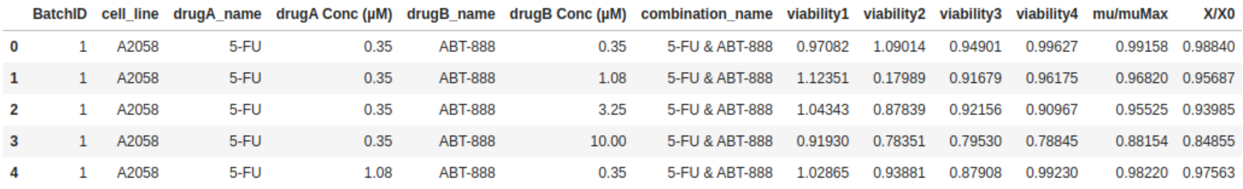
# Independent Project Report

**Title:** Integration and analysis of cancer drug response data

**Submitted By:**
Sehban Fazili (MT21143)
Mohammad Osama Ataullah (MT21127)
Sayan Mitra (MT21142)
Shoumik Bhattacharya (MT21144)

# 1. Dataset

The dataset is **CombinationalDrugDataResponse** which is taken from the journal https://aacrjournals.org/mct/article/15/6/1155/92159/An-Unbiased-Oncology-Compound-Screen-to-Identify . The dimensions of the dataset is (368832 X 13). It gives information on the different drug combinations that were tested on different cancer cell lines. Figure-1 shows the first five rows of the dataset.



| | BatchID | cell_line | drugA_name | drugA Conc (μM) | drugB_name | drugB Conc (μM) | combination_name | viability1 | viability2 | viability3 | viability4 | mu/muMax | X/X0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | A2058 | 5-FU | 0.35 | ABT-888 | 0.35 | 5-FU & ABT-888 | 0.97082 | 1.09014 | 0.94901 | 0.99627 | 0.99158 | 0.98840 |
| **1** | 1 | A2058 | 5-FU | 0.35 | ABT-888 | 1.08 | 5-FU & ABT-888 | 1.12351 | 0.17989 | 0.91679 | 0.96175 | 0.96820 | 0.95687 |
| **2** | 1 | A2058 | 5-FU | 0.35 | ABT-888 | 3.25 | 5-FU & ABT-888 | 1.04343 | 0.87839 | 0.92156 | 0.90967 | 0.95525 | 0.93985 |
| **3** | 1 | A2058 | 5-FU | 0.35 | ABT-888 | 10.00 | 5-FU & ABT-888 | 0.91930 | 0.78351 | 0.79530 | 0.78845 | 0.88154 | 0.84855 |
| **4** | 1 | A2058 | 5-FU | 1.08 | ABT-888 | 0.35 | 5-FU & ABT-888 | 1.02865 | 0.93881 | 0.87908 | 0.99230 | 0.98220 | 0.97563 |

Figure -1

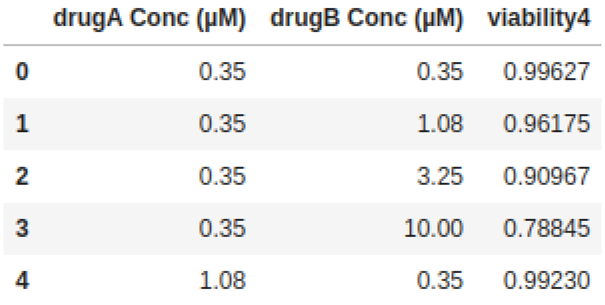
# 2. Data Preprocessing and Modeling

## 2.1. Feature Selection

The dataset only contained three main features that could be used for the regression task. The features selected and the relevant information about them are as follows:

- drugA Conc(uM): Drug A concentration used at micro molar concentration.
- drugB Conc(uM): Drug B concentration used at micro molar concentration.
- viability4: The combined effectiveness of the combination.

Figure -2 represents the dataset after feature selection of the first 5 rows.

