Divyang Dangi

Prediction of MEchanism of action of a drug

# Table of Contents

Abstract2

Literature Review3

Dataset4

Approach6

Exploratory Data Analysis8

Data Engineering16

Results21

Conclusion25

References26

# Abstract

In terms of pharmacology, Mechanism of Action (MoA) refers to the specific biochemical interaction through which a drug substance produces its effect.  Mechanism of action of drug is most important thing which needs to be identified for efficient use of that drug. Knowing mechanism of action of a drug would help scientist identify and modify a protein target associated with the diseases. Gene expression and cell viability pattern plays a major role in determining target protein for a drug and by using the same information from a genomic database we will predict MoA of a given drug. The dataset “Mechanism of action (MoA) prediction data set”[[1]](#footnote-1), taken from Kaggle, gives us exposure to large genomic dataset with gene expression and cell viability patters of drug with known mechanism of actions. The dataset will allow us to determine effect of gene and cell viability data on different target sites and perform bivariant classification using Logistic regression, KNN and Neural Network. The model with higher F1 score will be chosen to perform prediction of MoA of drugs. The above analysis is planned to be achieved by using Python and tools such as Numpy, Pandas, Seaborn, Matplotlib, Tensorflow and Scikit learn.

# Literature Review

In the research paper for “Deep learning applications for predicting pharmacological properties of drug and drug repurposing using transcriptomatic data”2, the researcher had successfully demonstrated use of deep neural network model (DNN) which was trained on large genetic activity of different cell line dataset was able to classify therapeutic category of drug molecule. The research done in this paper would help largely in drug discovery and development.

The article “Applications of machine learning in drug discovery and development”3 gives us a nice bifurcations of different Machine learning process suitable to solve and answer questions related to different drug development stages and clinical trials. The drawback of the research is lack of interpretability and repeatability of ML generated results and lack of high dimension data.

The review article “Bayer’s *in silico* ADMET platform: A journey of machine learning over the past two decades”4 gives us a in depth view of a platform named ADMET (absorption, distribution, metabolism and excretion) created by Bayer pharma to produce models of different physical, chemical properties and movement of drugs within the body. Company uses its own high quality tailored data to obtain experimental endpoint using deep neural network.

# Dataset

The dataset “Mechanism of action (MoA) prediction “1 taken from Kaggle is already been divided into test data and train data. The train dataset consists of 23,814 rows and 876 features for each row. Each row in training dataset is a unique molecule of drug from known genomic database named as sig\_id.

In training dataset, we could see three categorical features namely “cp\_type”, “cp\_time” and “cp\_dose”. There are 772 gene expression columns and each drug molecule gives a specific gene expression for different gene. These gene expression is denoted by ‘g- ‘. The dataset also has 100 Cell viability column. Cell viability is an assay performed to measure cell survival following treatment with drug molecules and is denoted by ‘c- ‘. Training data also has 206 scored target columns which gives us information about target protein molecule in relation to MoA of a drug molecule.

Test dataset consist 3983 rows and have similar features of “cp\_type”, “cp\_time”, “cp\_ dose”,” g- “, “c- “as in training data set. Test dataset does not have scored target columns of protein molecule as they need to be predicted.

**Descriptive statistics:**

Both Train and Test dataset has three categorical features of ‘cp\_type’, ‘cp\_time’, ‘cp\_dose’.

**Cp\_type:** Cp\_type is categorized into treatment i.e cp\_vehicle and controlled i.e ctrl\_vehicle drug molecules. The ctrl\_vehicle is a placebo and does not have any MoA, thus it’s going to be removed from both test and train data and take cp\_vehicle only for our analysis.

**Cp\_time:** It represents the time elapsed between the doses given. The drug was given three times, in time span of 24 hrs,48 hrs. and 72 hrs.

**Cp\_dose**: It represent the level of dose used in the experiment i.e high dose and low dose.

**g-:** Gene expression value is the amount of protein expressed when drug interacts with that particular gene. The dataset has pool of 772 gene expression for every drug molecule in both test and train dataset and is of numeric data type containing both negative and positive values.

**C-:** Cell viability assay is the measure of overall health of viable living cells after treatment with drug in the experiment. Dataset consist of 100 cell lines indicated that the value for cell viability in both test and train data and is of numeric data type containing both negative and positive values.

**Score Target features:** These are the different target protein molecules through which the drug produces its intended effect. There are 206 sites of action and each drug can have more than one target sites through which it produces its intended effect. These features are present in train data only since, it’s our dependent variable and we will be predicting target features for the drugs in test dataset.

**Table 1: Features and its descriptive statistics**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Descriptive Statistics** |
| Cp\_type | Categorical | Two Categories: cp\_vehicle and ctrl\_vehicle |
| Cp\_time | Categorical | Three categories: 24hrs, 48 hrs and 72 hrs |
| Cp\_dose | Categorical | Two categories: D1(High dose) and D2(Low dose) |
| Gene (g-) | Numeric | 772 gene expressions data for each drug |
| Cell viability(c-) | Numeric | 100 cell lines data for each drug |
| Scored Target Features | Multi-Labelled Classification converted to Binary classification | 206 sites of actions converted to action or no action. |

# Approach

The approach to predict MoA of unknown drug molecule is represented as follow:

Exploratory Data Analysis

Pre processing data for analysis

(Multi Labelled to Binary)

Perform SMOTE, Feature selection and PCA

PCA

Perform prediction to determine drug action.

**Step 1: Exploratory Data Analysis**

A detailed Exploratory data analysis will be performed on test data, train data and scored target data. Distribution of the features namely “cp\_type”, “cp\_time” “cp\_dose”,”g-“, “c-“ will be plotted for both test and train data. We would also see the effect of “cp\_time” and “cp\_dose” on gene data and cell viability data along with correlation within the features of train data. We will also have a look at distribution of scored target data and correlation between target data and train data features.

**Step 2: Preprocess data for analysis**

Before performing analysis of a dataset, it needs to be processed to make it more accurate and legible. Features namely “cp\_type”, “cp\_time” and “cp\_dose” needs to be converted into factors or categorical data types for both test and train data. “ctr\_vehicle” which is a placebo, comes under “cp\_type” feature will be removed from both test and train data as it does not give any information on MoA of drug. Test features distribution shows dataset with multi-labelled classification and it will be converted to bivariant classification for further analysis.

**Step 3: Perform SMOTE, PCA and Feature selection**

Target features after conversion to bivariant data seems quite imbalanced, so to counter the challenge SMOTE will be used to perform oversampling and balance the data. Since we have 876 features for each attribute it is difficult to perform analysis on such large data. To mitigate that challenge, Feature selection will be done to select all important features for our prediction and PCA will be done to reduce dimensionality of data. Both dataset one with PCA and other without PCA will be used for comparison purpose.

**Step 4: Train and evaluate model for prediction**

I will be making four different models for analysis namely Normal data with PCA, normal data with feature selection, SMOT data with PCA and SMOT data with feature selection. As we already have our dataset divided into test and train data, we will be using train data to train and validate our model. Use of algorithm namely Logistic regression, KNN classification and Neural Network will be used to train and validate results. Cross validation will be done on train data and model performance will be evaluated. Model with higher F1 rate will be chosen for prediction.

