Sequence Alignment using ML

Jyoti Kataria
Bioinformatics Analyst, Medgenome Labs Ltd.
M.Sc. Biotechnology (IIT Bombay)

Lets recap!

DNA & RNA

Central Dogma

Gene

Introns & Exons

Dna replication

Transcription

Translation

Pairwise Alignment

MSA

Indels

SNPs

Codons

Databases

Announcements

ML Assignments:

Sep 3rd: Allotment

Sep 24th: Submission (Group wise)

Internal Assessment:

Full marks (if class participation is there)

ML Techniques

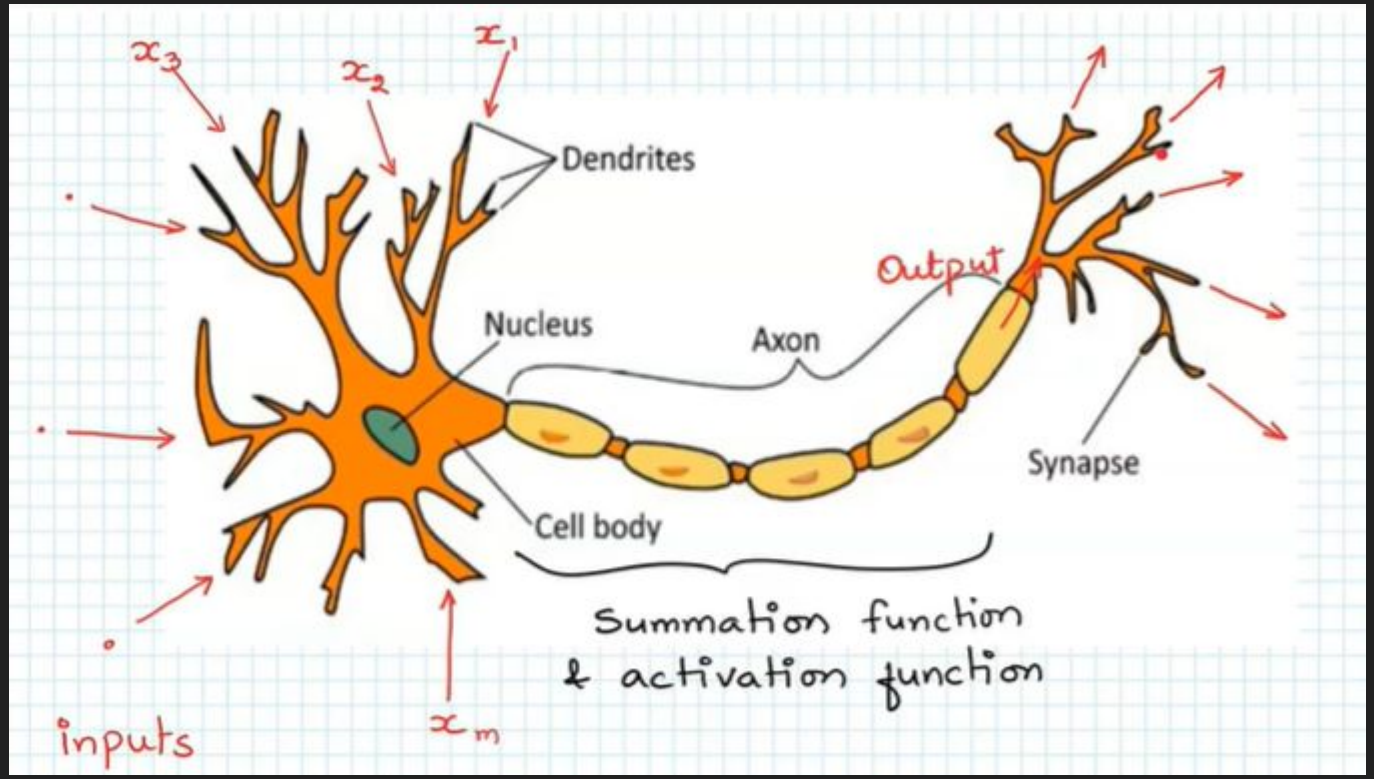
- 15 techniques for predicting optimal sequence alignment
- Based on NW and SW algorithm
- Ex: Multilayer perceptron, XG Boost, Support Vector Machine

