

TABLA DE CONTENIDOS

	página
Dedicatoria	I
Agradecimientos	II
Tabla de Contenidos	III
Índice de Figuras	VI
Índice de Tablas	X
Resumen	XI
Abstract	XII
1. Introducción	1
1.1. Contexto del proyecto	1
1.1.1. Trabajo relacionado	2
1.2. Definición del problema	4
1.3. Propuesta de solución	4
1.4. Hipótesis	5
1.5. Objetivos	5
1.6. Alcances	6
1.7. Resumen del Capítulo	7
2. Marco teórico	8
2.1. Análisis de datos	8
2.1.1. ¿En qué consiste el análisis de datos?	9
2.1.2. Etapas del proceso de análisis de datos	9
2.1.3. Principales problemas que resuelve el análisis de datos	12
2.2. Algoritmos de regresión	15
2.2.1. Regresión lineal	15
2.2.2. Algoritmo PLS	16
2.2.3. Algoritmo RGML	17

2.2.4. Ridge	20
2.3. Visualización de datos	20
2.4. Diseño web	28
2.4.1. Usabilidad y satisfacción de uso	29
2.4.2. Arquitectura cliente-servidor	30
2.5. Resumen del Capítulo	33
3. Metodología de desarrollo	34
3.1. Ambiente de desarrollo	34
3.1.1. RStudio	35
3.1.2. R	35
3.1.3. Shiny	36
3.1.4. Amazon Web Services (AWS)	37
3.2. Metodología de desarrollo de software	38
3.2.1. Metodología SCRUM	38
3.2.2. Implementación	41
3.2.3. Control de Versiones	43
3.3. Resumen del Capítulo	46
4. Desarrollo de la Solución	47
4.1. Historias de usuario	47
4.2. Mock-up	51
4.3. Arquitectura	57
4.4. Modularización del sistema	59
4.5. Resumen del Capítulo	60
5. Diseño de interfaz	61
5.1. Home	61
5.2. Data	63
5.2.1. Interfaz del módulo Data	65
5.3. Preprocessing	73
5.3.1. Interfaz del módulo Preprocessing	76
5.4. Transformation	82
5.4.1. Interfaz del módulo Transformation	84
5.5. Regression	91

5.5.1. Interfaz del módulo <i>Regression</i>	92
5.6. <i>Linear Model Evaluation</i>	103
5.6.1. Interfaz del módulo <i>Linear Model Evaluation</i>	104
5.7. <i>Report</i>	107
5.8. Resumen del Capítulo	110
6. Validación	111
6.1. Datos faltantes y Regresión Lineal	111
6.2. Validación por aproximación	120
6.3. Componentes Principales	128
6.4. <i>Ridge y PLS</i>	132
6.5. Evaluación Heurística	138
6.5.1. Visibilidad del estado del sistema	139
6.5.2. Partido entre el sistema y el mundo real	139
6.5.3. Control del usuario y la libertad	140
6.5.4. Consistencia y estándares	140
6.5.5. Prevención de errores	141
6.5.6. Reconocimiento más que Recordación	141
6.5.7. La flexibilidad y la eficiencia del uso	142
6.5.8. Diseño estético y minimalista	142
6.5.9. Ayude a los usuarios a reconocer, diagnosticar y recuperarse de errores	143
6.5.10. Ayuda y documentación	143
6.5.11. Notas	143
6.6. Resumen del Capítulo	144
7. Conclusiones	145
7.1. Trabajo Futuro	146
Bibliografía	147
Anexos	
A: Estructura del prototipo web	151
B: Estructura del repositorio <i>GitHub</i>	154

ÍNDICE DE FIGURAS

	página
2.1. Disciplinas que engloba la minería de datos	9
2.2. Etapas del KDD	10
2.3. Técnicas utilizadas para el análisis de datos	11
2.4. Fases del análisis de datos	12
2.5. Ejemplo de histograma, se puede ver un histograma de <i>Petal Width</i> con 10 y 20 barras respectivamente	22
2.6. Ejemplo de PDF y CDF	23
2.7. Ejemplo de ECDF	23
2.8. Ejemplo básico de diagrama de caja	24
2.9. Ejemplo de comparación de atributos en un diagrama de caja	24
2.10. Ejemplo de gráfico de dispersión	25
2.11. Ejemplo de histograma para temperatura superficial del mar (TSM)	26
2.12. Ejemplo de <i>matrix plots</i> con iris	26
2.13. Ejemplo de <i>matrix plots</i> de correlación con iris	27
2.14. Ejemplo de coordenadas paralelas con iris	27
2.15. Ejemplo de gráfico de estrella	28
2.16. Ejemplo de <i>Chernoff Faces</i>	28
2.17. Experiencia del usuario frente a un sitio web	30
2.18. Funcionamiento de la arquitectura cliente/servidor	33
3.1. Visión esquemática del funcionamiento de R	36
3.2. Estructura de la metodología SCRUM	39
3.3. Relaciones entre el equipo de trabajo	40
3.4. Flujo de sprints	41
3.5. Estructura del prototipo web	42
3.6. Git almacena la información como instantáneas del proyecto a lo largo del tiempo	44
3.7. Flujo de trabajo entre la máquina local y <i>Amazon</i>	45
4.1. Interfaz de la sección Data del Mock-up	52
4.2. Interfaz de la sección proyectos del módulo Data del Mock-up	53

