Drug-Disease Prediction and its Feature Selection Analysis

Mayank Murali mqm6516

May 6, 2020

Contents

1 Abstract

With growing number of diseases, finding the right drug to treat it is becoming a growing concern. The process of drug repositioning or drug repurposing identifies new indications for marketed drugs after validation to treat a particular disease.

The term project works on a computational approach to predict and analyze the results of drug repositioning based on few machine-learning algorithms such as Logistic Regression, Support Vector Machine and Random Forest as part of the ML classifiers. The computation is done with respect to features such as similarity between drug chemical structures, on how close are their targets within the drug- drug and disease-disease and on how correlated are the gene expression (miRNA).

2 Introduction

In spite of advancement in technology and improvement human disease identification, the progress to capitalizing this benefit to resolve any human disease is sluggish. The pharmaceutical industry faces quite a lot of challenge. Some of the notable challenged are increased time to bring new drugs to the market, high investment cost [1].

Drug repositioning or drug repurposing is the process of discovering, validating, and marketing previously approved drugs for new indications, is of growing interest to academia and industry due to reduced time and costs associated with repositioned drugs. Traditional drug development strategies usually take around 10-15 years for development whereas drug repositioning can be performed in 8-10 years. It consists of five stages: discovery and preclinical, safety review, clinical research, FDA review, and FDA post-market safety monitoring. However, there are only four steps in drug repositioning: compound identification, compound acquisition, development, and FDA post-market safety monitoring (Figure 7) [2].

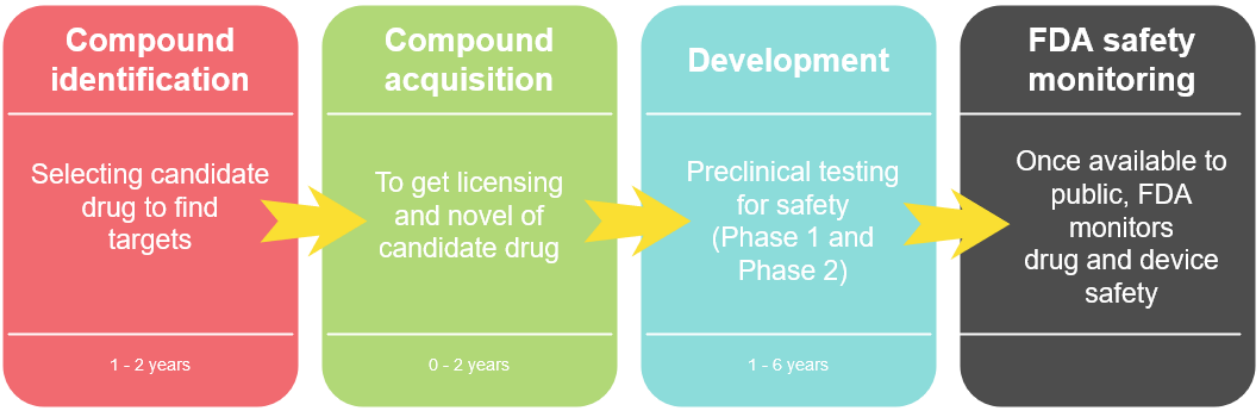


Figure 1: Flowchart of drug repositioning.

There are a variety of drug repositioning approaches such as computational approach,

biological experimental approach and mixed approach. Advancement in fields like data-mining, machine learning, and network-based approaches offer great opportunity to perform computational drug repositioning. Harnessing the data sources from genomics and biomedical public domains and developing predictive models has become even more accurate after applying applying novel algorithms and computational methods. Mostly, existing data such as chemical structure, profiles of side-effect, protein targets or gene expression are used for drug repositing.

Computational drug repositioning methods focus on shared characteristics between two drugs and depending on what kind of drug discovery either drug-based or disease-based [3]. Computational drug repositioning methods can be classified in to target-based, expression-based, knowledge-based, chemical structure-based, pathway-based and mechanism of action-based. In this project, we work on target-based method.

2.1 Goals

The main goal of this project is to develop an efficient model that can identify the type of drug for a particular disease. A lot of features are considered in this process.