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9497110

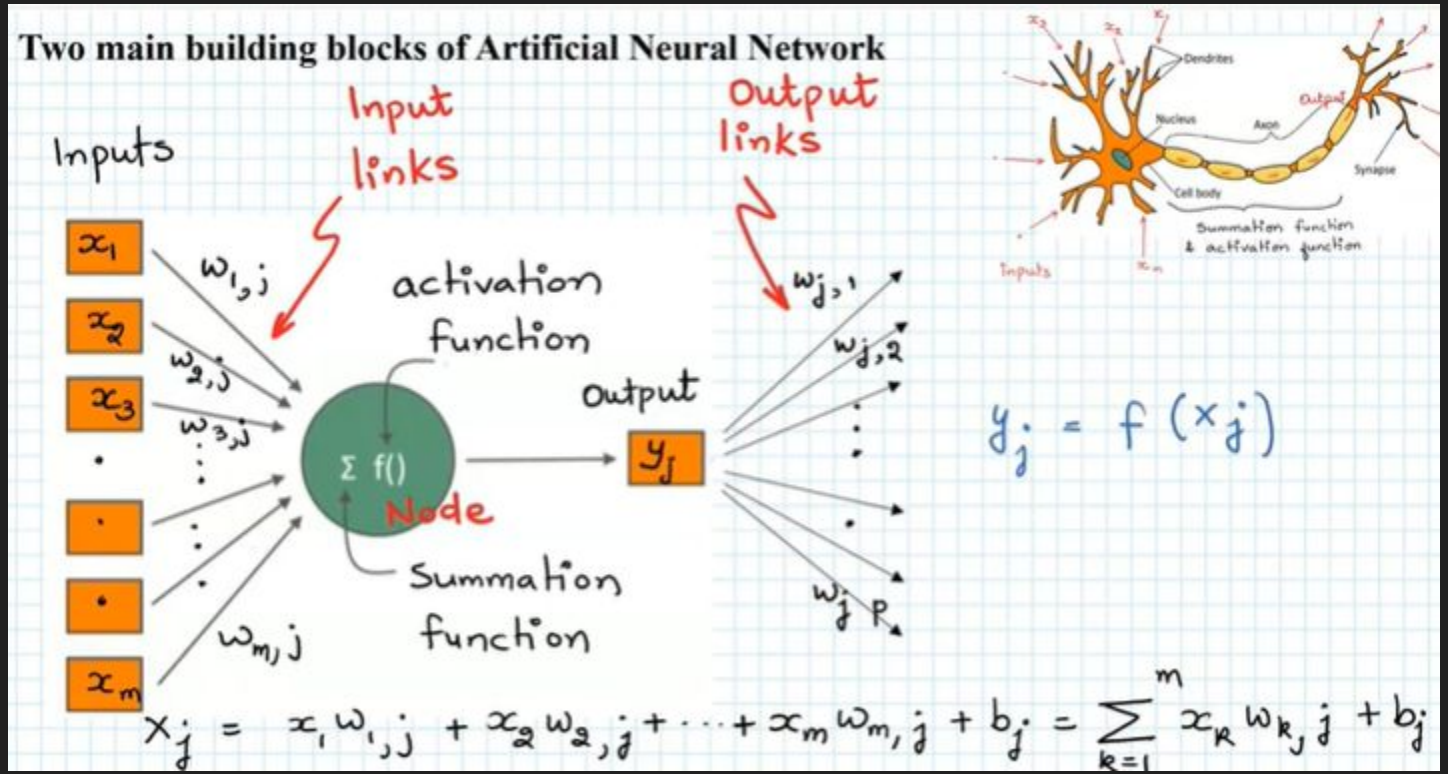
MLP
Multi-layer perceptron

- Type of ANN
-

