text{documentos recuperados}\}} \quad (2.1)$$

**Recall** Métrica que mide la razón de documentos relevantes recuperados con respecto al total de documentos relevantes. Representado en la Ecuación (2.2), el resultado de esta métrica es un valor continuo entre 0 y 1, mientras más cercano a 1, mayor fue la recuperación de documentos en base al total del universo disponible.

$$Recall = \frac{[\{\text{documentos relevantes}\} \cap \{\text{documentos recuperados}\}]}{\{\text{documentos relevantes}\}} \quad (2.2)$$

**F1** Métrica que considera los valores de *precision* y *recall* en un promedio ponderado. Representado en la Ecuación (2.3), el resultado de esta métrica es un valor continuo entre 0 y 1, en que un valor cercano a uno permite identificar a los estudiantes con una recuperación de documentos proporcional a su precisión, respecto a la relevancia de estos.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (2.3)$$

### 2.1.3. Alfabetización informacional

La alfabetización informacional (conocida en inglés *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”<sup>2</sup> (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.

### 2.1.4. 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”<sup>3</sup> (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).

### 2.1.5. Aprendizaje automático

El aprendizaje automático (*Machine Learning*, por su nombre en inglés) es un área de la Inteligencia Artificial enfocada al desarrollo de algoritmos capaces de generalizar comportamientos, a partir de información no estructurada suministrada en forma de ejemplos, de manera que posean la capacidad de adaptarse en base a experiencia adquirida y no deban ser reprogramados. Los diversos algoritmos de esta rama se diferencian en su forma de llevar a cabo el aprendizaje, algunas de estas son:

---

<sup>2</sup>Traducción libre.

<sup>3</sup>Traducción libre.

**Aprendizaje supervisado** Se realiza mediante un entrenamiento controlado por un agente externo (supervisor, maestro), que determina la respuesta que se debería generar a partir de una entrada determinada. El supervisor controla la salida y en caso de que esta no coincida con la deseada. Se modifican los parámetros usados, con el fin de conseguir que la salida obtenida se aproxime a la deseada.

**Aprendizaje no supervisado** Se realiza mediante un entrenamiento sin conocimiento a priori de la salida deseada, es decir, solo son conocidas las entradas, por lo tanto el aprendizaje se basa en el grado de familiaridad o similitud entre la información que presenta una entrada y la información recolectada por entradas anteriores.

Uno de los problemas que intenta solucionar la rama de aprendizaje automático supervisado, es la clasificación, cuya definición en este contexto corresponde al problema de identificar a qué categoría pertenece una nueva observación.

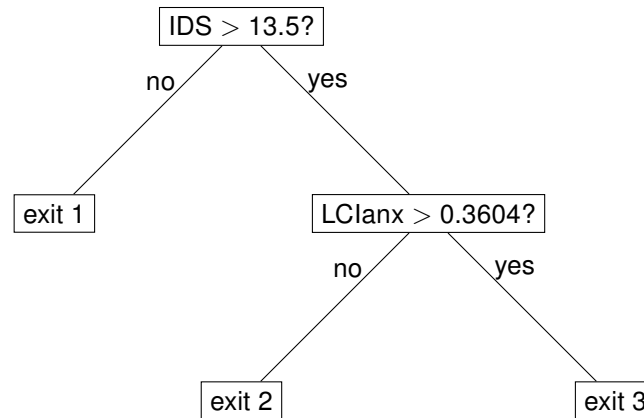
Un clasificador se define como cualquier algoritmo que resuelva el problema de clasificación. Para ello, este tipo de algoritmo necesita ser entrenado con un conjunto de observaciones que ya estén etiquetadas en una categoría, de forma que pueda corregir su aprendizaje. Las observaciones son representadas a través de un conjunto de propiedades que permiten determinar a qué categoría pertenece. Por su parte, las propiedades pueden ser de tipo categórico (por ejemplo el tipo de sangre), ordinal (“grande”, “medio”, “pequeño”), valores enteros o reales o incluso utilizando la diferencia y similitud entre la observación actual y las observaciones previas (Vilches, 2015). Como consecuencia del entrenamiento, se consigue un modelo del clasificador capaz de asignar observaciones desconocidas a una categoría conocida específica, solo sabiendo sus propiedades.

La terminología usada en esta área, suele llamar a las observaciones “instancias”, “ejemplos” o “sujetos”, a las propiedades “características” o “atributos”, a las categorías “clases”, al conjunto de observaciones “conjunto de entrada” y al conjunto de observaciones usadas en el entrenamiento “conjunto de entrenamiento”.

Debido a las características de este proyecto, se utilizan técnicas de aprendizaje automático supervisado.

#### **2.1.6. Técnicas de aprendizaje automático supervisado**

A continuación, se describen los algoritmos y datos seleccionados para esta investigación.