**Step 5: Performing prediction to determine MoA of unknown drug**

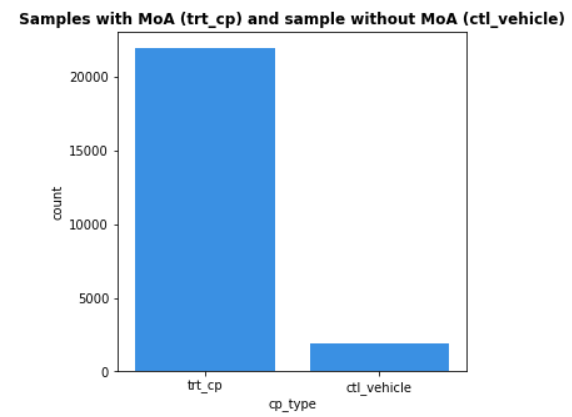
Model with higher accuracy, precision, recall and F1 score will be selected to perform prediction of test data on drugs with unknown MoA.

# Exploratory Data Analysis

Mechanism of action dataset has two files in it, namely MoA dataset (dependent variable) and Target dataset (independent variable). MoA dataset consist of 3 categorical variables, namely cp\_type, cp\_time and cp\_dose and 2 numeric variables gene data readings (g-) and cell viability readings (c-). Target data is classified into multi-labelled data and will be converted to binary data for our analysis. Following is the descriptive statistics of both MoA dataset and Target dataset.

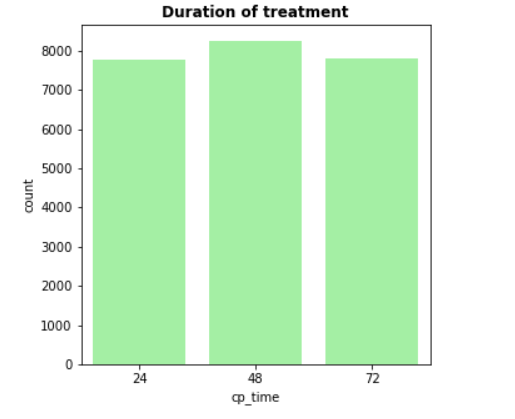
**Cp\_type**: Cp\_type is one of the categorical variables of MoA dataset. It has trt\_cp which represents drug which shows mechanism of action and ctl\_vehicle which does not have any mechanism of action (placebo). The graph, figure 1 below suggest that the data is largely imbalanced but since ctl\_vehicle is (placebo) and is of no importance to us, it will be avoided in further analysis.

Figure 1: Plot for cp\_type for MoA data.



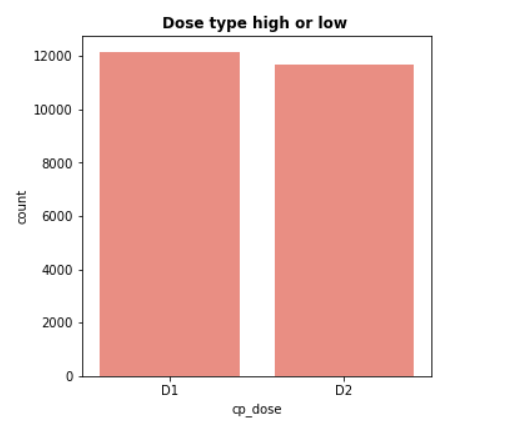
**Cp\_time:** Cp\_time categorized into three types namely 24hr, 48hrs and 72 hrs. an seems to be quite balanced for all categories as per figure 2.

Figure 2: Plot for cp\_time for MoA data.



**Cp\_dose**: Cp\_dose is categorized into two categories of high dose (D1) and low dose (D2). The data for cp\_dose seems to be well balanced for both categories as per figure 3.

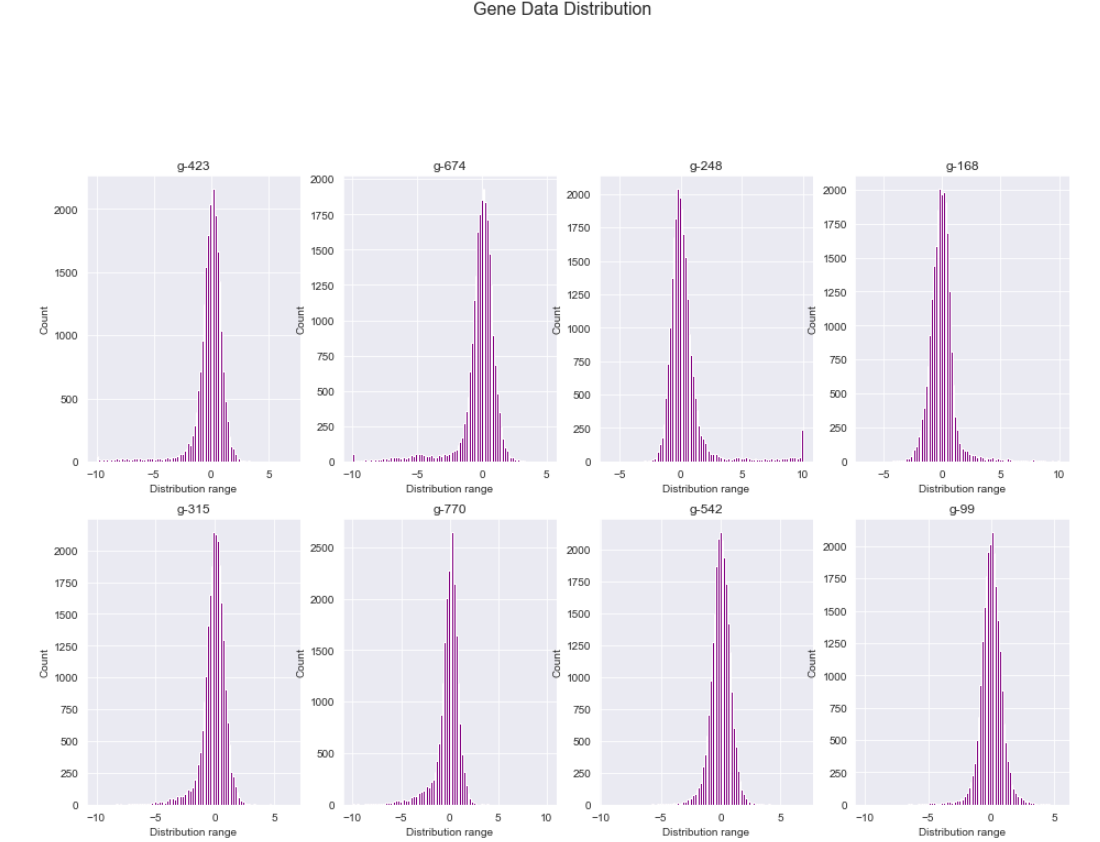
Figure 3: Plot for cp\_dose for MoA data.



## **Gene data:**

Gene data is continuous numeric data and has readings for 772 different gene responses for a particular drug. As we cannot have a look at response of all gene expression, a random sample was taken and distribution range was observed within -10 to +10 range which can be seen in figure 4.

