Problem statement

A new pharmaceutical startup is recently acquired by one of the world's largest MNCs. For the acquisition process, the startup is required to tabulate all drugs that they have sold and account for each drug's effectiveness. A dedicated team has been assigned the task to analyze all the data. This data has been collected over the years and it contains data points such as the drug's name, reviews by customers, popularity and use cases of the drug, and so on. Members of this team are by the noise present in the data.

Your task is to make a sophisticated NLP-based Machine Learning model that has the mentioned features as the input. Also, use the input to predict the base score of a certain drug in a provided case.

Data

The dataset has the following columns:

Variable Name	**Description**
patient_id	ID of patients
name_of_drug	Name of the drug prescribed
use_case_for_drug	Purpose of the drug
review_by_patient	Review by patient
drug_approved_by_UIC	Date of approval of the drug by UIC
number_of_times_prescribed	Number of times the drug is prescribed
effectiveness_rating	Effectiveness of drug
base_score	Generated score (Target Variable)

Data description

The data folder consists of the following two **.csv** files:

- **train.csv** - (32165x 7)
- **test.csv** - (10760x6)

The **sample_submission** is described as follows:

```
patient_id,base_score
206461,9.05
95260,8.85
```

92703,5.26
138000,8.03