

TABLA DE CONTENIDOS

	página
Dedicatoria	I
Agradecimientos	II
Tabla de Contenidos	III
Índice de Figuras	VII
Índice de Tablas	X
Resumen	XVI
1. Introducción	17
1.1. Descripción de la propuesta	17
1.2. Contexto del proyecto	18
1.3. Definición del problema	19
1.4. Propuesta de solución	21
1.5. Preguntas de investigación	22
1.6. Objetivos	22
1.7. Alcances	23
1.8. Descripción de contenido	23
2. Marco teórico	24
2.1. <i>Machine Learning</i>	24
2.2. Aprendizaje supervisado	25
2.2.1. Regresión	25
2.2.2. Regresión lineal	26
2.2.3. <i>Ordinary Least Squares Regression</i>	28
2.2.4. <i>Ridge Regression</i>	29
2.2.5. <i>Lasso Regression</i>	30
2.2.6. <i>Partial Least Squares</i>	30
2.2.7. Regresión no lineal	31

2.2.8.	<i>Support Vector Machines</i>	32
2.2.9.	<i>Support Vector Regression</i>	34
2.3.	Aprendizaje no supervisado	35
2.3.1.	<i>Clustering</i>	35
2.3.2.	K-means	36
2.3.3.	<i>Clustering</i> jerárquico	36
2.4.	Evaluación de modelos	39
2.4.1.	Métricas de desempeño	39
2.4.2.	<i>k-Fold Cross-validation</i>	40
2.4.3.	<i>Nested Cross-validation</i>	41
2.4.4.	<i>5x2 Cross-validation</i>	43
2.5.	Preprocesamiento de datos	43
2.5.1.	Normalización Z-Score	43
2.5.2.	Coefficiente de correlación de Pearson	43
2.5.3.	Algoritmo de Selección de Atributos Relief	44
2.6.	Antecedentes del dominio de aplicación	45
2.6.1.	Estabilidad conformacional	45
2.6.2.	Mutaciones de proteínas y estabilidad	45
2.6.3.	Vectores de Autocorrelación de Secuencia de Aminoácidos	46
2.6.4.	Investigaciones anteriores relacionadas con <i>machine learning</i> y Vectores AASA	48
2.7.	Conceptos para el desarrollo de la investigación	49
2.7.1.	Metodología de Minería de Datos CRISP-DM	49
2.7.2.	Tecnologías	51
3.	Metodología	55
3.1.	Metodología experimental	55
3.2.	Entendimiento del dominio	56
3.3.	Entendimiento de los datos	56
3.3.1.	Análisis Exploratorio de Datos	57
3.4.	Preparación de los Datos	57
3.4.1.	Normalización de datos	57
3.4.2.	Reducción de la dimensionalidad	58
3.5.	Modelado	60

3.5.1.	Algoritmos de entrenamiento	60
3.5.2.	Métricas de evaluación	60
3.5.3.	Optimización de hiperparámetros	61
3.5.4.	Entrenamiento y evaluación	62
3.6.	Evaluación	64
3.6.1.	Criterios de evaluación de modelado predictivo	64
3.6.2.	Criterio de evaluación de reproducibilidad	65
3.6.3.	Interpretación y análisis de resultados	65
3.7.	Despliegue	65
4.	Desarrollo y análisis de resultados	66
4.1.	Entendimiento de los datos	66
4.1.1.	Conjuntos de datos	66
4.1.2.	Correlación entre descriptores	67
4.1.3.	Valores distintos de la estabilidad	67
4.1.4.	Distribución de la estabilidad	68
4.1.5.	Detección de observaciones atípicas	69
4.1.6.	Correlación lineal entre descriptores y estabilidad	70
4.1.7.	<i>Clustering</i>	70
4.2.	Preparación de los datos	74
4.2.1.	Limpieza e integridad de los datos	74
4.2.2.	Descarte de descriptores intercorrelacionados	75
4.2.3.	Selección de atributos	75
4.3.	Modelado y evaluación de modelos	77
4.3.1.	Grilla de valores para hiperparámetros	77
4.3.2.	Experimento 1 (E1): Modelado usando todos los descriptores .	77
4.3.3.	Experimento 2 (E2): Modelado usando datos agregados con media aritmética	79
4.3.4.	Experimento 3 (E3): Modelado usando datos reducidos por selección de atributos	80
4.4.	Discusión	87
4.5.	Aspectos metodológicos mejorables	90