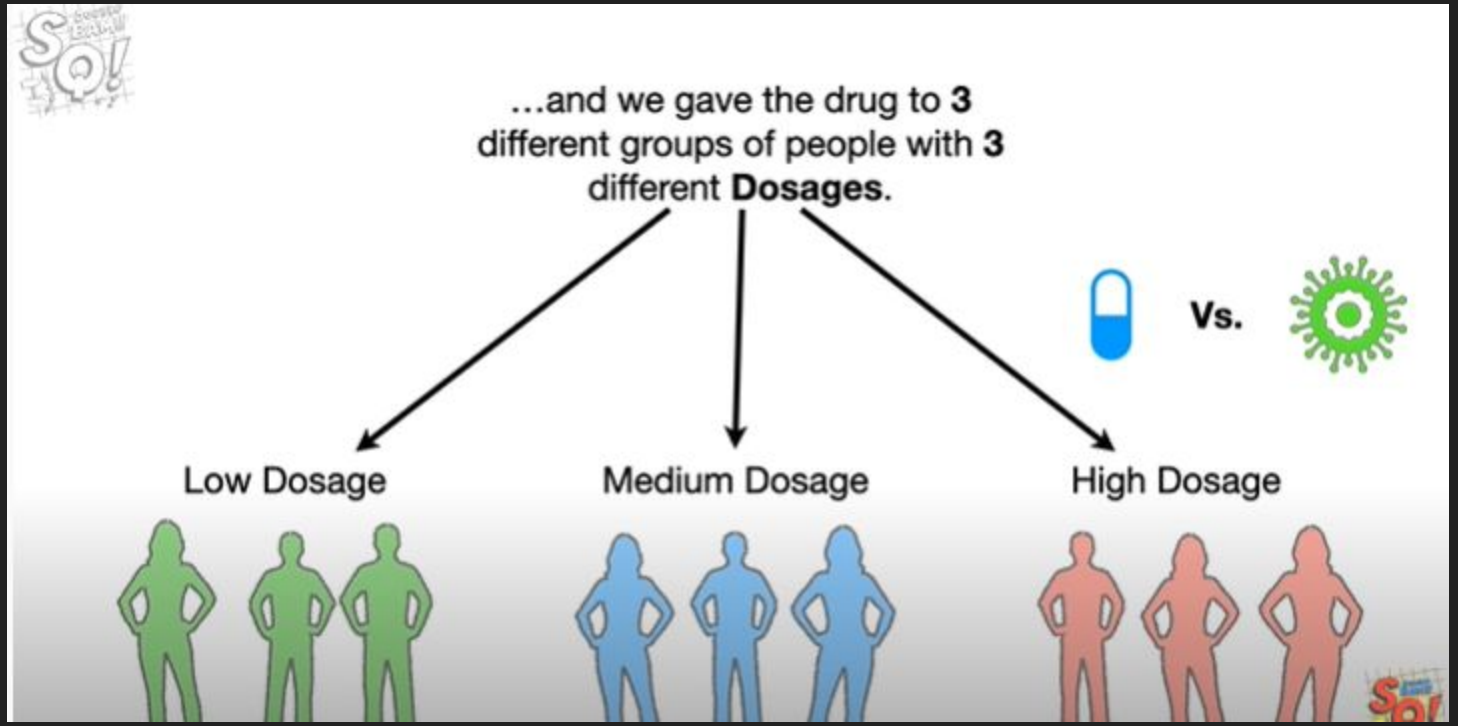
ANN


x_3
x_2
x_1
Dendrites
Output
Nucleus
Axon
Synapse
Cell body
Summation function
& activation function
inputs
x_m

