

# Tabla de Contenido

<b>1. Introducción</b>	<b>1</b>
1.1. Antecedentes . . . . .	1
1.1.1. WIC . . . . .	2
1.1.2. AKORI . . . . .	3
1.2. Justificación . . . . .	4
1.3. Objetivos . . . . .	5
1.3.1. Objetivo General . . . . .	5
1.3.2. Objetivos Específicos . . . . .	5
1.4. Hipótesis de investigación . . . . .	5
1.5. Resultados Esperados . . . . .	6
1.6. Alcances . . . . .	6
1.7. Metodología . . . . .	6
1.8. Estructura del informe . . . . .	7
<b>2. Marco Conceptual</b>	<b>8</b>
2.1. Internet . . . . .	8
2.1.1. Web . . . . .	8
2.1.2. Sitio Web . . . . .	9
2.1.3. Página web . . . . .	9
2.1.4. URL . . . . .	9
2.2. Web Crawling . . . . .	10
2.3. Web Scraping . . . . .	10
2.4. Web mining . . . . .	10
2.5. Vector de Características . . . . .	11
2.6. Proceso KDD . . . . .	11
2.6.1. Minería de datos . . . . .	12
2.6.2. Minería de textos . . . . .	20
2.6.3. Evaluación de clasificadores . . . . .	23
2.6.4. Métodos de validación . . . . .	24
2.7. Arquitectura de seguridad . . . . .	25
2.7.1. Certificado de seguridad X.509 . . . . .	27
2.7.2. Metodología de Rating según seguridad web por SSL LABS . . . . .	30
<b>3. Categorización Web</b>	<b>34</b>
3.1. Categorización por Servicio . . . . .	35
3.2. Enfoques para Clasificar Páginas Web . . . . .	37

3.2.1.	Clasificación por Texto URL . . . . .	37
3.2.2.	Clasificación por Contexto . . . . .	37
3.2.3.	Clasificación por Contenido . . . . .	38
3.3.	Clasificación propuesta . . . . .	38
3.4.	Vector de Características . . . . .	38
3.4.1.	Contenido Web . . . . .	38
3.4.2.	Certificado de seguridad . . . . .	39
<b>4.</b>	<b>Implementación del vector de características</b>	<b>40</b>
4.1.	Construcción de juego de datos . . . . .	40
4.1.1.	Etiquetado de juego de datos . . . . .	41
4.2.	Características de sitios web por categoría . . . . .	42
4.2.1.	Características según diseño . . . . .	42
4.2.2.	Desarrollo de Software y Características según contenido HTML . . . . .	45
4.2.3.	Seguridad según SSL/TLS . . . . .	57
<b>5.</b>	<b>Minería de datos</b>	<b>61</b>
5.1.	Especificaciones técnicas . . . . .	61
5.1.1.	Hardware . . . . .	61
5.1.2.	Software . . . . .	62
5.2.	Modelamiento . . . . .	62
5.2.1.	Algoritmos de minería de datos . . . . .	63
5.3.	Minería de textos . . . . .	73
5.3.1.	Algoritmos de minería de datos en text mining . . . . .	75
5.3.2.	Minería de datos y Seguridad web como variable de decisión . . . . .	87
<b>6.</b>	<b>Resultados</b>	<b>92</b>
6.1.	Análisis de resultados esperados . . . . .	92
6.1.1.	R1: Definir las categorías a considerar de sitios web Chilenos . . . . .	92
6.1.2.	R2: Definir los parámetros del Vector de Características (Feature Vector) . . . . .	93
6.1.3.	R3: Clasificar páginas web . . . . .	94
6.1.4.	R5: Validar la hipótesis de investigación . . . . .	99
<b>7.</b>	<b>Conclusiones</b>	<b>100</b>
7.1.	Conclusiones Generales . . . . .	100
7.2.	Trabajo futuro y recomendaciones . . . . .	102
	<b>Bibliografía</b>	<b>104</b>
	<b>Anexos</b>	<b>109</b>
<b>A.</b>	<b>Juego de Variables</b>	<b>109</b>
A.1.	Juego de Variables de contenido HTML . . . . .	109
A.2.	Juego de Variables de texto . . . . .	110
A.3.	Variable de seguridad . . . . .	111
<b>B.</b>	<b>Reducción de dimensionalidad</b>	<b>112</b>