*Árboles de decisión**Figura 2.1: Árbol de decisión*

Fuente: Elaboración propia, (2017)

*Máquinas de vectores de soporte (SVM)*

Las máquinas de vectores de soporte o SVM (del inglés *Support Vector Machine*) son un conjunto de algoritmos de aprendizaje supervisado que se enfocan en resolver problemas de clasificación, regresión y agrupamiento. La SVM destinada para clasificación, es un clasificador lineal binario que busca encontrar un hiperplano que separe de forma óptima un conjunto de datos, maximizando la distancia entre las dos clases.

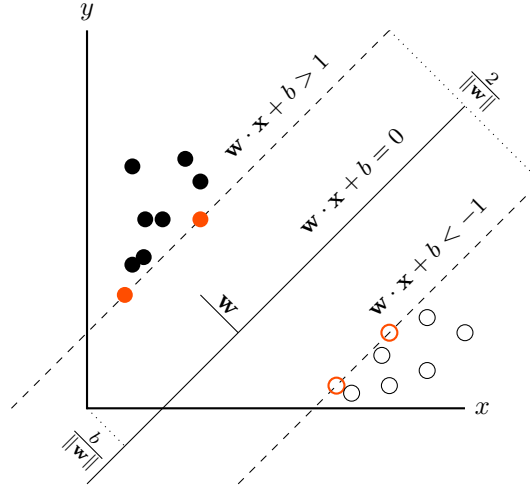


Figura 2.2: SVM

Fuente: Elaboración propia, (2017)

### Perceptrón

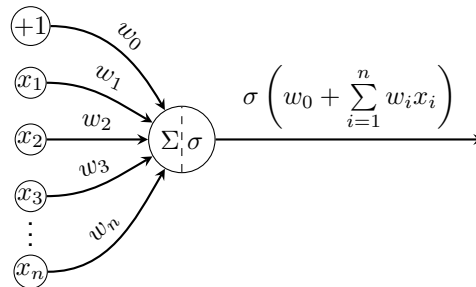


Figura 2.3: Perceptrón

Fuente: Elaboración propia, (2017)

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix} \quad (2.4)$$

### Perceptrón multicapa

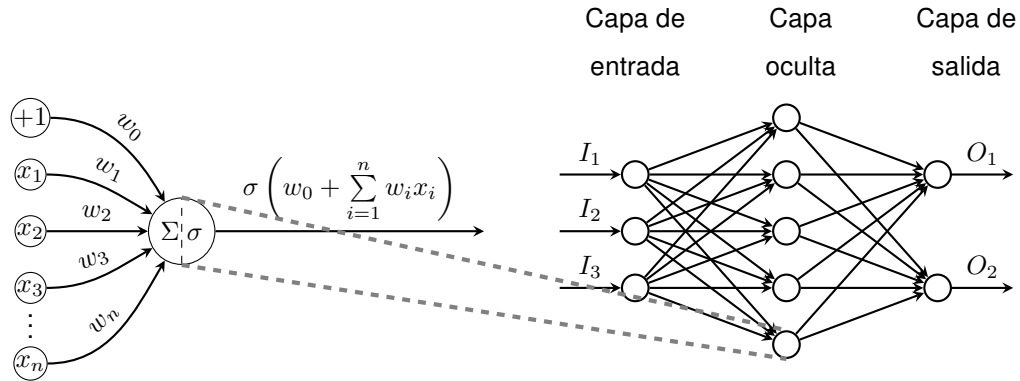


Figura 2.4: Perceptrón multicapa

Fuente: Elaboración propia, (2017)

### Naïve Bayes

El clasificador Naïve Bayes, también conocido como clasificador bayesiano, es un clasificador probabilístico basado en el teorema de Bayes descrito en la eq:naive-bayes, donde  $P(A|B)$  es la probabilidad de la hipótesis

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \quad (2.5)$$

#### 2.1.7. Técnicas de reforzamiento

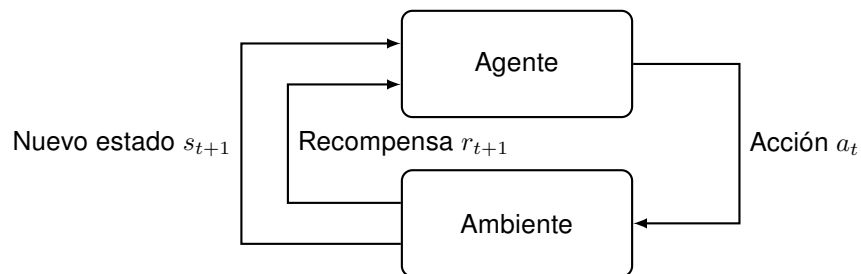


Figura 2.5: Aprendizaje por reforzamiento

Fuente: Elaboración propia, (2017)

### 2.1.8. 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”<sup>4</sup>. 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.