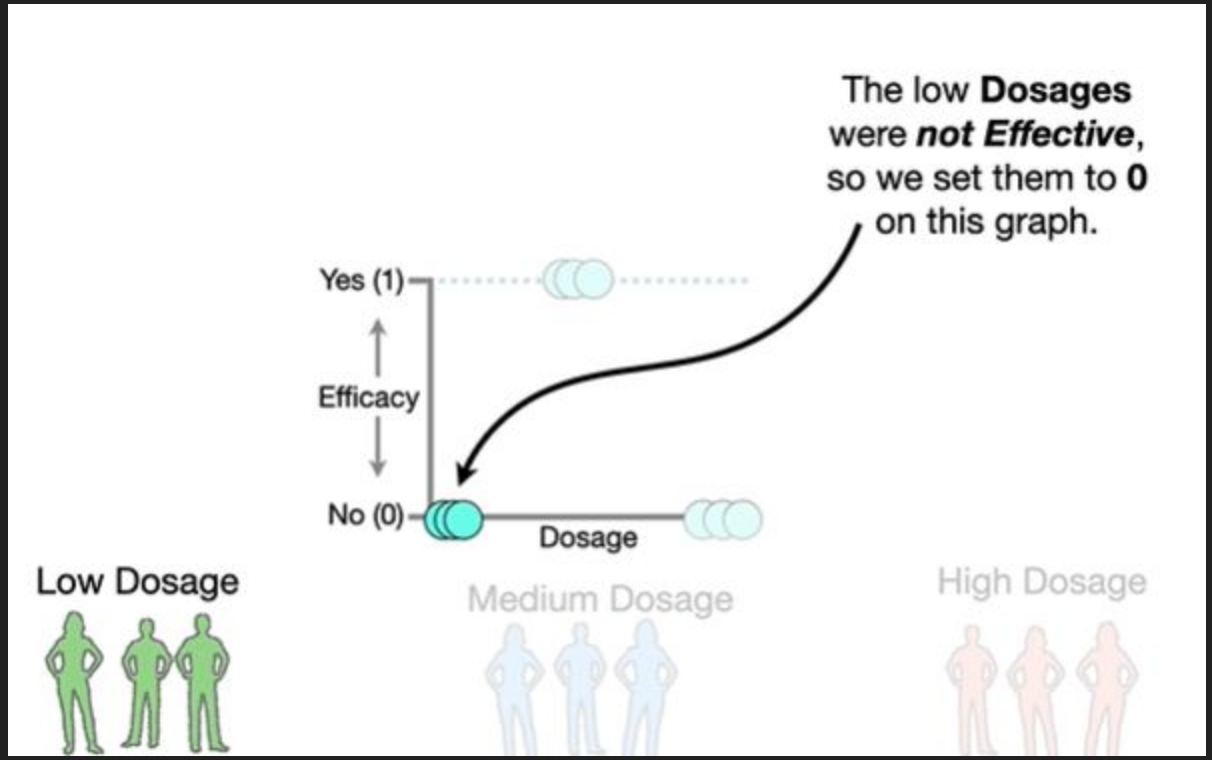
ANN (Contd..)


Two main building blocks of Artificial Neural Network
Inputs
Input links
Output links
activation function
output
Σ f()
Node
summation function
Dendrites
Nucleus
Axon
Synapse
Cell body
Summation function & activation function
$y_j = f(x_j)$
$x_j = x_1 w_{1,j} + x_2 w_{2,j} + \cdots + x_m w_{m,j} + b_j = \sum_{k=1}^{m} x_k w_{k,j} + b_j$