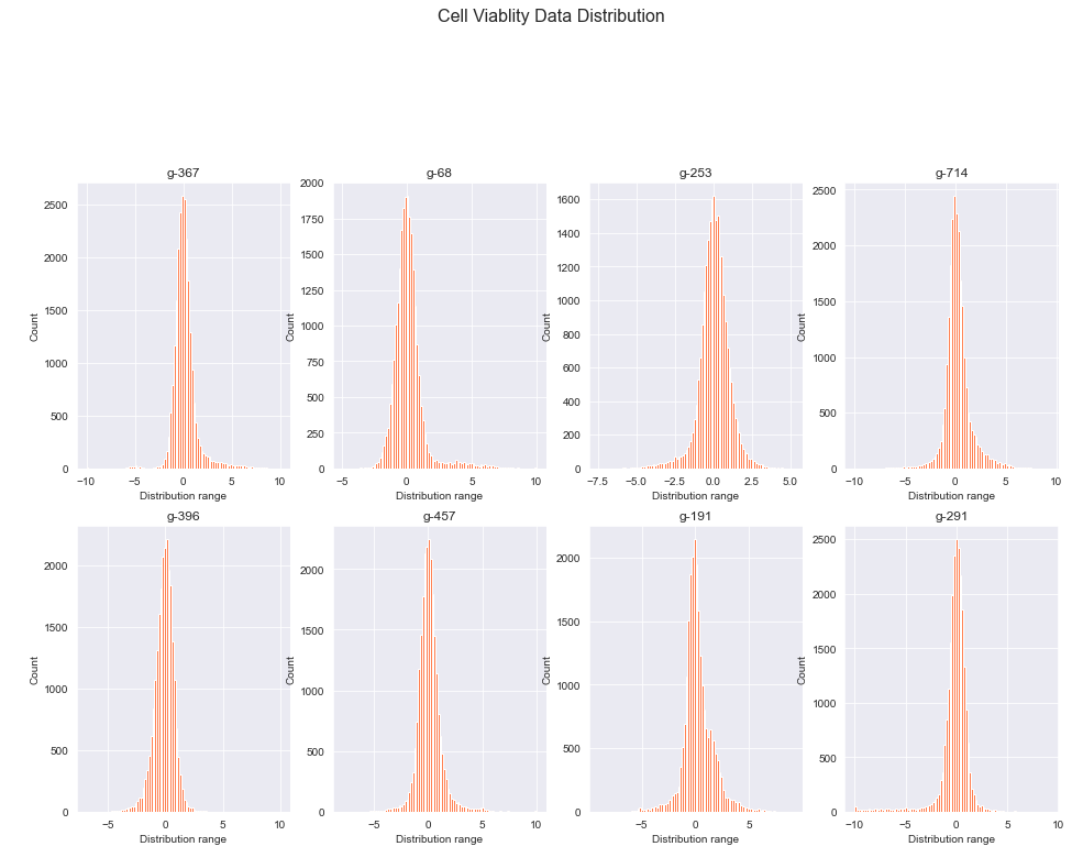
Figure 4: Plots for Gene data distribution



## **Cell viability data:**

Cell viability data is also a continuous numeric data and has readings for 100 different cell viability data responses for a particular drug. As we cannot have a look at response of all cell viability data thus, a random sample was taken and distribution range was also observed within -10 to +10 range which can be seen in figure 6.

Figure 6: Plots for Cell viability data distribution



## **Target Data:**

Out Target dataset had 206 different target sites where a drug can act. A drug can act on more than one site which is said to be a multi-Labelled problem (figure 7) but for our analysis we will be classifying the data into binary data (figure 8) i.e., the drug will show any effect or there will be no effect.

Figure 7: Multi-labelled data count plot Figure 8: Binary data count plot

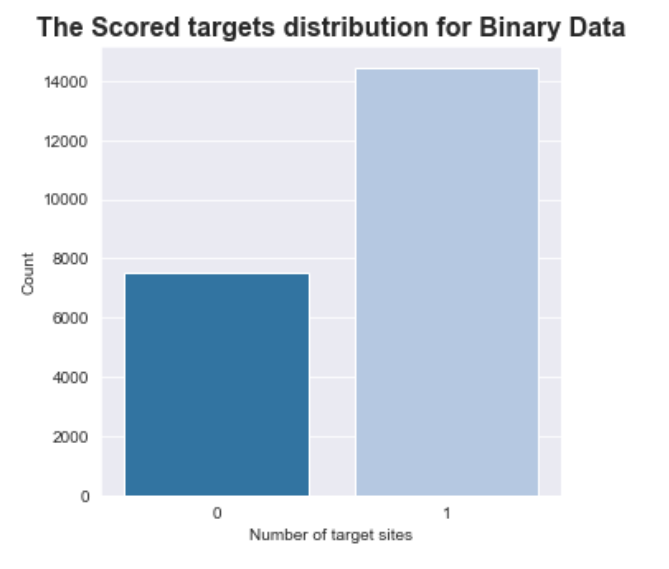
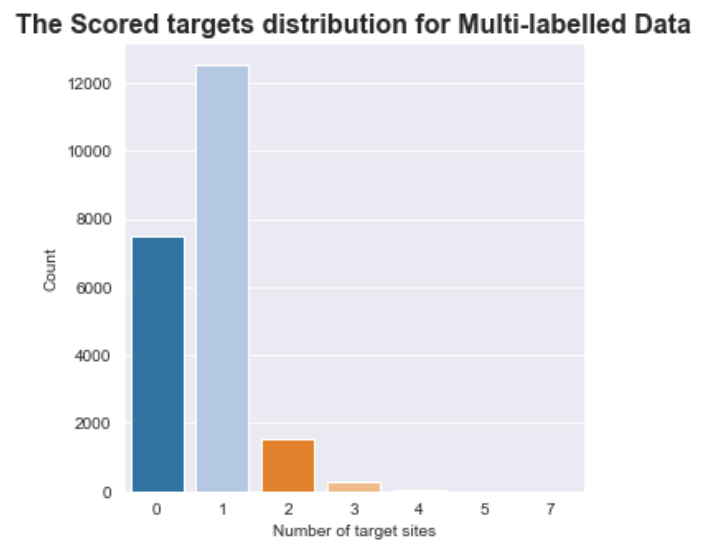
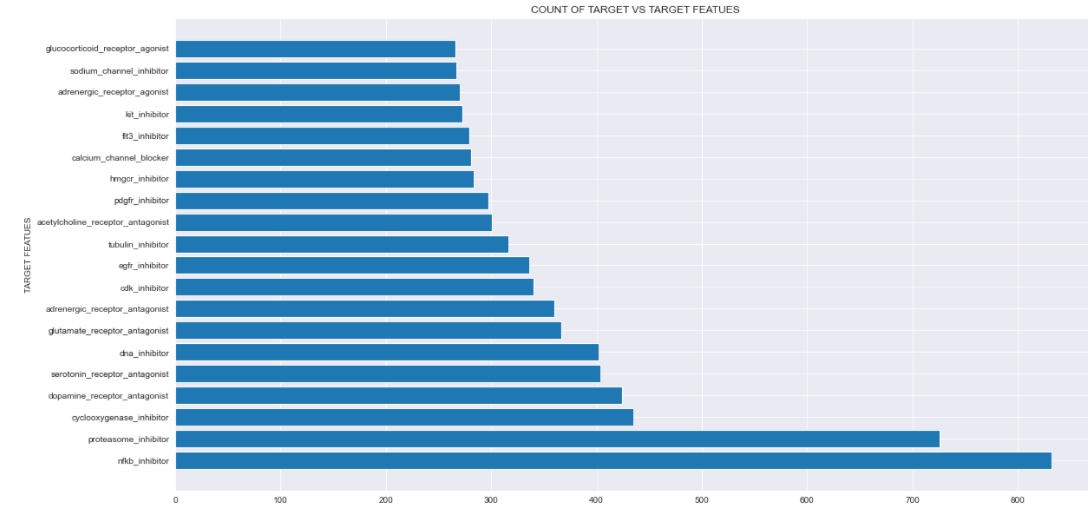


Figure 9: Top 20 Target features with counts



## **Correlations**

Correlations were taken using heatmap for Gene data, cell viability data and target data.