4.3.	Interfaz de la sección Análisis exploratorio del Mock-up	54
4.4.	Interfaz del módulo Entrenar del Mock-up	55
4.5.	Interfaz del módulo Predecir del Mock-up	56
4.6.	Interfaz del módulo Validación del Mock-up	57
4.7.	Arquitectura del prototipo web	58
4.8.	Modularización del sistema	60
5.1.	Página principal del prototipo web	62
5.2.	Güiña, imagen obtenida de Fauna Australis	62
5.3.	interfaz del módulo Data	64
5.4.	Conjunto de datos pre-cargados en el sistema	66
5.5.	Conjunto de datos pre-cargados en el sistema	67
5.6.	Incorporar nuevos conjunto de datos al sistema	68
5.7.	Incorporar un nuevo conjunto de datos al sistema mediante una <i>URL</i>	69
5.8.	Editar un conjunto de datos	70
5.9.	Visualizar un conjunto de datos	71
5.10.	interfaz mediante un <i>Parallel plot</i>	71
5.11.	Herramientas para manejo dinámico de la interfaz	72
5.12.	Opcion para alterar la paleta de colores	73
5.13.	Interfaz del módulo Preprocessing	74
5.14.	Resultado tras aplicar LOF al conjunto de datos airquality	75
5.15.	Visualización de datos faltantes en un <i>Box Plot</i>	77
5.16.	Interfaz de datos faltantes en un Histograma	78
5.17.	Interfaz de datos faltantes en un <i>Scatter plot</i>	79
5.18.	Interfaz del submódulo LOF	80
5.19.	Interfaz del submódulo eliminación de ruido	82
5.20.	Interfaz del submódulo PCA	83
5.21.	Interfaz del módulo Transformation	85
5.22.	Interfaz del submódulo Normalización	86
5.23.	Conjunto de normalizaciones adicionales	87
5.24.	Visualización del submódulo SVD	89
5.25.	Visualización del submódulo Selección de Atributos	90
5.26.	Tipos de validaciones y parámetros generales de los modelos lineales	93
5.27.	Interfaz del módulo Regression	94

5.28. Regresión Lineal sobre el conjunto de datos <i>airquality</i>	95
5.29. Predicción de Ozono aplicando Regresión Lineal	96
5.30. Regresión lineal por mínimo cuadrados sobre el conjunto de datos <i>airquality</i>	97
5.31. Predicción de Ozono aplicando PLS	98
5.32. Ridge sobre el conjunto de datos <i>airquality</i>	99
5.33. Parámetro de ajuste lambda de Ridge para el conjunto de datos <i>air-</i> <i>quality</i>	100
5.34. Predicción de Ozono aplicando Ridge	101
5.35. RGLM sobre el conjunto de datos <i>airquality</i>	102
5.36. Predicción de Ozono aplicando RGLM	103
5.37. Interfaz del módulo Linear Model Evaluation	104
5.38. Visualización del gráfico <i>Residuals vs Fitted</i>	105
5.39. Visualización del gráfico <i>Normal Q-Q</i>	106
5.40. Visualización del gráfico <i>Residuals vs Leverage</i>	107
5.41. Reporte del trabajo realizado en el prototipo web parte 1	108
5.42. Reporte del trabajo realizado en el prototipo web parte 2	109
6.1. Cargar el conjunto de datos <i>algae</i> a Güiña	113
6.2. Datos faltantes del conjunto de datos <i>algae</i>	114
6.3. Estructura interna del conjunto de datos <i>algae</i>	115
6.4. Aplicando Atribute Selection sobre el conjunto de datos <i>algae</i>	118
6.5. Aplicando Linear Model sobre el conjunto de datos <i>algae</i>	119
6.6. Aplicando Linear Model sobre el conjunto de datos <i>algae</i>	120
6.7. Cargar el conjunto de datos <i>Auto</i> a Güiña	123
6.8. Aplicando Linear Model sobre el conjunto de datos <i>Auto</i>	124
6.9. Tipo de validación y conjunto de variables predictoras para aplicar Regresión Lineal	125
6.10. Predicción de las millas por galón del conjunto de datos <i>Auto</i>	126
6.11. Predicción gráfica de las millas por galón del conjunto de datos <i>Auto</i>	127
6.12. Componentes Principales para el conjunto de datos <i>USArrests</i>	130
6.13. Cargar el conjunto de datos <i>USArrests</i> a Güiña	131
6.14. Componentes Principales para el conjunto de datos <i>USArrests</i> en Güiña	131
6.15. <i>Lambda</i> obtenido por <i>Ridge</i> con validación cruzada	133

6.16. Cargar el conjunto de datos <i>Hitters</i> a Güiña	135
6.17. Datos faltantes del conjunto <i>Hitters</i>	135
6.18. <i>Ridge</i> sobre el conjunto de datos <i>Hitters</i>	136
6.19. <i>Lambda</i> obtenido por <i>Ridge</i> con validación cruzada	137
6.20. <i>PLS</i> sobre el conjunto de datos <i>Hitters</i>	138
6.21. Consistencia gráfica delm prototipo web	141
A.1. Primera estructura del prototipo web	151
A.2. Segunda estructura del prototipo web	152
A.3. Estructura final del prototipo web	153
B.1. Estructura del repositorio <i>Github</i>	154
B.2. Archivos del directorio <i>funciones</i>	155

ÍNDICE DE TABLAS

	página
3.1. Ambiente de desarrollo donde se trabajó el prototipo	34
6.1. Comparación de la predicción realizada mediante Validación por Aproximación	128