DNA Classification By IBM Granite Model

How Your DNA Can Predict Your Label and Risk of Disease

Get Started

Raw Dataset

This Project use raw synthetic dataset generated by AI. Datase can be acces s by kaggle on this link

https://www.kaggle.com/datasets/miadul/dna-classification-dataset

Project Overview

In this project, the goals is to make IBM Granite Model can classify DNA by the class label such as human, bacteria, virus, and plant, and disease risk by parameters such as kmers-3 freq and mutation status which classify to low, medium, and high potential risk.

From easier way to classify DNA, we can make more fast dan large prediction about DNA theory and relationship between protein component, mutation status, long sequence of DNA with disease risk

Analysis Process

Process start by download dataset and import it to Colab. Then use pandas to read the dataset and cleaning the dataset by removing duplicate dan drop the colomn not use.

After cleaning the dataset, we use it in training model about our dataset using batch size and other feature like Sequence, GC content, AT content, Num A, Num G, Num T, Num C, Mutation flag, and Kmer 3 freq.

Model IBM then classify the dataset by step:

1. Predict the class label such as human, bacterian, virus, and plant
2. Predict the disease risk status such as low, medium, and high
3. Print by format DNA [number]: Predicted Class Label - [Predicted Class], Predicted Disease Risk - [Predicted Risk]

After that, the process continue to fine tuning the parameters and test the model again to check the accuration

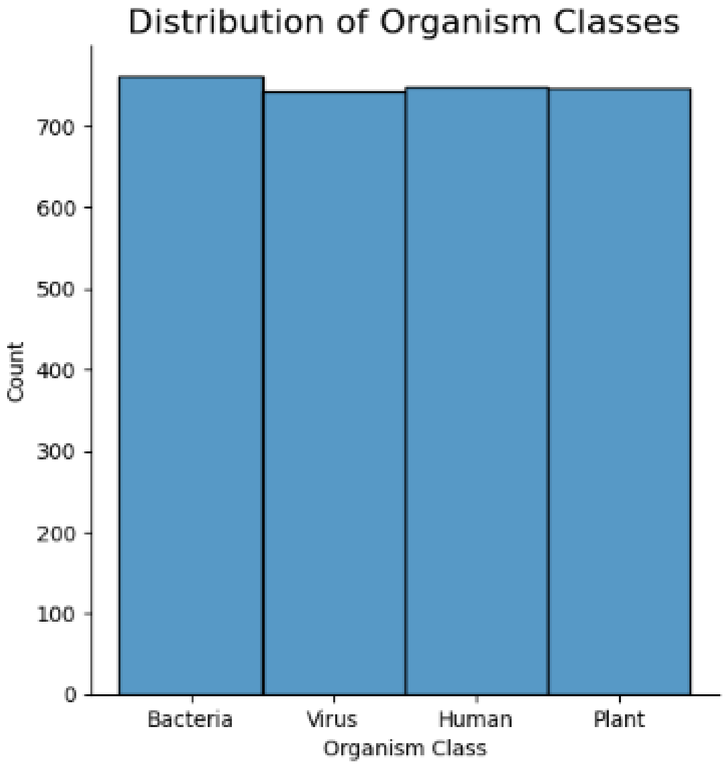
Insight and Findings

The insight is new to me because never try to training and testing with LLM such as IBM Granite Model.

From the analysis, fine tuning parameters give big chance to how model see and understood the dataset and task. But because limited of resource on IBM Granite Model, I can't much fine tuning the parameters. But from the result show with fine tuning the model give response almost completed in first fifty dna data. Which the base parameter give model response unknown. Even though it give not unknown response in fine tuning version, the precision and accuracy still not good and need more improvement