### **Gene Data correlations:**

Figure 10: Gene correlation data heatmap

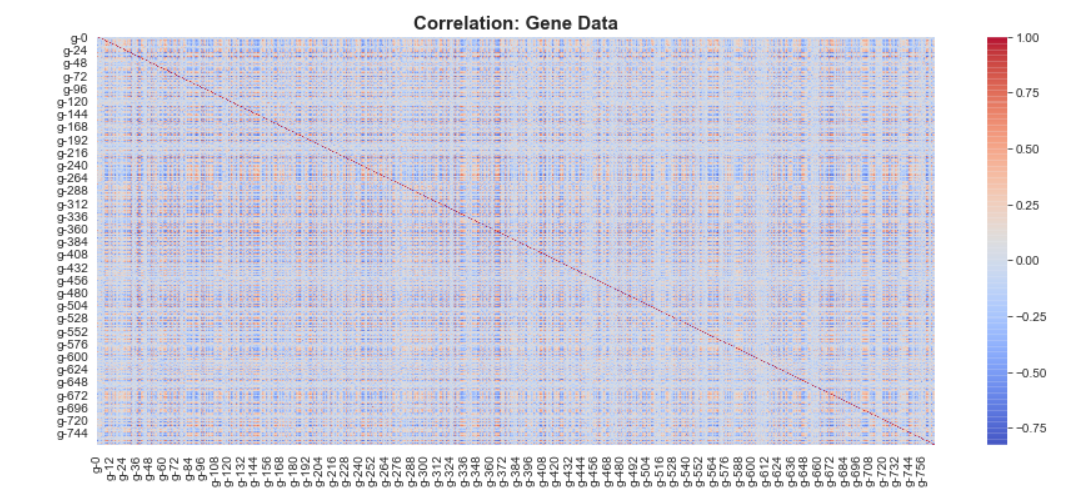


Table 2: Top 10 Gene Data correlation pairs

|  |  |  |
| --- | --- | --- |
| Correlation pairs | | Correlations |
| g37 | g50 | 0.906424 |
| g349 | g460 | 0.857138 |
| g50 | g489 | 0.851010 |
| g37 | g489 | 0.850305 |
| g369 | g569 | 0.846992 |
| g50 | g672 | 0.842562 |
| g248 | g460 | 0.840229 |
| g123 | g744 | 0.830518 |
| g63 | g195 | 0.828375 |
| g38 | g744 | 0.826617 |

Figure 10 shows heatmap for correlation within Gene data. Its difficult to exactly figure out the highest correlated pairs from the correlation matrix, hence a table with highest correlation is been made (Table 2). From all above data we can say that there is good amount of correlation in existence between gene data.

### **Cell Viability data correlations:**

Figure 11: Cell Viability correlation data heatmap

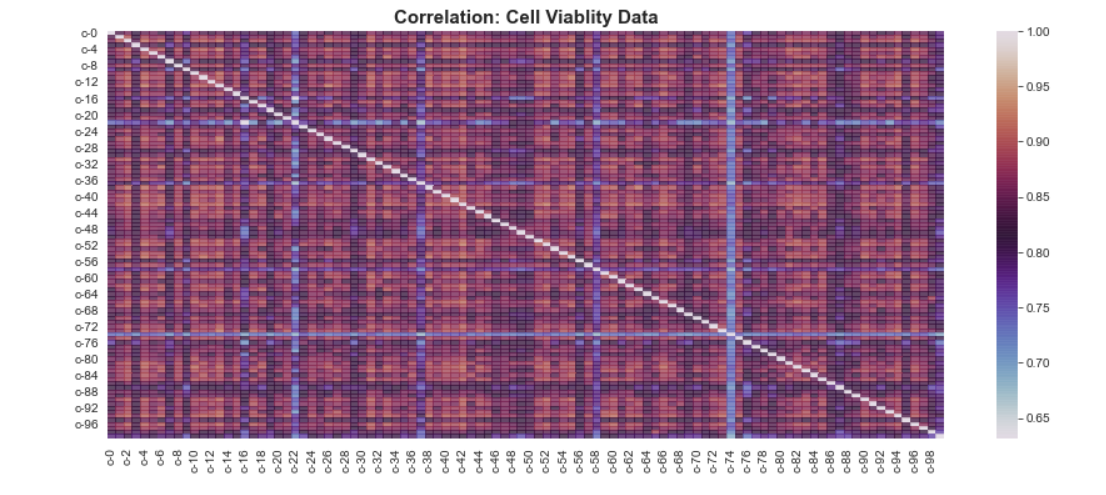


Table 3: Top 10 Cell viability Data correlation pairs

|  |  |  |
| --- | --- | --- |
| Correlation pairs | | Correlations |
| c42 | c52 | 0.928782 |
| c4 | c13 | 0.926462 |
| C13 | c26 | 0.924563 |
| c4 | c42 | 0.924094 |
| c38 | c63 | 0.923631 |
| c13 | c73 | 0.923532 |
| c4 | c52 | 0.920889 |
| c2 | c38 | 0.920750 |
| c11 | c55 | 0.920345 |
| c4 | c55 | 0.920303 |

Figure 11 shows heatmap for correlation within Cell Viability data. Heatmap plot for cell viability data shows good correlation within cells and same can be seen from table 3. We also observed that Cell viability data is more correlated within itself than Gene data.

