

TABLA DE CONTENIDO

1	Introducción.....	1
1.1	Contenido de los capítulos.....	1
1.2	Fundamentación.....	1
1.3	Objetivos.....	2
1.3.1	Objetivo General.....	2
1.3.2	Objetivos específicos.....	2
1.3.3	Propuesta.....	2
2	Contextualización, marco teórico y estado del arte.....	3
2.1	Educación Superior en Chile.....	3
2.1.1	Deserción en la Educación Superior.....	4
2.1.2	Posibles Causas de la deserción según Modelos Teóricos.....	6
2.1.3	Deserción en la Universidad de las Américas (UDLA).....	7
2.1.4	Acreditación en la Universidad de las Américas.....	8
2.2	Clasificación.....	9
2.3	Knowledge Discovery in Databases (KDD).....	10
2.4	Validación cruzada.....	11
2.5	Selección de Variables.....	13
2.5.1	Filtro.....	14
2.5.2	Iterativa.....	14
2.5.3	Métodos embebidos.....	15
2.6	Uso de variables categóricas.....	16
2.7	Valores faltantes.....	17
2.8	Métodos de clasificación.....	17

2.8.1	Árboles de decisión	18
2.8.2	Random Forest.....	20
2.9	Presentación de resultados.....	23
2.9.1	Curva ROC.....	23
2.9.2	Matriz de confusiones.....	25
2.10	Estado del arte de la predicción de deserción de estudiantes.....	26
3	Metodología.....	29
3.1	Estudio del contexto.....	29
3.1.1	Definición de objetivos.....	29
3.1.2	Antecedentes de UDLA	29
3.1.3	Análisis de información disponible	30
3.2	Selección	32
3.2.1	Falta de datos históricos.....	32
3.2.2	Información al avanzar el semestre	32
3.2.3	Definición de la base de datos.....	34
3.2.4	Definición de las etiquetas	35
3.3	Preprocesamiento de datos	37
3.3.1	Reparación de datos erróneos.....	37
3.3.2	Tratamiento de datos faltantes	37
3.4	Transformación	38
3.5	Data Mining.....	57
3.5.1	Variables	57
3.5.2	Separación de conjuntos	58
3.5.3	Selección de características	59
3.5.4	Clasificador.....	60

3.6	Interpretación y exploración	61
3.6.1	Evaluación de los modelos	61
4	Resultados	63
4.1	Parámetros de los modelos para cada etapa	63
4.2	Desempeño de los modelos con los datos de los conjuntos de prueba.....	65
4.2.1	Curvas ROC	65
4.2.2	Matrices de Confusión	67
4.3	VARIABLES MÁS RELEVANTES.....	70
4.4	Comparación de Promedio de variables más importantes en matriz de confusión 73	
5	Análisis de Resultados	77
5.1	Generales	77
5.2	Modelo inicio de semestre.....	77
5.3	Modelo cátedra 1	79
5.4	Modelo Cátedra 2.....	80
5.5	Modelo Fin de semestre.....	82
6	Conclusiones.....	84
7	Glosario de términos y abreviaciones	86
8	Bibliografía	87
Anexo 1.	Tipos de Variable disponibles para cada modelo.....	93
Anexo 2.	Metodologías utilizadas que no entregaron mejoras	95
Anexo 2.1	SOM Based Stratified Sampling	95
Anexo 2.1.1	Mapas autoorganizativos.....	95
Anexo 2.1.2	SOM Based Stratified Sampling	97
Anexo 2.2	Balanceo de bases de datos.....	97

Anexo 2.2.1	Resultados con SMOTE	98
Anexo 3.	Resultados con Redes Neuronales.....	100
Anexo 3.1	Generalidades	100
Anexo 3.2	Normalización.....	102
Anexo 3.3	Matrices de Confusión de las Redes Neuronales	102
Anexo 3.3.1	Inicio de semestre	102
Anexo 3.3.2	Cátedra 1	103
Anexo 3.3.3	Cátedra 2	104
Anexo 3.3.4	Fin de semestre.....	104
Anexo 4.	Entrenamiento con retropropagación del error.....	105