Assignment 4

In this assignment, we are asked to analyze a dataset from Kaggle and make inferences. The dataset contains various parameters relating to a student, namely the GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA and Research, and we were asked to develop a function of these which is able to predict the Chance of Admit. The thought process followed is outlined below:

Preprocessing

- Each set of values are normalized to bring them between 0 and 1.
- The percentage error is calculated by the calculating the average of (*y_pred* - *y_actual*) / *y_actual* * 100 for each datapoint.

Approach 1 - Linear Model

- In this approach, we assume the output to be a linear function of the input.
- We pass the coefficients of our variables to the `curve_fit` function from the `scipy.optimize` library.
- From an analysis of the optimized parameters, we can make the following conclusions about parameter dependence:
  - The linear function depends the most on `CGPA`, with a coefficient value of 1.2.
  - The next highest dependency is on `TOEFL` and `GRE` score, with parameter values of 0.64 and 0.34.
  - There is a very less dependency on `University Rating`
  - The percentage error by this method is 6.869%
- The constant and the coefficients of the GRE Score, TOEFL Score, University Rating, SOP, LOR ,CGPA, Research are -1.31312734e+00, 6.44957000e-01, 3.47707313e-01, 4.44143771e-09, 2.10539024e-02, 8.88706993e-02, 1.20549029e+00 and 2.48789408e-02 respectively.

Approach 2 - Non Linear Model

- In this approach, we assume that the output varies linearly with powers of the input parameters.
- We again use the `curve_fit` function for the same.
- Here, we need to keep in mind the fact that each parameter lies between 0 and 1, so a high value of power results in a lower power and thus a low dependence.
- Again, an analysis of the optimized parameters give us the following conclusions:
  - The function has a very high dependence on the `University Rating` parameter.
  - The function also depends on the `TOEFL` and the `GRE` score, as well as on `CGPA`.

    - This method gives a percentage error of 6.840%

- The constant and the coefficients of the GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research are -6.43996514e+00, 1.32694451e+00, 1.53674603e+00, 9.39165546e-08, 2.37966653e-02, 8.93684492e-02, 4.45897884e+00 and 2.56471247e-02.
- The powers of GRE Score, TOEFL Score, University Rating, SOP, LOR and CGPA are 4.67000243e-01, 2.16770765e-01, 7.18507274e-35, 1.00257233e+01, 9.51419394e-01 and 2.36193687e-01.

We explored other models, such as exponential and tanh models, but the results we obtained came out to be poorer. Another approach to the problem was to split the dataset into two portions, one with the points containing the `Research` parameter as 0 and the other containing the `Research` parameter as 1, and performing `curve_fit` on the two separate parts, which provided approximately similar percentage error.

In an ideal case, the relationship between the input and output parameters would be very complicated, with a vast interconnection between parameters. This relationship can be captured well by a neural network, a more preferable solution to this problem.