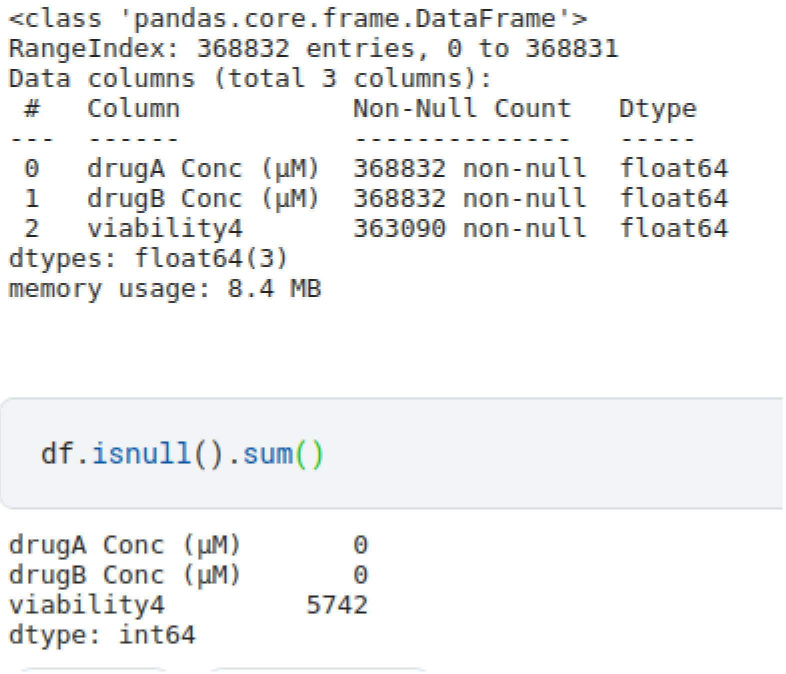


| | drugA Conc (μM) | drugB Conc (μM) | viability4 |
|---|---|---|---|
| **0** | 0.35 | 0.35 | 0.99627 |
| **1** | 0.35 | 1.08 | 0.96175 |
| **2** | 0.35 | 3.25 | 0.90967 |
| **3** | 0.35 | 10.00 | 0.78845 |
| **4** | 1.08 | 0.35 | 0.99230 |

Figure -2

## 2.2. Substituting Missing/NULL values with Median

During preprocessing we found NULL values only in the column[‘viability4’] which had 5742 null values which were replaced with median. The figure below gives us information about the same.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 368832 entries, 0 to 368831
Data columns (total 3 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   drugA Conc (µM)  368832 non-null  float64
 1   drugB Conc (µM)  368832 non-null  float64
 2   viability4       363090 non-null  float64
dtypes: float64(3)
memory usage: 8.4 MB
```