## 2.2. ESTADO DEL ARTE

El propósito de esta sección es presentar los últimos trabajos realizados en la línea de investigación que se ha planteado en la sección y capítulo anterior. La búsqueda de estos trabajos considera: la exploración de trabajos con estudios que midan experiencia de usuario y/o medidas de rendimiento en la utilización de interfaces no-tradicionales operadas con el cuerpo para la realización de actividades en distintos contextos.

### 2.2.1. Alfabetización informacional

La alfabetización informacional (conocida en inglés *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”<sup>5</sup> (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.

A nivel nacional, la enseñanza de estas competencias se realiza principalmente por parte de las bibliotecas universitarias, por lo cual, Marzal y Saurina (2015) realizan un análisis de la situación

---

<sup>4</sup>Traducción libre.

<sup>5</sup>Traducción libre.

actual de los cursos y políticas de alfabetización informacional de las bibliotecas pertenecientes al Consejo Nacional de Educación. Se concluye que no existe uniformidad en las definiciones y marcos teóricos utilizados. Por otro lado, los estudiantes universitarios chilenos presentan dificultades con las competencias informacionales, tales como aprenderlas, pero no utilizarlas, hecho que se vio reflejado por el estudio interno realizado por la Universidad de Playa Ancha (Urrea & Castro, 2016) sobre el impacto de la enseñanza de la alfabetización informacional en dicha universidad. Una de las posibles causas de por qué los estudiantes tienen dificultades es el hecho de que en los colegios no se prioriza la generación de conocimiento, sino la reiteración de la información.

### 2.2.2. 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”<sup>6</sup> (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 *et al.*, 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).

### 2.2.3. Comportamiento de búsqueda de información de estudiantes

Con la incorporación de herramientas digitales en la enseñanza escolar es necesario evaluar su aporte en el aprendizaje. En general, en la evaluación se analizan las mediciones respecto a los datos proporcionados por los estudiantes, ya sea de forma directa, como cuestionarios o indirecta, como datos generados al utilizar un sistema computacional pero con previa autorización del estudiante.

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,

---

<sup>6</sup>Traducción libre.



tiempos de permanencia o de respuestas y uso de sitios *web*, entre otros. Tomando tiempos de permanencia y el uso de sitios, Shah, Hendaheya 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.

Siguiendo este tipo de análisis sobre datos generados por estudiantes, en la prueba internacional PISA (OECD, 2015) se analizaron los registros de la navegación en un sitio *web* ficticio, con múltiples hipervínculos internos. El sitio estaba dividido en diferentes páginas unidas por *links*, clasificados en tres niveles, siendo las de nivel 1 las más fáciles de acceder (variados hipervínculos conducían a éstas), en cambio, la ruta para llegar a las de nivel 3 era específica. Para contestar las preguntas de comprensión lectora era necesario encontrar la página con la información necesaria, la cual, dependiendo de su nivel (1 a 3) podía transformar la tarea de una navegación simple a una compleja. El objetivo de el análisis fue determinar qué tan fluida es la lectura en línea, los diferentes desempeños de los estudiantes y la persistencia al tratar de responder.

#### **2.2.4. 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 Alkhozai (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.

*Predicción del desempeño estudiantil*

La predicción está enfocada en pronosticar el desempeño estimando valores desconocidos de variables que caracterizan al estudiante. Estos valores normalmente corresponden al rendimiento, conocimiento o calificaciones. También se utiliza para detectar estilos de aprendizaje, predecir si contestará correctamente una pregunta, modelar los cambios en el conocimiento al adquirir un nuevo aprendizaje y determinar variables del aprendizaje no observables en la conducta en línea de los estudiantes (Romero Ventura, 2010).

En general el enfoque es detectar de forma temprana quienes probablemente tendrán un bajo desempeño o encontrar los indicadores que permitan identificar a estos alumnos de tal forma que el profesor o la entidad educacional correspondiente pueda intervenir a tiempo y prevenir por ejemplo la deserción escolar.

Por ejemplo, el trabajo de Antunes (2010) se centró en anticipar el fracaso de los estudiantes lo más pronto posible en cursos de fundamentos de la programación. Se utilizó la información de los estudiantes universitarios (12 atributos observables y 4 de calificaciones) que han cursado dicha asignatura a lo largo de cinco años. Con el algoritmo J48 se construyeron tres clasificadores bajo distintos enfoques para determinar reglas de decisión que permitan predecir el fracaso. Por otro lado, el estudio realizado por Kumar (2016) identificó a los estudiantes con mayor probabilidad de aprobar el examen final de una carrera de ingeniería. En base a atributos numéricos como las notas de secundaria y de semestres previos al examen, en conjunto a variables demográficas (p. ej. género, estado civil, ocupaciones de los padres, entre otras) se realizó la predicción del desempeño de los estudiantes, siendo los algoritmos J48 y REPTree los con mejor desempeño.