5. Conclusiones y trabajo futuro	92
5.1. Conclusiones	92
5.2. Trabajos futuros	94

Anexos

A: Anexos Análisis Exploratorio de Datos	99
A.1. Distribución de estabilidad	99
A.2. Correlación lineal entre descriptores y estabilidad	102
A.3. <i>Clustering</i>	104
B: Anexos Preparación de los datos	115
B.1. Selección de descriptores con algoritmo SURF	116
C: Anexos Modelado y Evaluación	118
C.1. Anexos Experimento 1	118
C.2. Anexos Experimento 2	120
C.3. Anexos Experimento 3	122

ÍNDICE DE FIGURAS

	página
1.1. Equipo de trabajo y labores de investigación.	19
2.1. Ejemplo de hiperplano de dos dimensiones para una variable de respuesta Y , formado dos predictores, X_1 y X_2 . Figura extraída de [15, pág. 73].	27
2.2. Ejemplo de la transformación realizada por PLS antes de ajustar el modelo. A la izquierda se grafican observaciones en función de dos predictores, <i>Population</i> y <i>Ad Spending</i> . A la derecha se grafican las observaciones en función de sus componentes principales calculadas a partir de dichos predictores. Figura extraída de [15, pág. 232].	31
2.3. Diferencia entre un márgenes suaves y rígidos para el caso de un hiperplano separador unidimensional	32
2.4. Ejemplo de <i>kernel trick</i> . Los datos no son separables en \mathbb{R}^2 . Sin embargo, sus proyecciones en \mathbb{R}^3 dadas por una función <i>kernel</i> sí lo son.	33
2.5. Ejemplo de <i>kernel trick</i> en SVR. El modelo no lineal en el espacio original corresponde al modelo modelo lineal en el espacio extendido.	35
2.6. Ejemplo de funcionamiento de <i>K-means</i> con $K = 3$. Figura extraída de [15, pág. 389].	37
2.7. Ejemplo de resultado <i>clustering</i> jerárquico aglomerativo. De izquierda a derecha, se grafican las <i>clusterizaciones</i> correspondientes para un valor de k entre 1 y 3, respectivamente. Figura extraída de [15, pág. 392].	38
2.8. Diagrama explicativo de k-Fold Cross-validation para un valor k igual a 10.	41
2.9. Diagrama de Nested Cross-validation. Figura extraída de [18].	42
2.10. Ejemplo de estructura de un conjunto de datos codificado como vectores AASA.	47
2.11. Cálculo de un vector AASA con un $l = 5$ para la propiedad p_k	48
2.12. Ilustración de metodología CRISP-DM.	50
2.13. Tecnologías a usar para el desarrollo de la investigación.	51
3.1. Esquema de metodología de investigación.	56
3.2. Esquema de evaluación de desempeño de modelos entrenados.	61

3.3.	Esquema del ciclo interno del proceso de <i>Nested Cross-validation</i> , correspondiente a la optimización de hiperparámetros.	62
3.4.	Esquema del ciclo externo del proceso de <i>Nested Cross-validation</i> , correspondiente a la estimación del desempeño de generalización de un algoritmo de entrenamiento.	64
4.1.	<i>Boxplots</i> de las distribuciones de estabilidad para cada conjunto.	69
4.2.	Agrupaciones arrojadas por <i>K-means</i> con 3 <i>clusters</i> para el conjunto de datos 4LYZ.	72
4.3.	Agrupaciones arrojadas por <i>clustering</i> jerárquico con 3 <i>clusters</i> para el conjunto de datos 4LYZ.	72
4.4.	Mapa de calor de la matriz de distancia euclidiana calculada para el conjunto de datos 4LYZ.	73
4.5.	Mapa de calor de la matriz de distancia <i>Manhattan</i> calculada para el conjunto de datos 4LYZ.	74
A.1.	<i>Boxplot</i> de la distribución de estabilidad para conjunto 1STN.	99
A.2.	<i>Boxplot</i> de la distribución de estabilidad para conjunto 4LYZ.	100
A.3.	<i>Boxplot</i> de la distribución de estabilidad para conjunto 1BPI.	101
A.4.	<i>Boxplot</i> de la distribución de estabilidad para conjunto HLYZ.	102
A.5.	Agrupaciones arrojadas por <i>K-means</i> con 3 <i>clusters</i> para el conjunto de datos 1STN.	104
A.6.	Agrupaciones arrojadas por <i>clustering</i> jerárquico con 3 <i>clusters</i> para el conjunto de datos 1STN.	105
A.7.	Agrupaciones arrojadas por <i>K-means</i> con 3 <i>clusters</i> para el conjunto de datos 1BPI.	106
A.8.	Agrupaciones arrojadas por <i>clustering</i> jerárquico con 3 <i>clusters</i> para el conjunto de datos 1BPI.	107
A.9.	Agrupaciones arrojadas por <i>K-means</i> con 3 <i>clusters</i> para el conjunto de datos HLYZ.	108
A.10.	Agrupaciones arrojadas por <i>clustering</i> jerárquico con 3 <i>clusters</i> para el conjunto de datos HLYZ.	109
A.11.	Mapa de calor de la matriz de distancia euclidiana calculada para el conjunto de datos 1STN.	109