### **Correlation between Target variables:**

Figure 12: Target data correlation heatmap

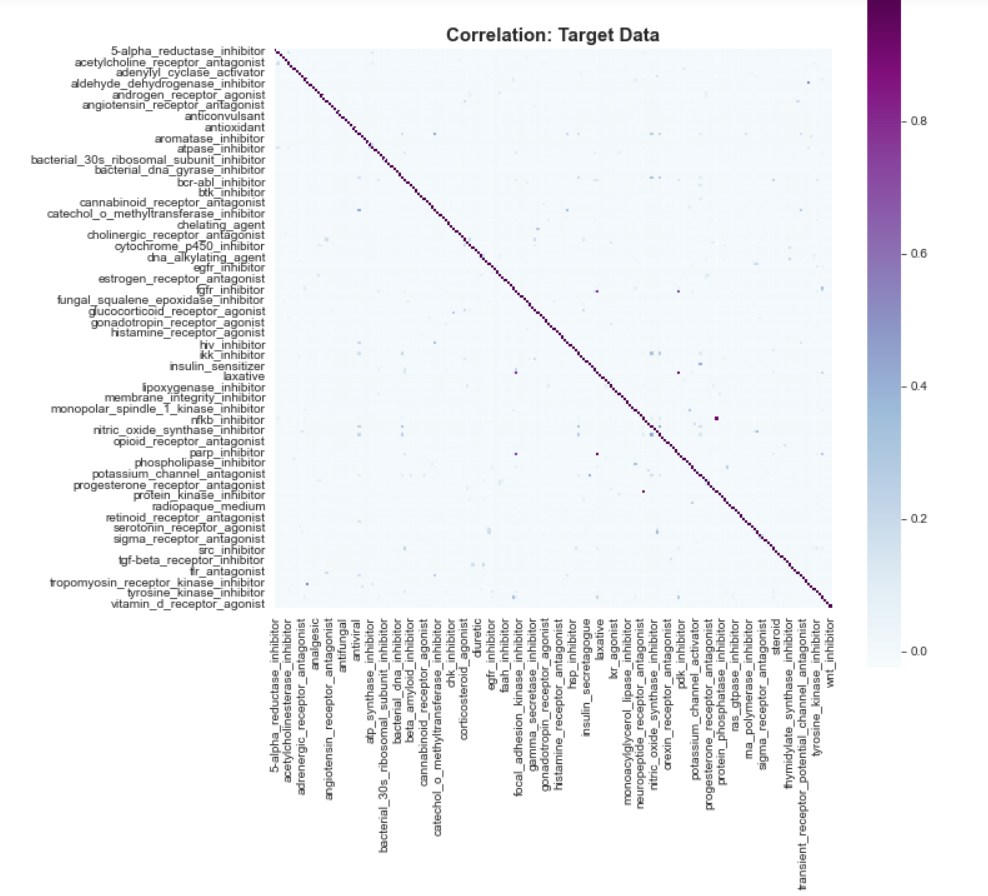


Table 4: Top 10 Target data correlation pairs

|  |  |  |
| --- | --- | --- |
| Correlation Pairs | | Correlations |
| kit\_inhibitor | pdgfr\_inhibitor | 0.994601 |
| nfkb\_inhibitor | proteasome\_inhibitor | 0.990562 |
| flt3\_inhibitor | kit\_inhibitor | 0.951336 |
| nfkb\_inhibitor | pdgfr\_inhibitor | 0.936127 |
| aldehyde\_dehydrogenase\_inhibitor | trpv\_agonist | 0.820699 |
| nitric\_oxide\_production\_inhibitor | nrf2\_activator | 0.743302 |
| insulin\_sensitizer | ppar\_receptor\_agonist | 0.666296 |
| apoptosis\_stimulant | caspase\_activator | 0.661034 |
| ikk\_inhibitor | nitric\_oxide\_production\_inhibitor | 0.631186 |
| bcl\_inhibitor | nitric\_oxide\_production\_inhibitor | 0.623892 |

Target data show good correlations between them.

### **Correlation between MoA data and target data:**

Figure 13: Correlation count plot between Target data and MoA features

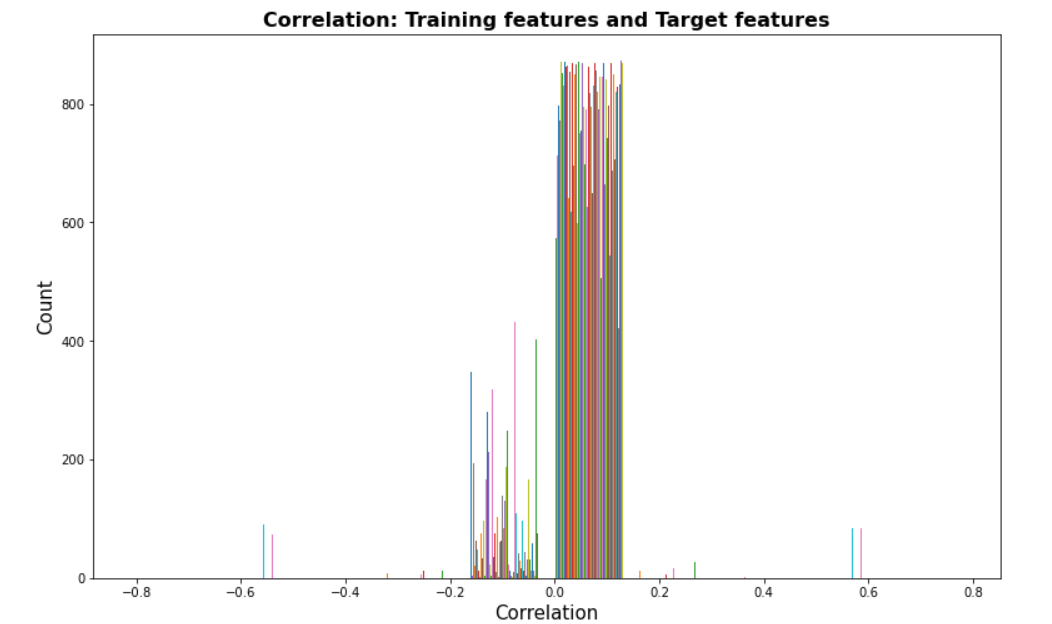
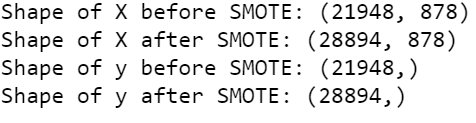


Figure 13 shows some that most of the correlated features are saturated within the range of 0 to 0.1 which show that there is weak positive correlation. We can also observe some high positive correlation along with high negative correlation.

# Data Engineering

## **SMOTE**:

* SMOTE (**Synthetic Minority Oversampling Technique) is a technique used to perform oversampling on minority class which in our case exist within target data.**
* **Shape of data was changed for both MoA data and target data.**

****

* X= MoA data and y= Target data

## **Feature importance:**

* Dataset has 875 features we need to take subset of relevant features in order to reduce background noise for our prediction.
* ExtraTreeClassifier was used to select important feature based on feature importance score. Feature importance was done on both Normal data and SMOTE data.
* Feature Importance on normal data and SMOTE data was dose using ExtraTreeClassifier and graph was plotted using feature importance score.

Figure 14: ExtraTreeClassifier score plot with Figure 15: ExtraTreeClassifier score plot

normal data with SMOTE data

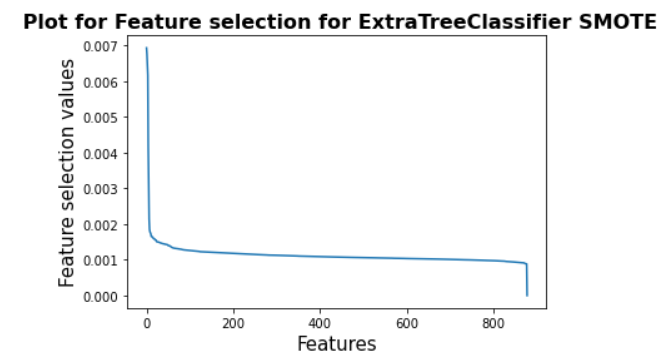
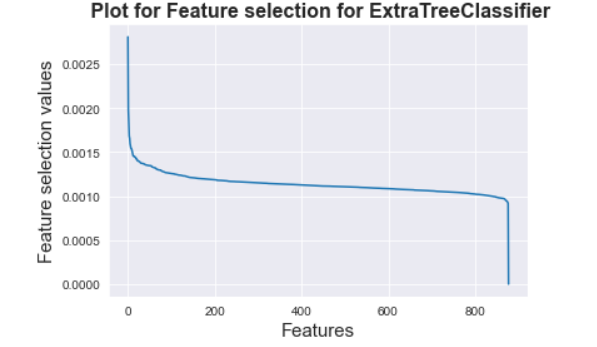
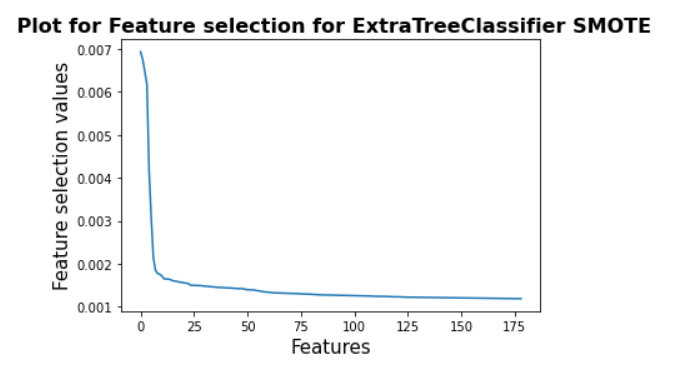


Figure 16: ExtraTreeClassifier score plot with Figure 17: ExtraTreeClassifier score plot

normal data (Zoomed) with SMOTE data (Zoomed)



* From above graphs for feature importance, we can observe a fall in feature selection values when the graph crosses 50 features, so in model building different features reading will be tried and model with best result will be taken for further analysis.