A nivel escolar, Márquez-Vera, Cano, Romero y Ventura (2013) buscaron predecir el bajo rendimiento escolar y asociarlo con las causas de la deserción escolar o repitencia. Realizaron una investigación con estudiantes mexicanos de 15 y 16 años, con el objetivo de determinar los factores más influyentes y las razones que llevan al fracaso, e identificar a los estudiantes que muestren estas características para ofrecerles la ayuda correspondiente. Se consideraron atributos socio-económicos, personales, sociales, familiares, escolares y calificaciones, haciendo uso de diez algoritmos de clasificación tradicionales, cinco para la construcción de árboles de decisión (J48, RandomTree, Reptree, SimpleCart y ADtree) y cinco para la generación de reglas asociación (JRIP, NNge, OneR, Prism y Ripdor). Además, crearon un algoritmo utilizando programación genética gramatical para predecir si un estudiante aprueba o reprueba. Los resultados muestran que con

este algoritmo evolucionado se obtienen resultados tan precisos como con los algoritmos tradicionales, pero con menos reglas y condiciones por regla, permitiendo así reducir el número de características, y utilizarlas para detectar a los estudiantes con mayor tendencia a repetir el año escolar.

En Shamsi Lakshmi (2016) se realiza un análisis comparativo de distintas técnicas de clasificación (Naïve Bayes, LibSVM, J48, Random Forest y JRIP) aplicado a determinar los factores claves que afectan el rendimiento en estudiantes de ingeniería en la India, bajo la premisa de que una inapropiada educación primaria y secundaria repercute en desempeño de la educación superior de estos estudiantes. Se consideraron 34 parámetros, en su mayoría variables descriptivas asociadas al entorno familiar. Los resultados obtenidos mostraron que Naïve Bayes es el más exacto para predecir un mal desempeño y JRIP el más preciso en la predicción de las calificaciones, junto con entregar los factores que afectan el rendimiento en un set de reglas.

### 2.2.5. 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<sup>7</sup> haciendo uso de las técnicas de maximización, 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.

---

<sup>7</sup><https://moodle.org/>

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.

Oskouei y Askari (2014) identifica los factores que afectan el rendimiento de los estudiantes en diferentes países, y aplica 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.

Roberto González-Ibáñez, Shah y White (2012) utiliza

Shah *et al.* (2016) utiliza

Roberto González-Ibáñez, Shah y White (2015) utiliza

Hassan, Shi, Craswell y Ramsey (2013) utiliza

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).

*Tabla 2.1: Clasificación del estado del arte*

Técnicas/Prototipo	Finalidad	P2	P3	P4
Árboles de decisión	✓	✓	✓	✓
Chapter 2: Background and experimental set-up		✓	✓	✓
Chapter 3: Evaluating and comparing outlier-selection algorithms		✓		
Chapter 4: Stochastic Outlier Selection			✓	
Chapter 5: Meta-features for one-class data sets				✓
Chapter 6: Meta-learning for one-class classifiers				✓
Chapter 7: Conclusions	✓	✓	✓	✓

Fuente: Elaboración propia, (2017)

## 2.3. MARCO DE INVESTIGACIÓN

En esta sección se propone y formula el marco de investigación que guía este estudio. Para esto se desarrollan las preguntas de investigación que motivan la realización del estudio. Además se presentan las hipótesis que se desean someter a prueba en base a los resultados producto de la realización del estudio que en este documento se propone.

**RQ 1** ¿De qué manera se puede estimar durante el proceso de aprendizaje de competencias informacionales la influencia de diversos factores 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?

## 2.4. RESUMEN

En el presente capítulo

Como marco conceptual se definen conceptos importantes que son utilizados en el estudio. Primero la experiencia de usuario dado que implica emociones y actitudes de una persona con respecto al producto o servicio que se esté utilizando suele medirse con instrumentos que requieren una participación directa del usuario como encuestas. Con respecto al rendimiento en el área de recuperación de información se utilizan métricas para evaluarlo a partir fórmulas que usan la cantidad de documentos clasificados y documentos correctamente clasificados.

Las competencias informacionales son un conjunto de habilidades asociadas al descubrimiento reflexivo de la información. Resulta fundamental comprender cómo la información se produce, evalúa, utiliza y comparte. Existen habilidades asociadas puntualmente a tareas de investigación, basadas en la indagación, desarrolladas en Internet (*online inquiry*) para encontrar, evaluar críticamente, sintetizar y comunicar la información de una manera correcta.

En el estado del arte se realiza una revisión de distintos estudios de usuario en los que se trabaja con representaciones alternativas de documentos, donde se observaron distintos resultados. Por ejemplo, Nguyen y Zhang (2006) encontraron que, al utilizar una representación visual de documentos, la clasificación de documentos relevantes mejoraba o Tilsner (2009) quien observó un rendimiento mayor con menores tiempos de respuesta utilizando una interfaz visual. Por otro lado, Sebrechts, Cugini, y Laskowski (1999) obtuvieron un rendimiento similar contrastando una interfaz visual con una tradicional. Por otro lado, se ve que en el mercado ya existen algunas herramientas de visualización como Zakta o oSkope las que pueden resultar beneficiosas si se usan para buscar ciertos tipos de información.