A.12. Mapa de calor de la matriz de distancia <i>Manhattan</i> calculada para el conjunto de datos 1STN.	110
A.13. Mapa de calor de la matriz de distancia euclidiana calculada para el conjunto de datos 1BPI.	111
A.14. Mapa de calor de la matriz de distancia <i>Manhattan</i> calculada para el conjunto de datos 1BPI.	112
A.15. Mapa de calor de la matriz de distancia euclidiana calculada para el conjunto de datos HLYZ.	113
A.16. Mapa de calor de la matriz de distancia <i>Manhattan</i> calculada para el conjunto de datos HLYZ.	114

ÍNDICE DE TABLAS

	página
2.1. Ejemplo de conjuntos de datos real codificado usando vectores AASA.	48
4.1. Dimensiones de los conjuntos de datos abordados.	67
4.2. Número de descriptores intercorrelacionados para umbrales de r de Pearson entre 0.90 y 0.99, para cada conjunto de datos.	67
4.3. Porcentaje de valor distintos para la estabilidad en cada uno de los conjuntos de datos.	68
4.4. Cantidad de observaciones según su variación de estabilidad, por cada conjunto de datos.	68
4.5. Los 10 descriptores más correlacionados con la estabilidad en el conjunto HLYZ.	70
4.6. Los 10 descriptores más correlacionados con la estabilidad en el conjunto 1STN.	71
4.7. Los 40 descriptores más relevantes seleccionados por SURF para el conjunto HLYZ.	75
4.8. Los 40 descriptores más relevantes seleccionados por SURF para el conjunto 1BPI.	76
4.9. Valores posibles para los hiperparámetros de cada algoritmo probados en la optimización de hiperparámetros con <i>Grid Search</i>	77
4.10. Resumen de puntajes de prueba obtenidos en Experimento 1 para conjunto HLYZ.	78
4.11. Resumen de puntajes de entrenamiento obtenidos en Experimento 1 para conjunto HLYZ.	78
4.12. Resumen de puntajes de prueba obtenidos en Experimento 1 para conjunto 1BPI.	79
4.13. Resumen de puntajes de prueba obtenidos en Experimento 2 para conjunto HLYZ.	80
4.14. Resumen de puntajes de entrenamiento obtenidos en Experimento 2 para conjunto HLYZ.	80
4.15. Resumen de puntajes de prueba obtenidos en Experimento 3 para conjunto HLYZ usando validación Nested CV.	81