```
df.isnull().sum()
```

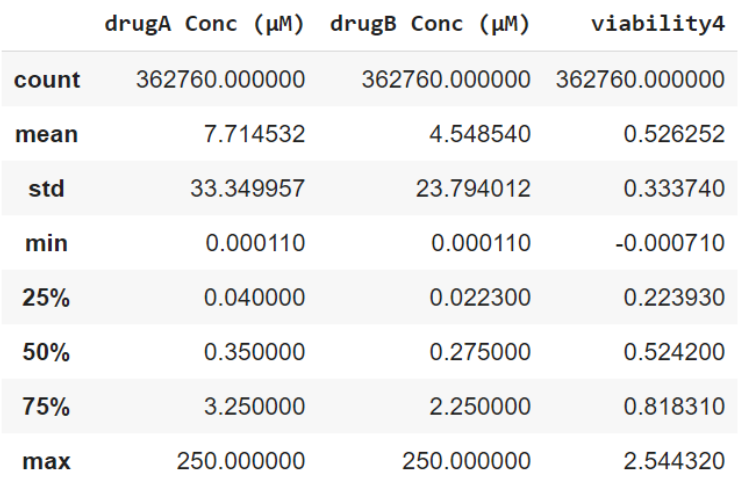
```
drugA Conc (µM)       0
drugB Conc (µM)       0
viability4         5742
dtype: int64
```

Figure-3

## 2.3. Removing Duplicate values:

A total of 6072 duplicate values were found and then removed.

Preprocessed Dataset Statistics



| | drugA Conc (µM) | drugB Conc (µM) | viability4 |
|---|---|---|---|
| **count** | 362760.000000 | 362760.000000 | 362760.000000 |
| **mean** | 7.714532 | 4.548540 | 0.526252 |
| **std** | 33.349957 | 23.794012 | 0.333740 |
| **min** | 0.000110 | 0.000110 | -0.000710 |
| **25%** | 0.040000 | 0.022300 | 0.223930 |
| **50%** | 0.350000 | 0.275000 | 0.524200 |
| **75%** | 3.250000 | 2.250000 | 0.818310 |
| **max** | 250.000000 | 250.000000 | 2.544320 |

### 2.4. Normalization:

First te dataset was divided features(X) and labels(y):
Features = 2 (drugA Conc (µM), drugB Conc (µM))
Features were then normalized using the Standard Scalar normalization technique.
Labels = 1 (viability4)

### 2.5. Splitting the data frames into train and test sets:

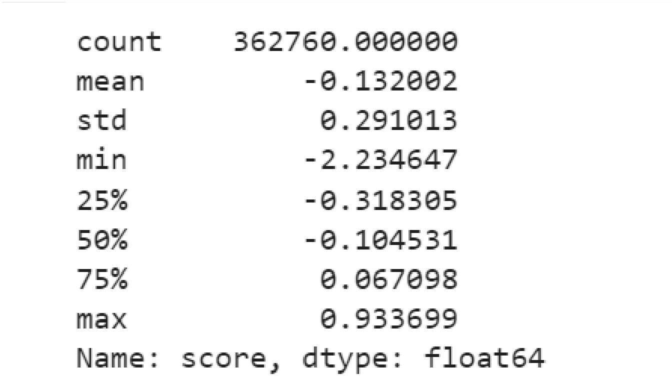
We splitted the dataset in the ratio of 7:3 i.e., 70% was used for training and 30% was used for testing.

## 3. Results

| Model | r2 score | MSE | MAE |