Finalmente, en el marco de investigación se plantean preguntas de investigación relacionadas con cómo las personas perciben su interacción con resultados de búsqueda de información a través de representaciones visuales y si es posible mejorar aspectos como la experiencia de usuario el rendimiento utilizando este tipo de representación. Lo anterior lleva a las dos hipótesis con las que se trabaja en este proyecto que de forma resumida son:

## REFERENCIAS BIBLIOGRÁFICAS

- Abdullah, A., Malibari, A. & Alkhozai, 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 18).
- Akinola, O., Akinkunmi, B. & Alo, T. (2012). A data mining model for predicting computer programming proficiency of computer science undergraduate students. (citado en página 20).
- 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 22).
- Association, A. L. *et al.* (2000). Information literacy competency standards for higher education. (citado en páginas 1, 11, 16).
- 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 18).
- Bichsel, J. (2012). *Analytics in higher education: Benefits, barriers, progress, and recommendations*. EDUCAUSE Center for Applied Research. (citado en página 18).
- 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 21).
- 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 21).
- Carroll, J. M. (1997). Human-computer interaction: psychology as a science of design. *Annual review of psychology*, 48(1), 61-83. (citado en página 10).
- 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 20).

- Council, N. R. *et al.* (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. National Academies Press. (citado en páginas 11, 17).
- Dutt, A., Ismail, M. A. & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*. (citado en página 22).
- Eisenberg, M. B. & Berkowitz, R. E. (1990). *Information problem solving: The big six skills approach to library & information skills instruction*. ERIC. (citado en páginas 11, 17).
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37. (citado en página 6).
- 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 3).
- González-Ibáñez, R. [Roberto], Shah, C. & White, R. W. (2012). Pseudo-collaboration as a method to perform selective algorithmic mediation in collaborative IR systems. *Proceedings of the American Society for Information Science and Technology*, 49(1), 1-4. (citado en página 21).
- González-Ibáñez, R. [Roberto], Shah, C. & White, R. W. (2015). Capturing Collabportunities: A method to evaluate collaboration opportunities in information search using pseudocollaboration. *Journal of the Association for Information Science and Technology*, 66(9), 1897-1912. (citado en página 21).
- Hassan, A., Shi, X., Craswell, N. & Ramsey, B. (2013). Beyond clicks: query reformulation as a predictor of search satisfaction. En *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 2019-2028). ACM. (citado en página 21).
- 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 17).



- Jayakameswaraiah, M. & Ramakrishna, S. (2014). A study on prediction performance of some data mining algorithms. *International Journal*, 2(10). (citado en página 21).
- Kelly, D. *et al.* (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2), 1-224. (citado en página 10).
- 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 20).
- Luan, J. (2002). Data mining and its applications in higher education. *New directions for institutional research*, 2002(113), 17-36. (citado en página 16).
- Marchionini, G. (2006). Toward human-computer information retrieval. *Bulletin of the American Society for Information Science and Technology*, 32(5), 20-22. (citado en página 9).
- Martin, J. (1991). *Rapid application development*. Macmillan Publishing Co., Inc. (citado en página 6).
- 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 páginas 1, 16).
- McConnell, S. (1996). *Rapid development: Taming wild software schedules*. Pearson Education. (citado en página 7).
- Merceron, A. & Yacef, K. (2005). Educational data mining: A case study. En *AIED* (pp. 467-474). (citado en página 18).
- 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 16).
- 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 20).
- 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 18).

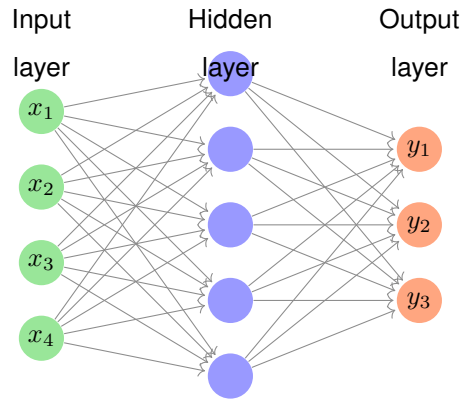
- OECD. (2015). Using Log-File Data to Understand What Drives Performance in PISA (Case Study). doi:<http://dx.doi.org/10.1787/9789264239555-10-en>. (citado en página 18)
- Oskouei, R. J. & Askari, M. (2014). Predicting academic performance with applying data mining techniques (generalizing the results of two different case studies). *Computer Engineering and Applications Journal*, 3(2), 79-88. (citado en página 21).
- Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (citado en página 10).
- 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 páginas 11, 17).
- Ricardo, B. & Berthier, R. (2011). Modern Information Retrieval: the concepts and technology behind search second edition. *Addision Wesley*, 84(2). (citado en página 9).
- Shah, C., Hendaheewa, 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 páginas 18, 21).
- 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 22).
- 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 21).
- 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 2).
- 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 22).