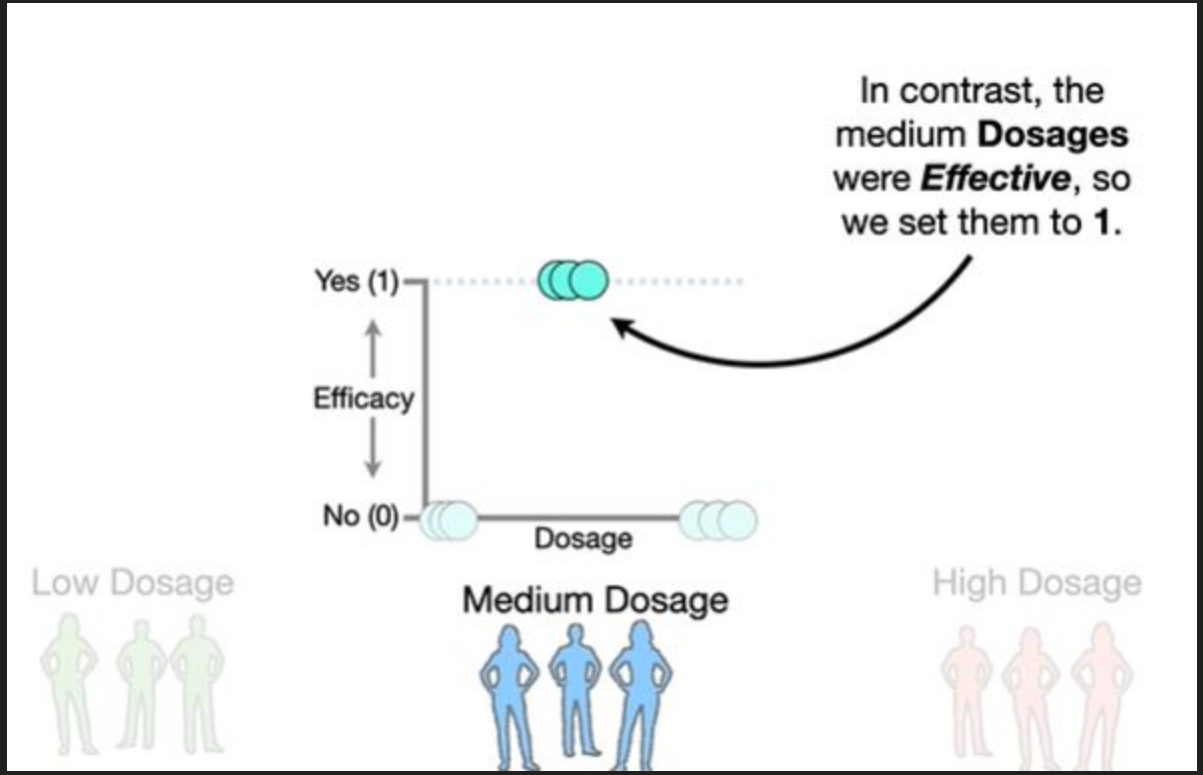
ANN (Contd..)


...and we gave the drug to 3 different groups of people with 3 different Dosages.
Vs.
Low Dosage
Medium Dosage
High Dosage

