BioInformática - Projeto Final

Classificação de Síndrome de Down em Camundongos

Integrantes

- Augusto Maillo Queiroga de Figueiredo
- Arthur de Brito Bonifácio
- João Antonio Oliveira Pedrosa

Resumo

Este projeto de Bioinformática aborda a classificação da síndrome de Down, mais especificamente o genótipo Ts65Dn, em camundongos com base nos níveis de expressão de 77 proteínas no córtex cerebral. Utilizando um conjunto de dados que inclui informações de 72 camundongos, divididos em controle e com síndrome de Down, o objetivo é desenvolver um classificador capaz de prever a presença da síndrome com alta precisão.

Introdução

A síndrome de Down é uma condição genética que afeta a expressão de diversas proteínas no organismo. Compreender a relação entre os níveis de expressão proteica e a presença da síndrome em camundongos é fundamental para avanços na área biomédica. Este projeto visa utilizar técnicas de aprendizado de máquina para criar um modelo preditivo robusto com base em dados de expressão proteica.

Metodologia

Estrutura do Dataset

O conjunto de dados utilizado é um arquivo CSV com 77 colunas representando diferentes proteínas. Cada linha corresponde a um camundongo, com informações sobre o genótipo ('Genotype') e os níveis de expressão dessas proteínas. Algumas colunas não relevantes foram removidas para simplificar o conjunto.

Pré-processamento

Antes da modelagem, realizamos as seguintes etapas de pré-processamento:

- Embaralhamento do conjunto de dados.
- Remoção de colunas não utilizadas (MouseID, Treatment, Behavior, class).
- Divisão do conjunto em features (X) e variável alvo (y).
- Preenchimento de valores ausentes utilizando a média como estratégia.

Modelagem e Treinamento

Utilizamos o algoritmo RandomForestClassifier para treinar o modelo, dividindo o conjunto de dados em conjuntos de treinamento e teste. O treinamento foi realizado e as previsões foram feitas no conjunto de teste.

Resultados

Para cada teste, foram realizados 5 diferentes experimentos e apresentamos as média para cada métrica. O conjunto de dados foi embaralhado no início de cada teste. A seguir são apresentados as médias desses resultados obtidos para diferentes tamanhos no conjunto de teste:

Percentual para Teste: 10%

	Geral	Control	Ts65Dn
Accuracy	1.0	-	-
F1	1.0	1.0	1.0
Precisão	1.0	1.0	1.0
Revocação	1.0	1.0	1.0
Número de Amostras	108	56	52

Percentual para Teste: 20%

	Geral	Control	Ts65Dn
Accuracy	0.9954	-	-
F1	0.9954	0.9953	0.9954
Precisão	0.9955	1.0	0.991
Revocação	0.9953	0.9906	1.0
Número de Amostras	216	106	110

Percentual para Teste: 30%

	Geral	Control	Ts65Dn
Accuracy	0.9845	-	-
F1	0.9845	0.9841	0.9849
Precisão	0.9845	0.9810	0.9879
Revocação	0.9846	0.9872	0.9820
Número de Amostras	324	157	167

Percentual para Teste: 50%

	Geral	Control	Ts65Dn
Accuracy	0.9777	-	-
F1	0.9777	0.9778	0.9777
Precisão	0.9779	0.9672	0.9887
Revocação	0.9779	0.9888	0.9669
Número de Amostras	540	268	272

Conclusão

Os resultados demonstram uma alta precisão do modelo na classificação da síndrome de Down em camundongos com base nos níveis de expressão proteica. Este projeto destaca a eficácia do uso de algoritmos de aprendizado de máquina na Bioinformática para resolver problemas relevantes em ciências da vida. A compreensão dessas relações pode ter implicações significativas na pesquisa biomédica, abrindo portas para futuras investigações e desenvolvimentos terapêuticos.