3 Bibliography

This paper [4] focuses on a drug-centered approach by predicting the therapeutic class of FDA-approved compounds using a state-of-the-art machine learning model. Important features considered in this paper includes the similarity between drug chemical structure, correlation between the gene expression patters etc. Enhancing the predictive power of computation techniques have removed the liability of deriving drug repositioning from drug-disease interactions, variability and sparsity of data currently available for the diseases which derive from patients already treated with other drugs in most of the cases. 410 different types of drugs have been studied using layers of information based on similarities in chemical structures, molecular targets and gene expression signatures, the drug-drug similarities are identified on the ranks of the genes. The machine learning model implements multi-class Support Vector Machine (SVM). Data integration, noise reduction and classification are performed as part of the model. To understand the results, Receiver Operating Characteristic (ROC) curve, similarity matrices were used to obtain a maximum accuracy of 78%.

This article [5] identifies repositioning opportunities for schizophrenia as well as depression and anxiety disorders using machine learning techniques such as deep neural networks (DNN), support vector machines (SVM), elastic net (EN), random forest (RF) and gradient boosted machines (GBM). These predictive models were run on the drug expression data obtained from Connectivity Map or CMap. This data was normalized using MAS5 algorithm to average the expression levels of the genes. Since the data was imbalanced, weighted and unweighted analysis was considered for prediction. The predictive model was implemented on Python using "scikit-learn". The hyperparameter selection for DNN was performed using grid search with 2 neural net layers and for SVM, RF and

GBM models, built-in function "GridSearchCV" and nested three-fold cross validation (CV) was used. In order to predict the performance, ROC curve and precision-recall curve was used.

For weighted analysis, it was observed that SVM gave the best results considering the log loss. Considering the ROC-AUC as performance metric, SVM and EN performed well. For PR-AUC, again SVM performed the best. Weighted analysis saw improvement in other models. SVN and EN performed well in general. DNN performed best considering the ROC-AUC curve and SVM achieved the best for PR-AUC.

This paper [6] proposes an integrative computational framework to predict novel drug indications for both approved drugs and clinical molecules by integrating chemical, biological and phenotypic data sources. They define similarity measures for each of 1007 drugs against 719 diseases date and utilized a weighted k-nearest neighbor algorithm to transfer similarities of nearest neighbors to prediction scores for a given compound. SLAMS algorithm is introduced for drug repositioning by integrating multiple data sources. The features or similarity measures of drug chemical structure, protein targets, side-effect profile are considered.

The weighted k-nearest neighbor (k-NN) was used after optimization of few model parameters. The model was used on dataset obtained from DrugBank, PubChem, UMLS. Evaluation measures used include classification evaluation metrices, recall, F-score, precision. Overall, 18392 unique drug disease associations was acquired. The paper concludes that the false positive novel uses predicted by the method attained significant coverage of drug-disease associations tested in clinical trials.

Computational drug repositioning strategies and ways to improves this to develop more powerful approaches are discussed in this paper [8]. Xue et al., also summarizes 76 important resources about drug repositioning and also describe the challenges and opportunities from different perspectives such as commercial models, technology, patents and investment. Different computational approaches such as network-based approach, network-based cluster approach, network-based propagation approach, text mining-based approach, semantics-based approach are discussed. The important challenges common to most of the mentioned approach includes selecting the appropriate approach to make full use of massive amounts of medical data, a growing need to develop a new commercial model because the traditional commercial model is a serial model and causes overlapping investment issues. From the market perspective, a large number of diseases require new drugs to be treated, which brings potential economic benefits.

Talevi et al.,2019 paper [9] talks about the challenges and opportunities with drug repositioning. Intellectual property and economic considerations, data and compound availability, exhaustion of repurposing space are few important challenges described and analysed in this paper. The opportunities include rare and neglected conditions of drug repositioning such as rare diseases identification, precision medicine where individual variability in genes, environment, and lifestyle for each person to decide on or pursue an

appropriate treatment, systems medicine which provides an integrative perspective on phenotypic-oriented and target-oriented, 'rational' drug discovery, collaborative models. Few of these challenges can be overcome by collaboration between the public and private sectors. Government agencies and organizations have the tools to elaborate solutions to overcome some of the legal and commercial barriers faced by drug repurposing projects. One another alternative strategy is network pharmacology which underlines the possible advantages of polypharmacology to treat complex disorders.

DrugPredict [7], a novel computational drug-repositioning approch to rapidly identify potent drug candidates for cancer treatment is developed. This is a very specific case where drug repositioning is used for a particular disease. DrugPredict represents an innovative computational drug- discovery strategy to uncover drugs that are routinely used for other indications that could be effective in treating various cancers, thus introducing a potentially rapid and cost-effective translational opportunity. This is a very specific and complex paper to understand which has little relevance to what I plan to implement but worth a read.