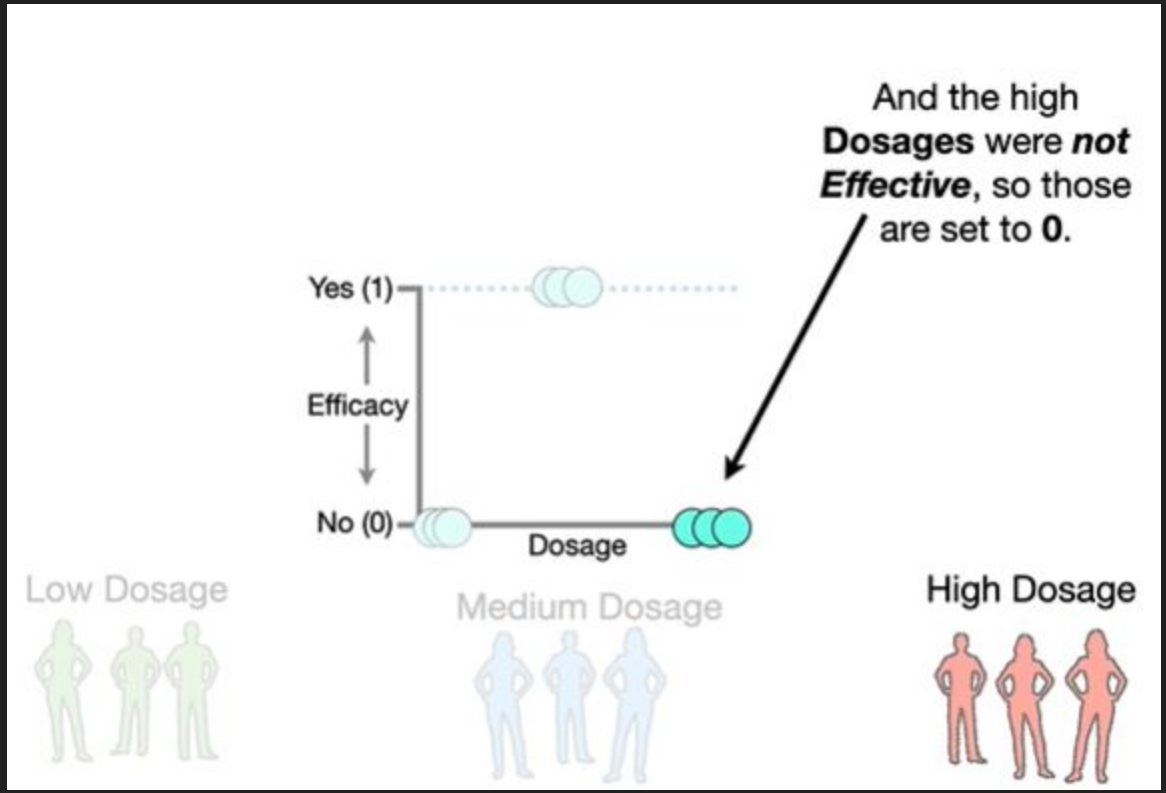
ANN (Contd..)


The low **Dosages** were ***not Effective***, so we set them to **0** on this graph.
Yes (1)
Efficacy
No (0)
Dosage
Low Dosage
Medium Dosage
High Dosage

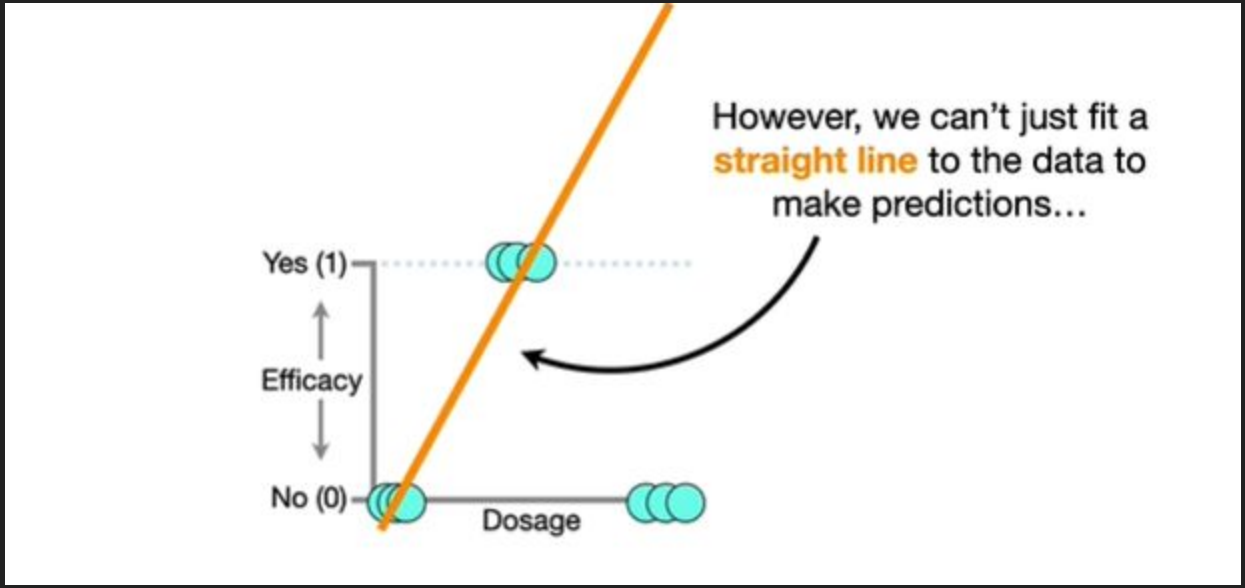
ANN (Contd..)