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 páginas 1, 17).

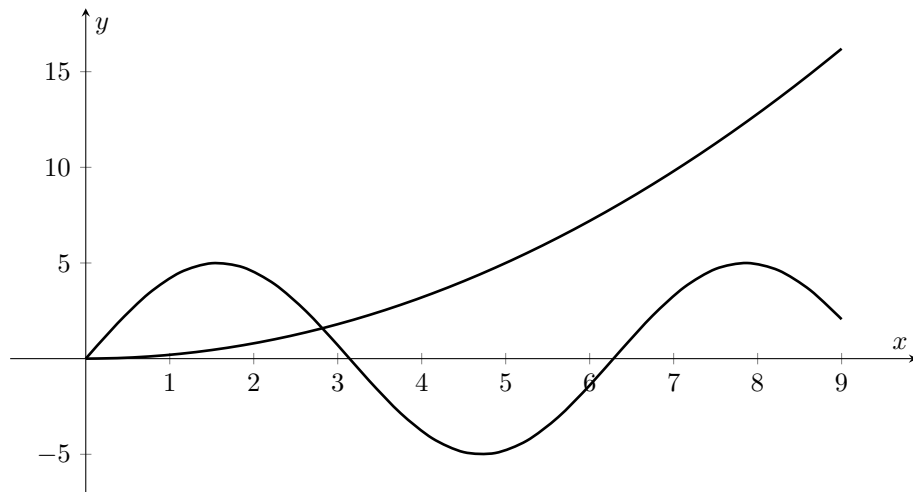
Weiner, S. A. (2014). Who teaches information literacy competencies? Report of a study of faculty. *College Teaching*, 62(1), 5-12. (citado en página 1).

## APÉNDICE A. CAPÍTULO APÉNDICE

### A.1. SECCIÓN DEL APÉNDICE



**Example Diagram with a Line Break in the Title  
(using the `text width` option in the title style)**



*Figura A.1: A scientific diagram using the `pgfplots` package by Christian Feuersaenger using the same colors which are also used for the layout*

Fuente: Elaboración propia, (2017)

**A.1.1. Subseccion del apéndice**

Como se ve en el Algoritmo A.1

---

**Algoritmo A.1:** Escribiendo algoritmos usando  $\text{\LaTeX}2\epsilon$ 

---

**Entrada:** Esto es la entrada del algoritmo**Salida:** Esto es la salida del algoritmo

```
1 inicio
2    $V \leftarrow U$ ;
3    $S \leftarrow \emptyset$ ;
4   mientras not at end of this document hacer
5       read current;
6       si understand entonces
7           go to next section;
8           current section becomes this one;
9       en otro caso
10          go back to the beginning of current section;
11       fin
12   fin
13 fin
```

---

# APÉNDICE B. ANOTHER APPENDIX CHAPTER

Como se puede apreciar en la Tabla B.1.

Tabla B.1: Ejemplo de una tabla

header1	header2	header3
1	2	3
4	5	6
7	8	9

Fuente: Elaboración propia, (2017)