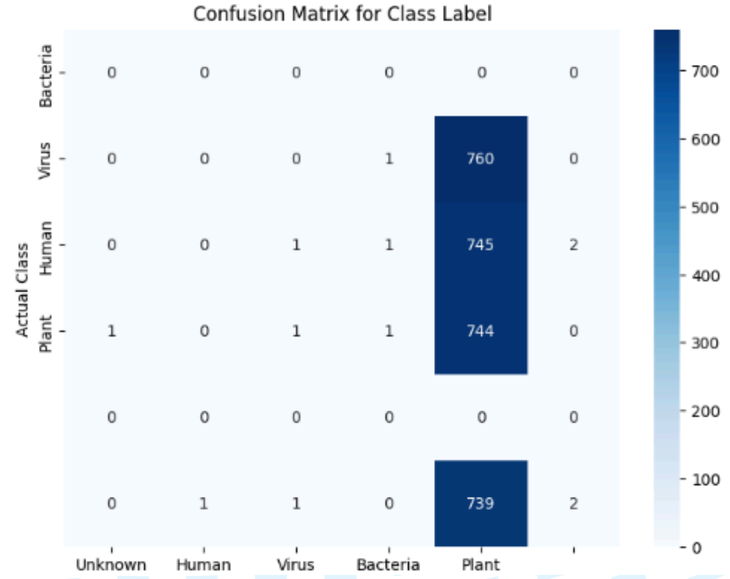
	Sequence	GC_Content	AT_Content	Num_A	Num_T	Num_C	Num_G	kmer_3_freq	Mutation_Flag	Class_Label	Disease_Risk
0	CTTTCGGGATACTTTTGGGATGGTCTTGGTCAAGGGTTTTAGCCCG...	50.0	50.0	22	28	19	31	0.986	0	Bacteria	High
1	TTGACCAAATTTGATTGGAAGTGGTAAGCGCGTATTCCTAGCATCA...	45.0	55.0	27	28	22	23	0.486	1	Virus	Medium
2	GCGTGAGTTCTAATTTAAAAAGTCGTAACACGTACCCCGGCGTGTA...	51.0	49.0	26	23	30	21	0.367	1	Bacteria	Low
3	ACTACGCGGACAAGAACCAACAGAACCTGGTTTTCGCAAGGGAGTG...	55.0	45.0	28	17	23	32	0.404	0	Human	Medium
4	TTCAATGCAGATTGAAAGTTACTTTCATCTGCCCTATGGGTCCCTT...	46.0	54.0	24	30	25	21	0.818	0	Human	High
...	...	...	...	...	...	...	...	...	...	...	...
2995	GATCAGCCCATACACCAAATCAATTGCATACATGTCCGATGTAACA...	46.0	54.0	30	24	27	19	0.786	1	Plant	Medium
2996	TGTTGTGTGTCTGATGATAGGTCATACCGCCTCGAAACATCACCAT...	49.0	51.0	28	23	24	25	0.831	0	Plant	Medium
2997	GACCCACTAAAAGTCTTCGTCTCCTTCCGATGGGAATTTTCGCCGA...	53.0	47.0	21	26	30	23	0.140	0	Virus	Medium
2998	CCAAAGGATATCTGTAATTGTTGCAGCGCCCCTACAATTTGAGCAC...	46.0	54.0	26	28	25	21	0.685	0	Plant	Medium
2999	CCGGATGCCGCTCTATACACCGTCATAGCCTGGAACACACAAATAA...	46.0	54.0	33	21	29	17	0.716	0	Plant	High

3000 rows × 11 columns




```
Accuracy for Class Label prediction: 0.0013
Accuracy for Disease Risk prediction: 0.0003

Confusion Matrix for Class Label:
```



Conclusion & Recommendation

fFom the experiment, IBM Granite Model can classify DNA to various class label and disease risk using feature DNA and mutation flag. With more larger dataset and more comprehensive code, IBM Garnite model can make large move in DNA learning field

AI support explanation

In this project, AI or Artificial Intelligence very impactful to make the code easier and made decision such as predict and classify faster but relliable. With AI, study will become more complex and deep but with easy way to make benefit in humans life

Thank You

By Santia