4.16. Resumen de puntajes de entrenamiento obtenidos en Experimento 3 para conjunto HLYZ usando validación Nested CV.	81
4.17. Resumen de puntajes de prueba obtenidos en Experimento 3 para conjunto 1STN, usando <i>Nested</i> , 5-Fold y 5x2 <i>Cross-validation</i>	83
4.18. Resumen de puntajes de prueba obtenidos en Experimento 3 para conjunto 4LYZ, usando <i>Nested</i> , 5-Fold y 5x2 <i>Cross-validation</i>	84
4.19. Resumen de puntajes de prueba obtenidos en Experimento 3 para conjunto 1BPI, usando <i>Nested</i> , 5-Fold y 5x2 <i>Cross-validation</i>	85
4.20. Resumen de puntajes de prueba obtenidos en Experimento 3 para conjunto HLYZ, usando <i>Nested</i> , 5-Fold y 5x2 <i>Cross-validation</i>	86
4.21. Hiperparámetros seleccionados por cada algoritmo de aprendizaje y por cada conjunto de datos, a partir de los resultados del Experimento 3 con métodos de validación alternativos.	87
A.1. Los 10 descriptores más correlacionados con la estabilidad en el conjunto 4LYZ.	102
A.2. Los 10 descriptores más correlacionados con la estabilidad en el conjunto 1BPI.	103
B.1. Los 40 descriptores más relevantes seleccionados por SURF para el conjunto 1STN.	116
B.2. Los 40 descriptores más relevantes seleccionados por SURF para el conjunto 4LYZ.	117
C.1. Resumen de puntajes de prueba obtenidos en Experimento 1 para conjunto 1STN.	118
C.2. Resumen de puntajes de entrenamiento obtenidos en Experimento 1 para conjunto 1STN.	119
C.3. Resumen de puntajes de prueba obtenidos en Experimento 1 para conjunto 4LYZ.	119
C.4. Resumen de puntajes de entrenamiento obtenidos en Experimento 1 para conjunto 4LYZ.	119
C.5. Resumen de puntajes de entrenamiento obtenidos en Experimento 1 para conjunto 1BPI.	120

C.6. Resumen de puntajes de prueba obtenidos en Experimento 2 para conjunto 1STN.	120
C.7. Resumen de puntajes de entrenamiento obtenidos en Experimento 2 para conjunto 1STN.	121
C.8. Resumen de puntajes de prueba obtenidos en Experimento 2 para conjunto 4LYZ.	121
C.9. Resumen de puntajes de entrenamiento obtenidos en Experimento 2 para conjunto 4LYZ.	121
C.10. Resumen de puntajes de prueba obtenidos en Experimento 2 para conjunto 1BPI.	122
C.11. Resumen de puntajes de entrenamiento obtenidos en Experimento 2 para conjunto 1BPI.	122
C.12. R^2 s de prueba en Experimento 3 para conjunto 1STN, usando 40 descriptores y 5-Fold CV.	123
C.13. R^2 s de prueba en Experimento 3 para conjunto 1STN, usando 40 descriptores y 5x2 CV.	123
C.14. R^2 s de prueba en Experimento 3 para conjunto 1STN, usando 40 descriptores y Nested CV.	123
C.15. RMSEs de prueba en Experimento 3 para conjunto 1STN, usando 40 descriptores y 5-Fold CV.	124
C.16. RMSEs de prueba en Experimento 3 para conjunto 1STN, usando 40 descriptores y 5x2 CV.	124
C.17. RMSEs de prueba en Experimento 3 para conjunto 1STN, usando 40 descriptores y Nested CV.	124
C.18. R^2 s de entrenamiento en Experimento 3 para conjunto 1STN, usando 40 descriptores y 5-Fold CV.	125
C.19. R^2 s de entrenamiento en Experimento 3 para conjunto 1STN, usando 40 descriptores y 5x2 CV.	125
C.20. R^2 s de entrenamiento en Experimento 3 para conjunto 1STN, usando 40 descriptores y Nested CV.	125
C.21. RMSEs de entrenamiento en Experimento 3 para conjunto 1STN, usando 40 descriptores y 5-Fold CV.	126
C.22. RMSEs de entrenamiento en Experimento 3 para conjunto 1STN, usando 40 descriptores y 5x2 CV.	126