## **PCA**:

* PCA is done to reduce dimension of the dataset with large amount of data.
* PCA helps in reducing dimensions and improve performance of the model by reducing background noise.
* For this project PCA was done using two methods. In first method PCA was combines with Logistic regression and for second method, grid search using logistic regression was used for selecting number of components.
* PCA was done on both normal data and SMOTE data.

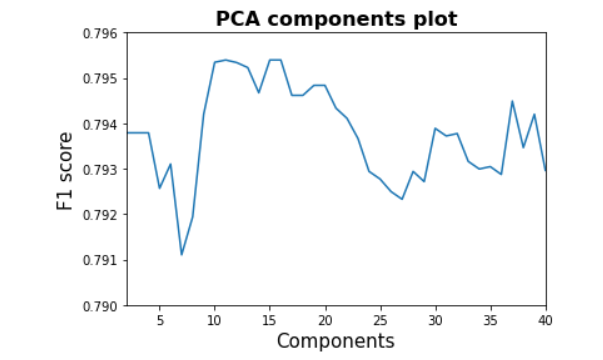
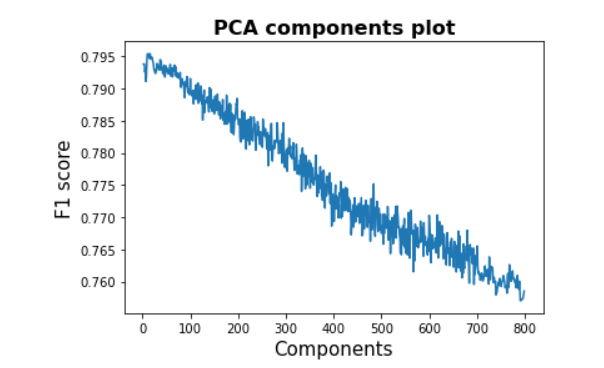
**PCA on normal data**:

1. PCA with logistic regression:

* PCA was done on normal data with logistic regression and F1 sore was plotted against different number of components.

Figure 18: PCA components plots with Normal Figure 19: PCA components plot on Normal

data data with 40 components



* Figure 18 represents a PCA plot of 800 components along with F1 score obtained after running Logistic regression on dataset.
* Figure 19 is a plot with less components than of figure 18 showing the peak of the graph which comes approximately between 10 to 15 components. In order to verify our selection of range of components grid search method was used with PCA.

1. PCA with grid search on normal data using logistic regression:

* In this method PCA was done using grid search. PCA explained variance ratio and classification accuracy was plotted against components and the best fitted line was plotted to indicate number of optimum PCA components.

Figure 19: PCA with component range 0 to 800 Figure 20: PCA with component range 0 to 70

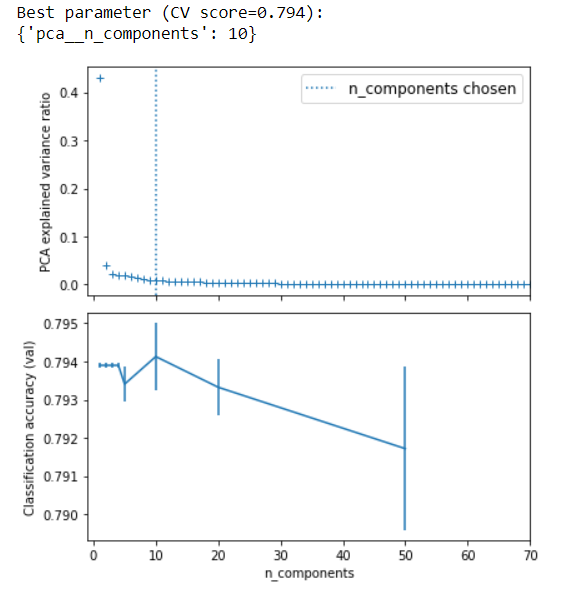
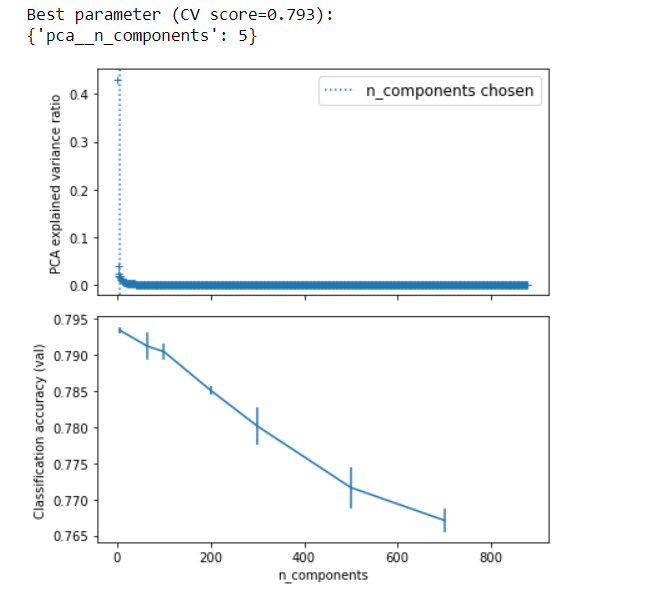
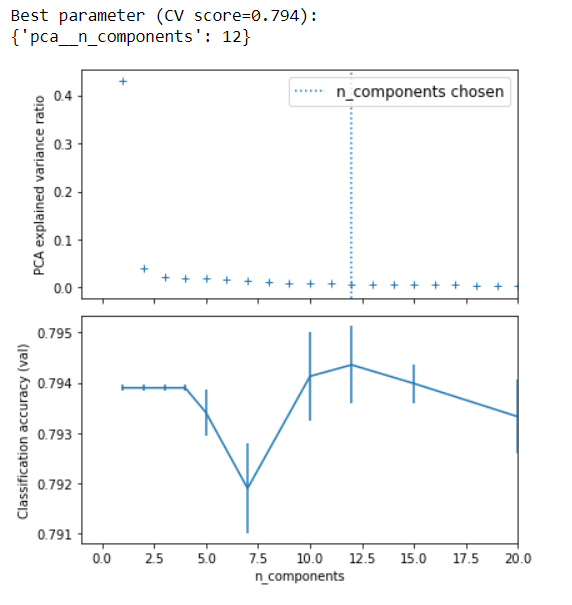


Figure 21: PCA with component range 0 to 20



* PCA with grid search was done using different range of components as input and best fitted lines can be observed in each graph above. The first graph in figure 19 shows PCA with 800 components and graph shows steep decline in classification accuracy.
* Optimum number of components suitable were 5. Since the graph was not much conclusive due to steep declining in accuracy, we narrowed down component range from 0 to 70 components and observed a good peak on figure 20 for classification accuracy plot with optimum number of components as 10.
* To fine tune the results more, we plotted a smaller range from 0 to 20 components and obtained optimum components as 12 as seen in figure 21.
* Comparing both methods used in PCA. Method number 1 gave us a range on 0 to 15 components and method 2 gave us optimum components as 12 which aligns in the range, we got in method 1. So henceforth, for model building, number of components from PCA will be taken as 12.