In contrast, the medium **Dosages** were ***Effective***, so we set them to **1**.
Yes (1)
Efficacy
No (0)
Dosage
Low Dosage
Medium Dosage
High Dosage

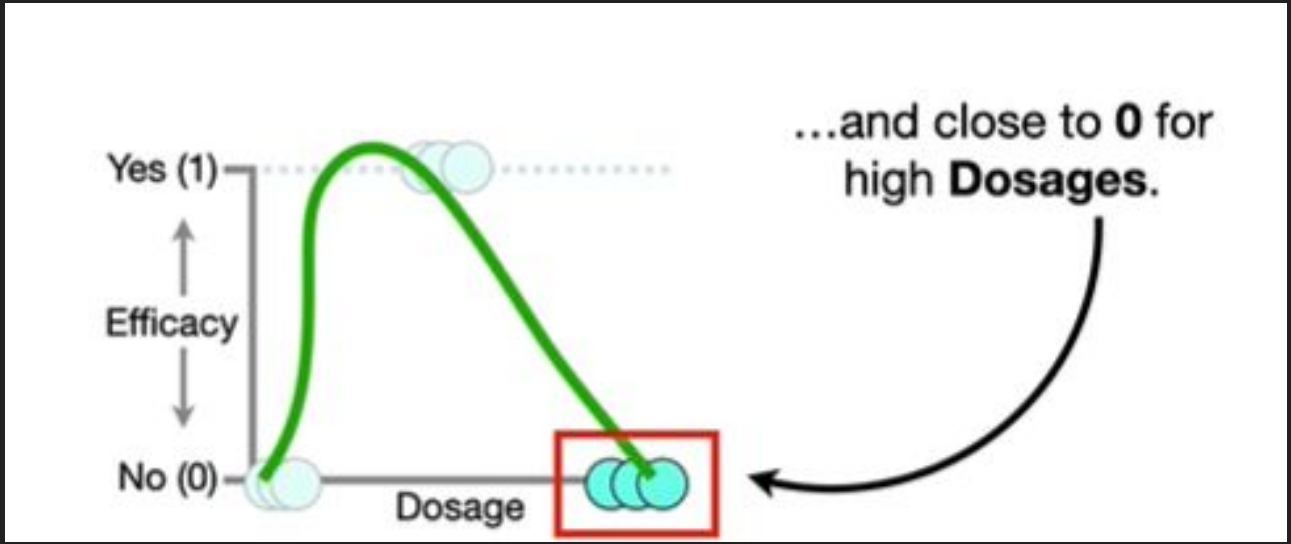
ANN (Contd..)


And the high
Dosages were not
Effective, so those
are set to 0.
Yes (1)
Efficacy
No (0)
Dosage
Low Dosage
Medium Dosage
High Dosage

ANN (Contd..)


However, we can't just fit a straight line to the data to make predictions...
Yes (1)
Efficacy
No (0)
Dosage

ANN (Contd..)



XGBoost

- Extreme Gradient Boosting
- Used for supervised learning problems where we use the training data with multiple features(xi) to predict a target variable (yi)
- Low accuracy classifiers > high accuracy classifiers
- Decision Tree based ensemble machine learning algorithm
- Faster than the traditional ML methods