C.23.RMSEs de entrenamiento en Experimento 3a para conjunto 1STN, usando 40 descriptores y Nested CV.	126
C.24. R^2 s de prueba en Experimento 3 para conjunto 4LYZ, usando 40 des- criptores y 5-Fold CV.	127
C.25. R^2 s de prueba en Experimento 3 para conjunto 4LYZ, usando 40 des- criptores y 5x2 CV.	127
C.26. R^2 s de prueba en Experimento 3 para conjunto 4LYZ, usando 40 des- criptores y Nested CV.	127
C.27.RMSEs de prueba en Experimento 3 para conjunto 4LYZ, usando 40 descriptores y 5-Fold CV.	128
C.28.RMSEs de prueba en Experimento 3 para conjunto 4LYZ, usando 40 descriptores y 5x2 CV.	128
C.29.RMSEs de prueba en Experimento 3 para conjunto 4LYZ, usando 40 descriptores y Nested CV.	128
C.30. R^2 s de entrenamiento en Experimento 3 para conjunto 4LYZ, usando 40 descriptores y 5-Fold CV.	129
C.31. R^2 s de entrenamiento en Experimento 3 para conjunto 4LYZ, usando 40 descriptores y 5x2 CV.	129
C.32. R^2 s de entrenamiento en Experimento 3 para conjunto 4LYZ, usando 40 descriptores y Nested CV.	129
C.33.RMSEs de entrenamiento en Experimento 3 para conjunto 4LYZ, usando 40 descriptores y 5-Fold CV.	130
C.34.RMSEs de entrenamiento en Experimento 3 para conjunto 4LYZ, usando 40 descriptores y 5x2 CV.	130
C.35.RMSEs de entrenamiento en Experimento 3a para conjunto 4LYZ, usando 40 descriptores y Nested CV.	130
C.36. R^2 s de prueba en Experimento 3 para conjunto 1BPI, usando 40 des- criptores y 5-Fold CV.	131
C.37. R^2 s de prueba en Experimento 3 para conjunto 1BPI, usando 40 des- criptores y 5x2 CV.	131
C.38. R^2 s de prueba en Experimento 3 para conjunto 1BPI, usando 40 des- criptores y Nested CV.	131
C.39.RMSEs de prueba en Experimento 3 para conjunto 1BPI, usando 40 descriptores y 5-Fold CV.	132

C.40.RMSEs de prueba en Experimento 3 para conjunto 1BPI, usando 40 descriptores y 5x2 CV.	132
C.41.RMSEs de prueba en Experimento 3 para conjunto 1BPI, usando 40 descriptores y Nested CV.	132
C.42. R^2 s de entrenamiento en Experimento 3 para conjunto 1BPI, usando 40 descriptores y 5-Fold CV.	133
C.43. R^2 s de entrenamiento en Experimento 3 para conjunto 1BPI, usando 40 descriptores y 5x2 CV.	133
C.44. R^2 s de entrenamiento en Experimento 3 para conjunto 1BPI, usando 40 descriptores y Nested CV.	133
C.45.RMSEs de entrenamiento en Experimento 3 para conjunto 1BPI, usan- do 40 descriptores y 5-Fold CV.	134
C.46.RMSEs de entrenamiento en Experimento 3 para conjunto 1BPI, usan- do 40 descriptores y 5x2 CV.	134
C.47.RMSEs de entrenamiento en Experimento 3a para conjunto 1BPI, usando 40 descriptores y Nested CV.	134
C.48. R^2 s de prueba en Experimento 3 para conjunto HLYZ, usando 40 descriptores y 5-Fold CV.	135
C.49. R^2 s de prueba en Experimento 3 para conjunto HLYZ, usando 40 descriptores y 5x2 CV.	135
C.50. R^2 s de prueba en Experimento 3 para conjunto HLYZ, usando 40 descriptores y Nested CV.	135
C.51.RMSEs de prueba en Experimento 3 para conjunto HLYZ, usando 40 descriptores y 5-Fold CV.	136
C.52.RMSEs de prueba en Experimento 3 para conjunto HLYZ, usando 40 descriptores y 5x2 CV.	136
C.53.RMSEs de prueba en Experimento 3 para conjunto HLYZ, usando 40 descriptores y Nested CV.	136
C.54. R^2 s de entrenamiento en Experimento 3 para conjunto HLYZ, usando 40 descriptores y 5-Fold CV.	137
C.55. R^2 s de entrenamiento en Experimento 3 para conjunto HLYZ, usando 40 descriptores y 5x2 CV.	137
C.56. R^2 s de entrenamiento en Experimento 3 para conjunto HLYZ, usando 40 descriptores y Nested CV.	137

C.57.RMSEs de entrenamiento en Experimento 3 para conjunto HLYZ, usando 40 descriptores y 5-Fold CV.	138
C.58.RMSEs de entrenamiento en Experimento 3 para conjunto HLYZ, usando 40 descriptores y 5x2 CV.	138
C.59.RMSEs de entrenamiento en Experimento 3a para conjunto HLYZ, usando 40 descriptores y Nested CV.	138