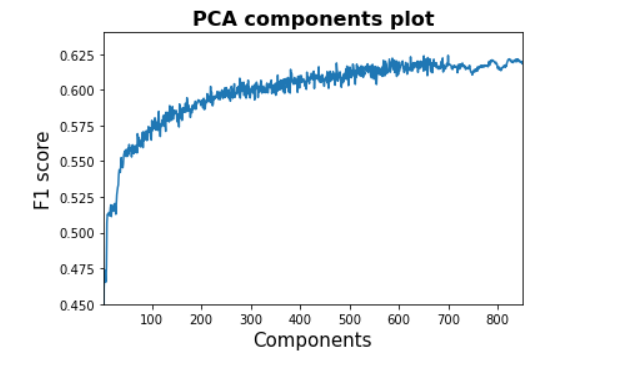
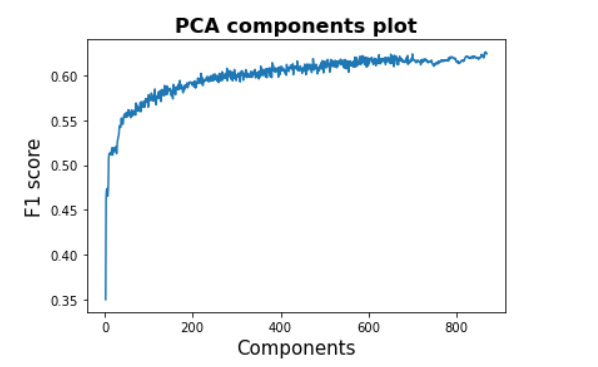
**PCA on SMOTE data:**

1. PCA with logistic regression:

* PCA was done on SMOTE data with logistic regression and F1 sore was plotted against different number of components.

Figure 22: PCA on SMOTE data Figure 23: PCA on SMOTE data with 800

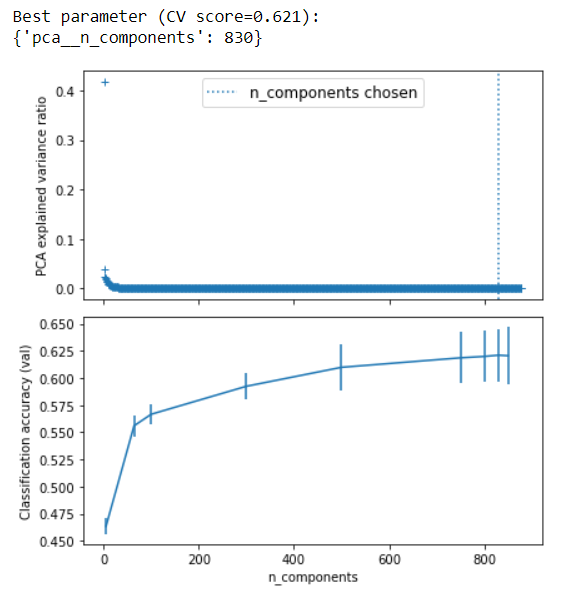
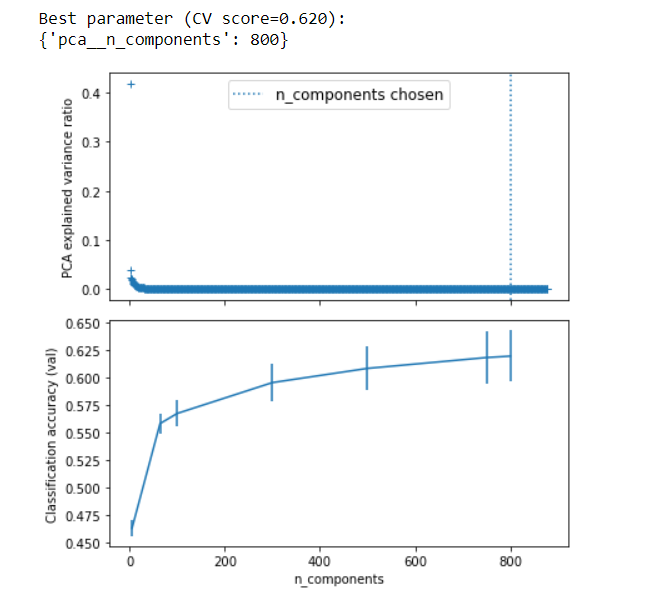
components



* SMOTE data on PCA with logistic regression showed a gradual increase in F1 score as the components increases.
* The reason for plotting PCA with 800 components in figure 23 was to plot up till endpoint of peak in figure 22 which was observed around 800 components.

1. PCA with logistic regression using grid search method:

Figure 24: PCA with component range 0 to 875 Figure 25: PCA with component range 0 to 850



* PCA on SMOTE data with logistic regression from above graph of figure 24 and 25 gives an optimum value of 830 components.
* Comparing two methods of PCA with SMOTE data we end up with optimum components in range of 800 to 830, and for our model building we would try both value of components and compare the F1 score.

# Results

* For the scope of this experiment, we are classifying results as binary i.e either the drug shows any action or there is no action.
* PCA and ExtraTree Classification for dimensionality reduction and SMOTE for balancing the dataset. For this experiment, 6 different datasets were used namely:

1. Normal MoA data
2. SMOTE MoA data
3. Normal MoA data with PCA
4. SMOTE MoA data with PCA
5. Normal MoA data with Classifier
6. SMOTE MoA data with Classifier.

* And 3 different Machine learning algorithms were used namely:

1. Logistic regression and logistic regression with cross validation
2. Neural Network
3. KNN and KNN with cross validation

* Each of the 6 different dataset was ran on all three algorithms

## **Evaluation metrics:**

* **Precision**: Precision score is between 0 and 1. Precision score will tell us, about how much the model will be correct in predicting that the drug will have one or more mechanism of action. So higher the precision score, higher the chances of model predicting it correctly6.
* **Recall**: Recall score is between 0 and 1. Recall score will tell us how correctly it can identify that mechanism of action of a drug out of all drug having mechanism of action. Higher the recall score, higher chances of identifying a drug having mechanism of action6.
* **F1 score**: F1 score conveys the balance between the precision and the recall. Selecting a model based on balance between precision and recall is a good way to rank models6.

## **Logistic regression:**

* Logistic regression and logistic regression with cross validation was performed on all 6 different datasets.
* For logistic regression Normal MoA dataset with PCA was found to have better results than any other datasets. The model was evaluated using F1 score, recall and precision.
* Table 2: Evaluation results for Logistic regression for Normal MoA dataset with PCA

|  |  |  |
| --- | --- | --- |
| Precision | Recall | F1 score |
| 0.663 | 0.992 | 0.795 |

* Precision score of 66 % means, the model will predict that the drug will have one or more mechanism of action, it would be correct 66% of the time and 99% recall means it correctly identifies 99% of all drugs having mechanism of action. F1 score of model is 79.5%
* For logistic regression with 5-fold with cross validation, Normal MoA dataset with classifier showed better results than any other dataset. It had an F1 score of 79.4%.
* Table 3: Evaluation results for Logistic regression with cross validation, Normal MoA dataset with classifier

|  |  |  |
| --- | --- | --- |
| Precision | Recall | F1 score |
| 0.665 | 0.984 | 0.794 |

## **Neural Network:**

* Neural networks (NNs) are considered a powerful technology in diverse types of biological research.
* The application of NNs is expected to exhibit excellent performance in predicting therapeutic effects and investigating relationships between genomic features, cell viability data with mechanism of action of drug7.
* When neural network was performed on the datasets, the best output was observed with SMOTE MoA data with a goof F1 score and well-balanced precision and recall.
* Table 4: Evaluation results for Neural Network, SMOTE MoA dataset

|  |  |  |
| --- | --- | --- |
| Precision | Recall | F1 score |
| 0.688 | 0.780 | 0.772 |

* Neural network models had a low recall rate, means its less like to correctly identify drugs having mechanism of action than logistic regression.