|---|---|---|---|
| Symbolic Regressor | 0.021 | 0.113 | 0.291 |
| Random Forest Regressor | 0.491 | 0.057 | 0.188 |
| CatBoost Regressor | 0.480 | 0.057 | 0.188 |
| Gradient Boosting Regressor | 0.468 | 0.060 | 0.196 |

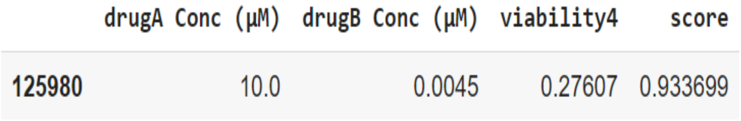
Note: These are the best values found after hyperparameter tuning.

- **Bliss Score Statistics:**



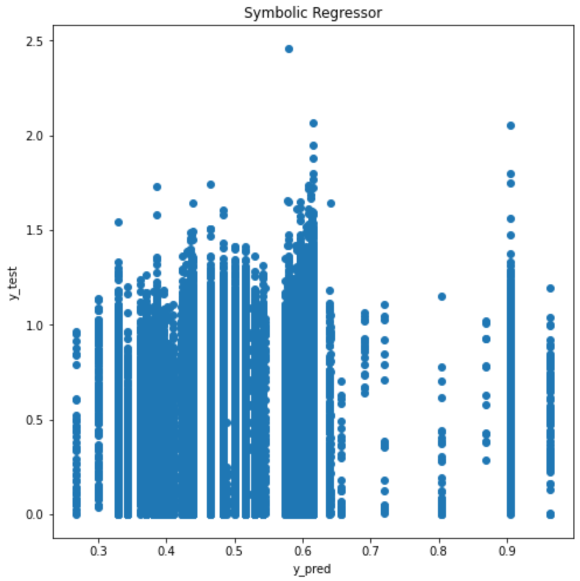
```
count    362760.000000
mean         -0.132002
std           0.291013
min          -2.234647
25%          -0.318305
50%          -0.104531
75%           0.067098
max           0.933699
Name: score, dtype: float64
```

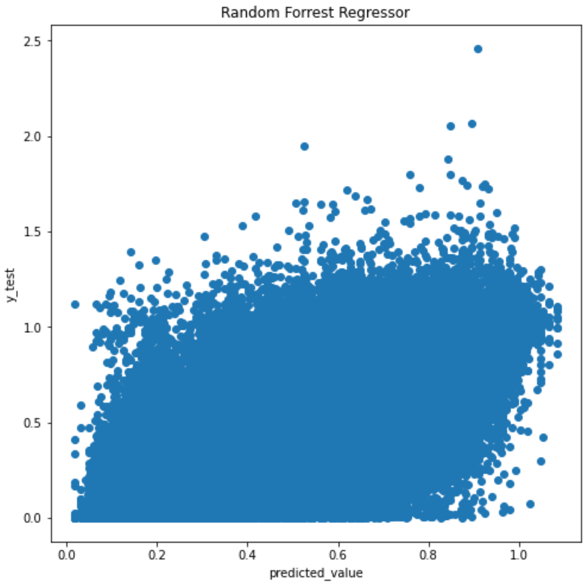
**<u>Note-</u>**For the Maximum Bliss Score which is coming as 0.933699 these are the following drugA and drugB concentrations as shown in figure-

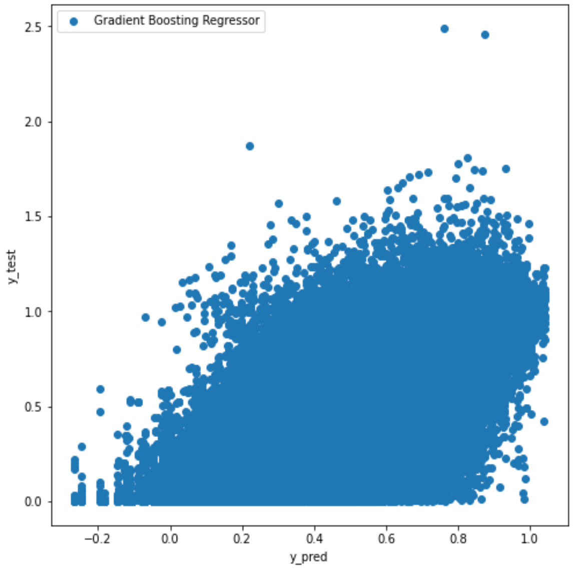


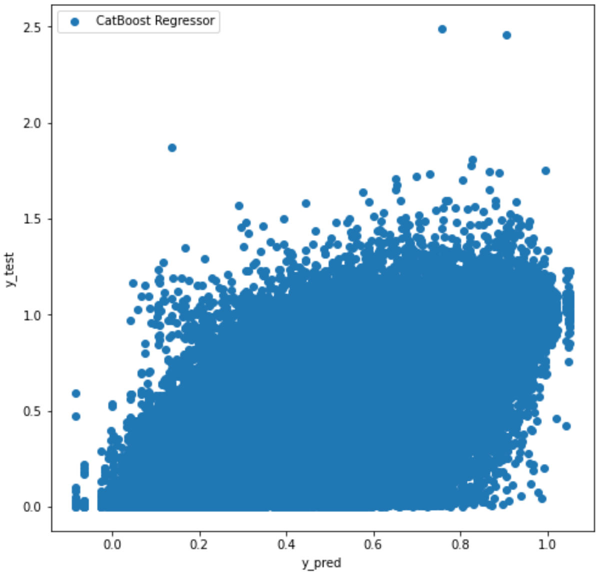
| | drugA Conc (µM) | drugB Conc (µM) | viability4 | score |
|---|---|---|---|---|
| **125980** | 10.0 | 0.0045 | 0.27607 | 0.933699 |

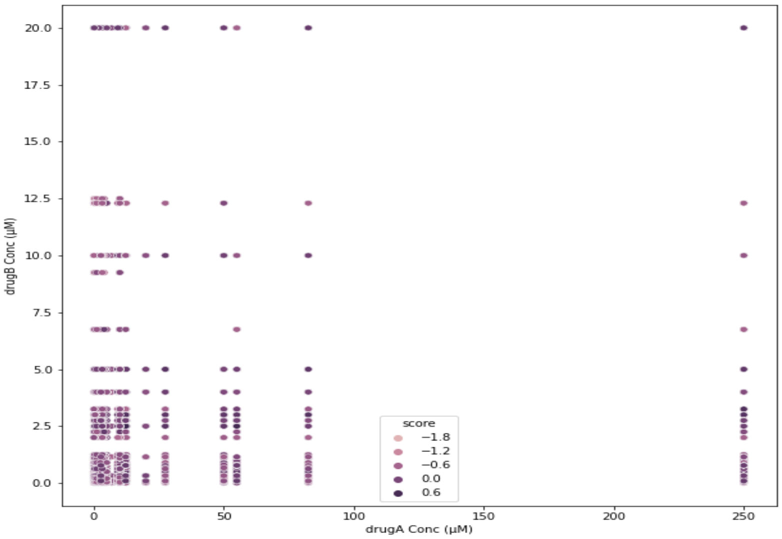
## 4. Scatter Plots:










20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
drugB Conc (μM)
0
50
100
150
200
250
drugA Conc (μM)
score
−1.8
−1.2
−0.6
0.0
0.6