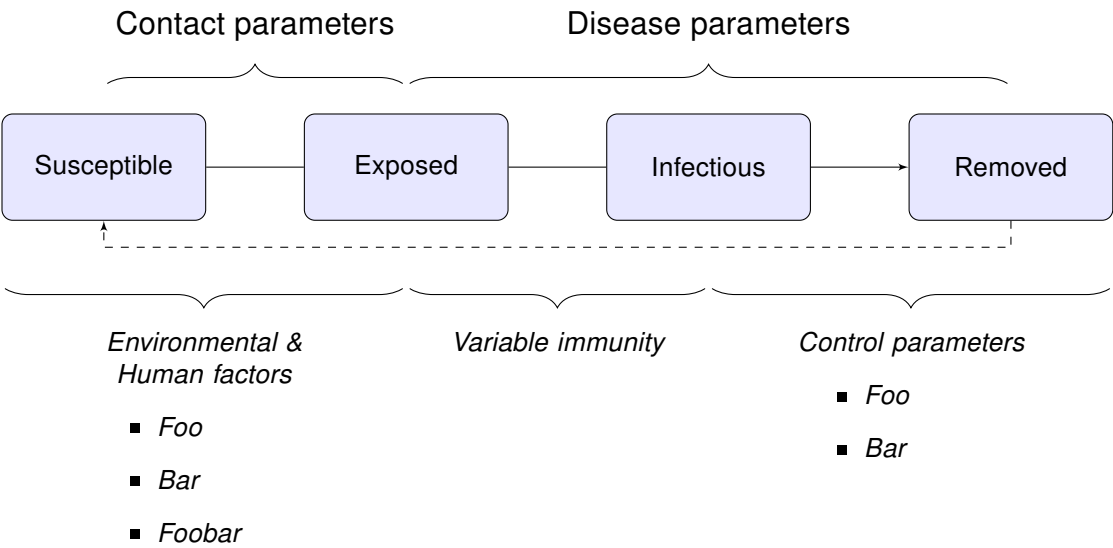
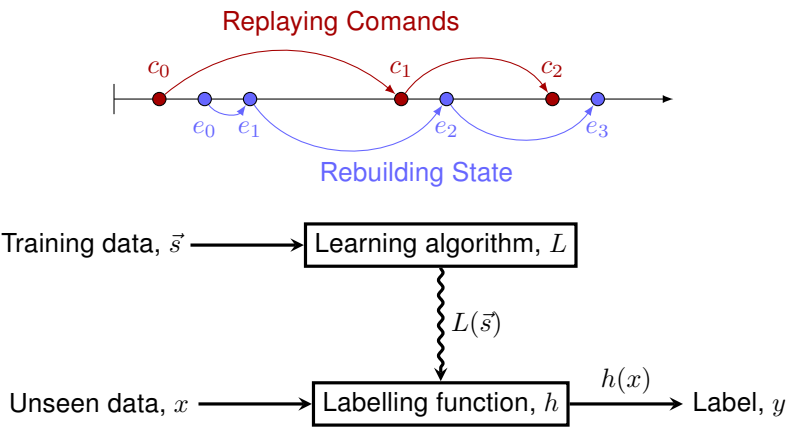
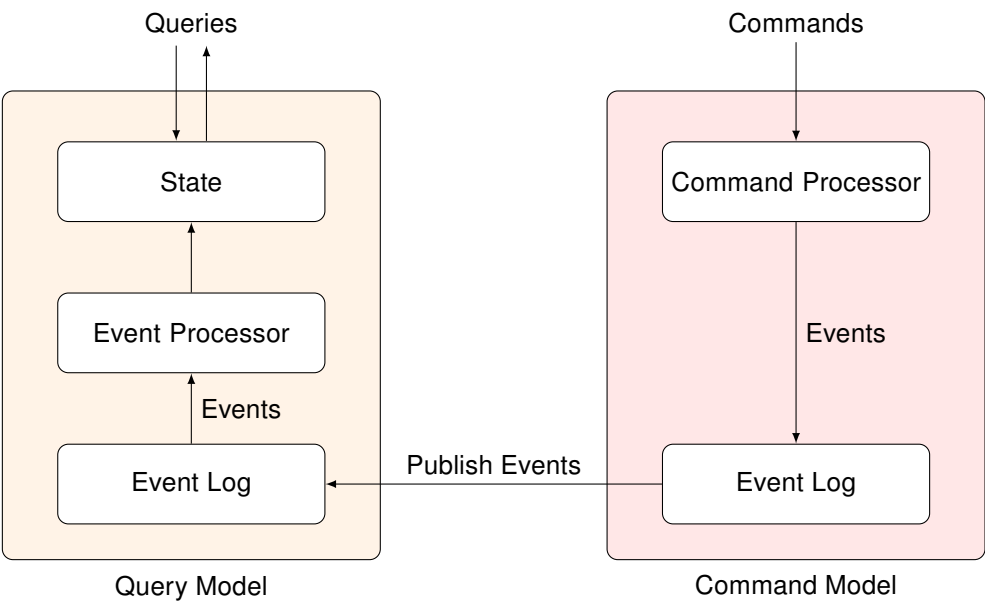