## **KNN:**

* KNN and KNN with cross validation was performed on all 6 datasets.
* For KNN, Normal MoA data showed better results than any other datasets with F1 score of 73.5%.
* Table 5: Evaluation results for KNN, Normal MoA dataset

|  |  |  |
| --- | --- | --- |
| Precision | Recall | F1 score |
| 0.688 | 0.792 | 0.735 |

* When using KNN with 5-fold cross validation similar results were obtained as of KNN and the optimum dataset was found to be Normal MoA dataset.
* When using KNN it was observed that, F1 score for all SMOTE datasets were very less, which is not good for our prediction.

## **Results of all experiments**

* Results for all experiments using all 6 datasets is as follow
* Table 6: Results of all models with different data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Normal MoA data | SMOTE MoA data | Normal MoA data with PCA | SMOTE MoA data with PCA | Normal MoA data with Classifier | SMOTE MoA data with Classifier |
| Logistic regression | F1- 0.753 | F1- 0.677 | F1- 0.795 | F1- 0.620 | F1- 0.794 | F1- 0.681 |
| Precision- 0.680 | Precision- 0.647 | Precision- 0.663 | Precision- 0.658 | Precision- 0.665 | Precision- 0.608 |
| Recall- 0.843 | Recall- 0.709 | Recall- 0.992 | Recall- 0.586 | Recall- 0.984 | Recall- 0.785 |
| Logistic regression  With cross validation | F1- 0.753 | F1- 0.580 | F1- 0.79 | F1- 0.621 | F1- 0.794 | F1- 0681 |
| Precision- 0.680 | Precision- 0.955 | Precision- 0.663 | Precision- 0.668 | Precision- 0.665 | Precision- 0.610 |
| Recall- 0.843 | Recall- 0.157 | Recall- 0.982 | Recall- 0.590 | Recall- 0.984 | Recall- 0.791 |
| Neural Network | F1- 0.720 | F1- 0.772 | F1- 0.708 | F1- 0.738 | F1- 0.705 | F1- 0.713 |
| Precision- 0.697 | Precision- 0.780 | Precision- 0.678 | Precision- 0.770 | Precision- 0.701 | Precision- 0.733 |
| Recall- 0.745 | Recall- 0.688 | Recall- 0.742 | Recall- 0.709 | Recall- 0.708 | Recall- 0.690 |
| KNN | F1- 0.735 | F1- 0.270 | F1- 0.718 | F1- 0.270 | F1- 0.726 | F1- 0.472 |
| Precision- 0.688 | Precision- 0.955 | Precision- 0.680 | Precision- 0.955 | Precision- 0.695 | Precision- 0.824 |
| Recall- 0.792 | Recall- 0.157 | Recall- 0.762 | Recall- 0.157 | Recall- 0.760 | Recall- 0.331 |
| KNN with cross validation | F1- 0.736 | F1- 0.270 | F1- 0.719 | F1- 0.270 | F1- 0.726 | F1- 0.478 |
| Precision- 0.688 | Precision- 0.955 | Precision- 0.681 | Precision- 0.955 | Precision- 0.694 | Precision- 0.842 |
| Recall- 0.791 | Recall- 0.157 | Recall- 0.762 | Recall- 0.157 | Recall- 0.761 | Recall- 0.313 |

* From all above results we can say that logistic regression on Normal MoA data with PCA gave the best results for our classification.
* PCA and feature importance both worked well on Normal datasets giving better results than using SMOTE for the datasets.

# **Conclusion**

The analysis for this experiment gave us lots of valuable insights from the data provided to us.

Initially we were able to establish that a correlation exists between gene data, cell viability data and target sites, which formed the base for our future analysis. Multi-labelled problem was transformed to binary problem and our focus was shifted from finding one or more than one target sites to finding if the drug exhibits any mechanism of action or it has no mechanism of action. Binary target was pretty imbalanced, so to make it balanced SMOTE was used to balance dataset. Since we had more than 800 features in our dataset, we had to perform PCA and feature importance using ExtraTree Classifier for dimensionality reduction.

Utilising all our analysis in models we found that Logistic regression on Normal data with PCA gave us best results for classification of mechanism of action. We found a good balance in recall and precision for the model along with good F1 score.

The experiment we performed would help scientists in their initial stage of analysis for any given drug which has gene and cell viability data. Scientist will be able to predict that a drug would have any mechanism of action or there would be no mechanism of action. This can save them a lot of time, energy and resources in real world scenario. It can also act as a catalyst to find multi-labelled classification and can majorly contribute to clinical study of drugs in Phase 1 analysis.

# References:

1. *Mechanisms of action (moa) prediction*. Kaggle. (n.d.). [https://www.kaggle.com/c/lish-moa/](https://www.google.com/url?q=https://www.kaggle.com/c/lish-moa/&sa=D&source=hangouts&ust=1628044904065000&usg=AFQjCNGMxqC5Fe8c8QT64jLZKDxfTnuG3Q).
2. Aliper, A. (2015). *Deep learning applications for predicting pharmacological* [https://pubs.acs.org/doi/abs/10.1021/acs.molpharmaceut.6b00248?src=recsys](https://www.google.com/url?q=https://pubs.acs.org/doi/abs/10.1021/acs.molpharmaceut.6b00248?src%3Drecsys&sa=D&source=hangouts&ust=1628045302012000&usg=AFQjCNHeoqPJb3CSga0rSs-DIXfmoAlB8Q)
3. Vamathevan J; Clark D; Czodrowski P; Dunham I; Ferran E; Lee G; Li B; Madabhushi A; Shah P; Spitzer M; Zhao S; (n.d.). *Applications of machine learning in drug discovery and development*. Nature reviews. Drug discovery. [https://pubmed.ncbi.nlm.nih.gov/30976107/](https://www.google.com/url?q=https://pubmed.ncbi.nlm.nih.gov/30976107/&sa=D&source=hangouts&ust=1628045422914000&usg=AFQjCNGjGmPKkjj5pEO0q2oFyFZFDnlrUg)
4. Göller, A. H, Kuhnke L, Montanari F, Bonin A, Schneckener S, Laak A ter, Wichard, J, Lobell, M., & Hillisch, A. (2020, July 9). *Bayer's in silico admet platform: A journey of machine learning over the past two decades*. Drug Discovery Today. [https://www.sciencedirect.com/science/article/pii/S1359644620302609](https://www.google.com/url?q=https://www.sciencedirect.com/science/article/pii/S1359644620302609&sa=D&source=hangouts&ust=1628045532772000&usg=AFQjCNExqHAISI0NuOMFRucpNjeLcSAIxg).
5. Google. (n.d.). *Classification: Precision and recall | machine learning crash course*. Google. [https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall](https://www.google.com/url?q=https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall&sa=D&source=hangouts&ust=1628045875544000&usg=AFQjCNHuhLYSWqgrJvXNof9H6lrSAtDfEA).
6. Choi, J., Park, S., & Ahn, J. (2020, February 5). *Refdnn: A reference drug based neural network for more accurate prediction of anticancer drug resistance*. Nature News. [https://www.nature.com/articles/s41598-020-58821-x](https://www.google.com/url?q=https://www.nature.com/articles/s41598-020-58821-x&sa=D&source=hangouts&ust=1628045944811000&usg=AFQjCNGdKJf4tBZkKoKzIvFIn5ia_3bYfA).

1. [↑](#footnote-ref-1)