

# Tabla de Contenido

<b>1. Introducción</b>	<b>1</b>
1.1. Motivación	1
1.2. Objetivos	3
1.3. Estructura del trabajo	4
<b>2. Marco Teórico</b>	<b>5</b>
2.1. Proteínas y Enzimas	5
2.1.1. Estructura primaria, secundaria y terciaria	5
2.1.2. Dominios	7
2.1.3. Números EC	7
2.1.4. UnitProtKB	8
2.2. Búsqueda por Similitud: BLAST	9
2.3. Acercamiento a Aprendizaje de Máquinas	10
2.3.1. Extracción de características	12
2.3.2. Escalar datos	13
2.3.3. Selección de características	13
2.3.4. Clasificadores	14
2.3.5. Selección de modelos	17
2.4. Predicción de funcionalidad de proteínas basados en Machine Learning	19
2.4.1. En la literatura	19
2.4.2. Critical assessment of functional annotation	20
<b>3. Herramientas utilizadas</b>	<b>22</b>
3.1. Lenguaje de programación	22
3.2. CD-HIT	23
3.3. Diamond	23
3.4. InterPro	23
<b>4. Metodología e Implementación</b>	<b>25</b>
4.1. Primeras pruebas	25
4.1.1. Uso de ProFET	26
4.1.2. Uso de InterPro	27
4.2. BLAST	28
4.2.1. Comportamiento en enzimas	28
4.2.2. Comportamiento sobre proteínas	31
4.3. Propuesta	32

<b>5. Resultados</b>	<b>37</b>
5.1. Modelo sin ajuste de parámetros . . . . .	37
5.2. Modelo con ajuste de parámetros . . . . .	38
5.2.1. Caso F-Test . . . . .	38
5.2.2. Caso Información mutua (MI) . . . . .	39
5.2.3. Porcentajes de mejora sobre clases . . . . .	40
5.2.4. Número de vecinos utilizados en KNN . . . . .	40
5.2.5. Número de características seleccionadas . . . . .	41
5.3. Diferencias con IPR . . . . .	43
5.4. Mezcla de características . . . . .	43
5.5. Pruebas a nuevos datos . . . . .	46
5.5.1. Actualizaciones de SwissProt . . . . .	46
<b>6. Discusión</b>	<b>48</b>
6.1. Secuencias no alineadas por BLAST . . . . .	48
6.2. Ventajas de BLAST-KNN en casos multiclase . . . . .	48
6.3. Limitación de SwissProt . . . . .	49
<b>7. Conclusiones y trabajo futuro</b>	<b>51</b>
<b>Bibliografía</b>	<b>53</b>