Figura B.1: Flowchart of fundamental disease transmission mechanisms



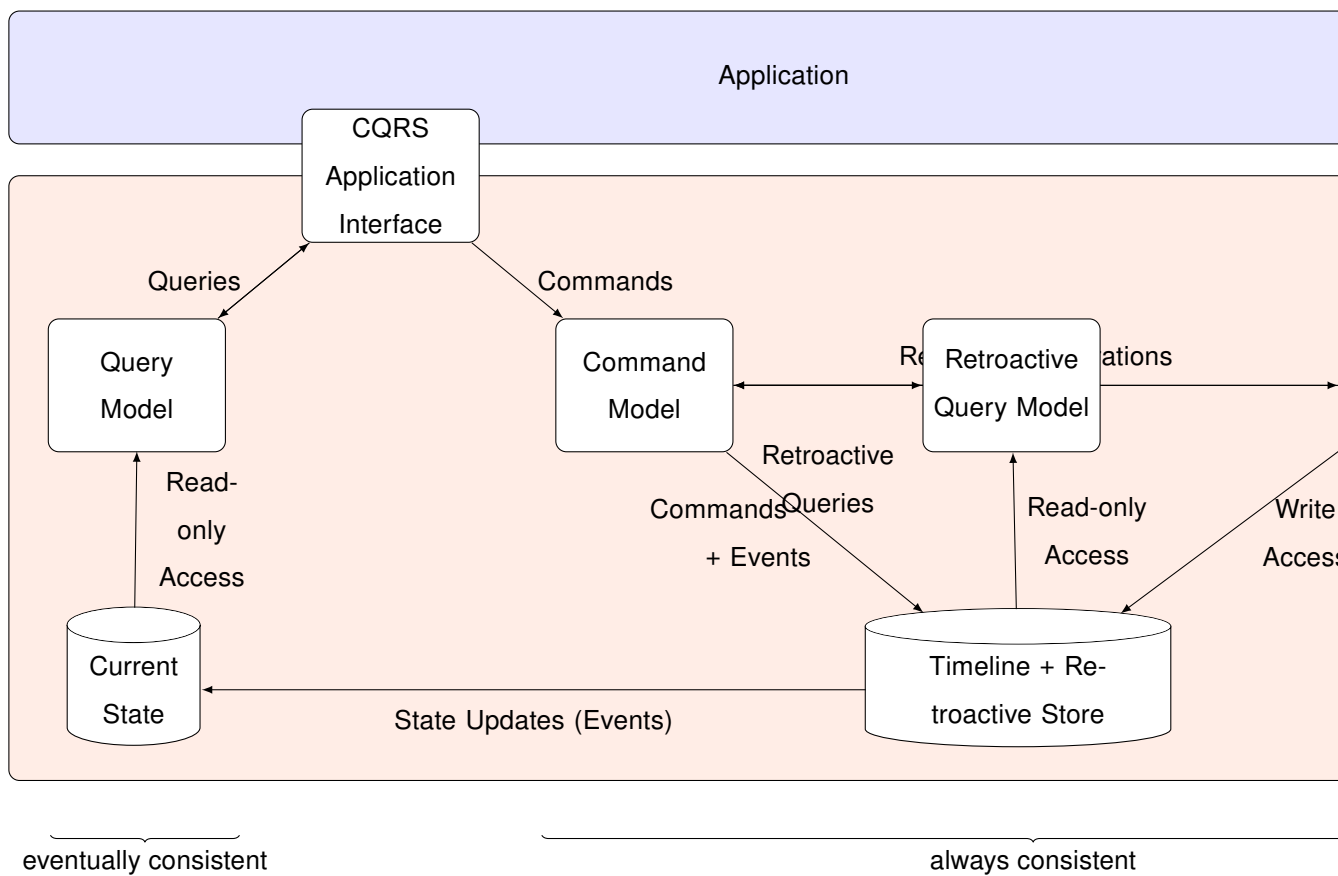
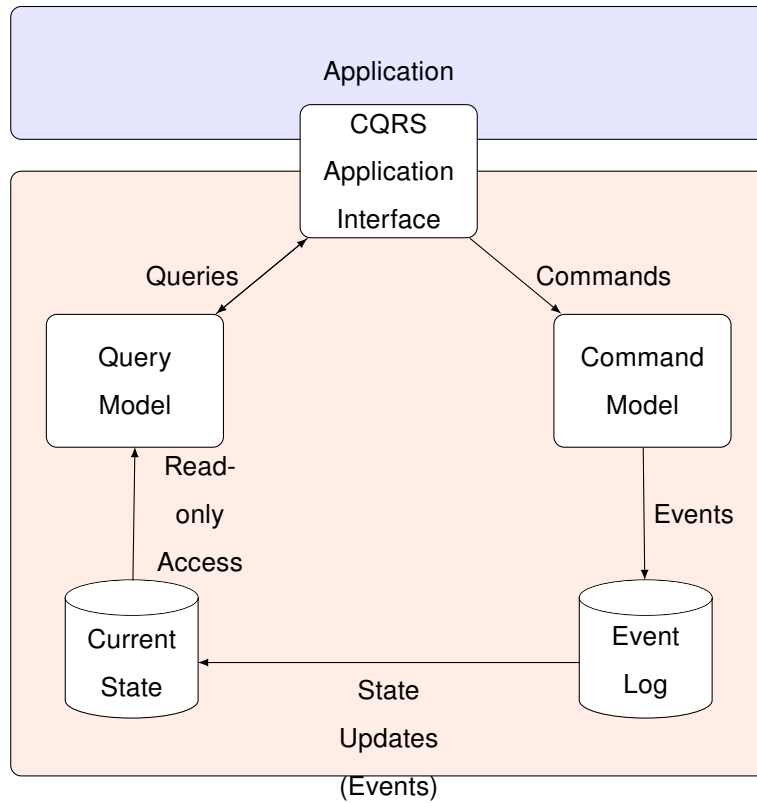


Expert labels the observation as  $a(n)$

		Anomaly ( $C_A$ )	Normality ( $C_N$ )
Algorithm classifies the data point as an	Outlier ( $C_O$ )	<b>hit</b> H	<b>false alarm</b> FA
	Inlier ( $C_I$ )	<b>miss</b> M	<b>correct reject</b> CR

Prediction outcome

		p	n	total
actual value	p'	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
total		P	N	





	ml		kdd	
Feature	kn	nddp	wdd	svddlofloci
Estimate local density	✓		✓	✓
Estimate global density		✓		
Domain based			✓	