3.1 Related Work on the dataset area

Computational drug repositioning has become a very common approach ever since vast amount of drug and disease knowledge databases have become available to public. Few of the well known sources are DrugBank, ChemBank, OMIM, KEGG, and PubMed.

Some familiar work on the dataset area are related to precision medicine, systems medicine [**?**], approaches to rapidly identify potent drug candidates for cancer treatment [7] etc.

3.2 Related work on the Pattern Recognition Approach

Related work on the pattern recognition approach include training deep neural networks to predict the therapeutic use of a large number of drugs using gene expression data. Another relevance of dimensionality reduction is seen in computational drug reposition where the dimensionality of the data is reduced while retaining biological relevance for training and developing deep neural networks/models.

4 Data

The dataset for this project is obtained from GitHub. Following are the dataset details: Features:

- Disease name
- Drug name
- 491 Different chemical substructure similarity between drugs
- 817 Side effect that describe drug-drug similarity

- Sharing 136 targets between drugs
- Sharing 64 genes between diseases
- 136 miRNA sharing between diseases

Class:

- 0 - the drug cannot be used to treat the disease.
- 1 - the drug can be used to treat the disease.

Figure 2 illustrates the length of the test dataset along with number of features.

```
In [77]: len(trainData)
Out[77]: 1000

In [78]: len(trainData.columns)

Out[78]: 1648
```

Figure 2: Dataset information.

Figure 3 portrays the raw dataset and Figure 4 represents the features mentioned in the dataset.

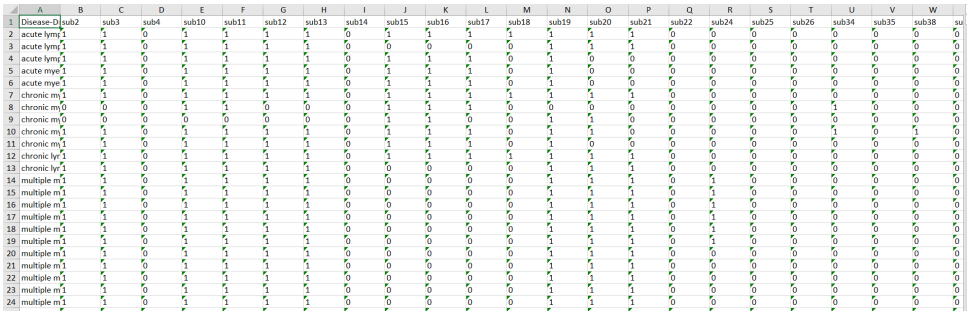


Figure 3: Image of raw dataset.

The below Figure 5 illustrates the mean and standard deviation of the features in the dataset.

Identifying features

```
In [9]:
feat = trainData.columns
#feat = trainData.iloc[0, : ]
feat

Out[9]: Index(['Disease-Drug', 'sub2', 'sub3', 'sub4', 'sub10', 'sub11', 'sub12',
       'sub13', 'sub14', 'sub15',
       ...
       'hsa-mir-7-1', 'hsa-mir-7-2', 'hsa-mir-7-3', 'hsa-mir-765',
       'hsa-mir-9-1', 'hsa-mir-9-2', 'hsa-mir-9-3', 'hsa-mir-92a-1',
       'hsa-mir-96', 'Class'],
      dtype='object', length=1648)
```

Figure 4: Types of features in the dataset.

Figure 5: Mean and standard deviation of all the features.

5 Methods

In this section, a brief plan of methods and approaches will be introduced.
The following are the main highlighted steps in proceeding to complete the project.

5.1 Steps to achieve the goal

- Step 1: Reorder data based on features

  I plan to reorder data based on the features to make sets of features in order to move forward with dimensionality reduction.

- Step 2: Feature dimensionality reduction

  Identify and chose the right features for dimensionality reduction.

- Step 3: Feature elimination

  Eliminate features based using RFE to 200, 100 and 20 (maybe 3 as well).

- Step 4: PCA

Additionly perform PCA to compare with RFE.

- Step 5: Classifier/ model training

  Use Support Vector Machine, Logistic classifier and Random Forest models to determine the best accuracy with different sets of features obtained using RFE and PCA.

The data set gets split into three ways

- 80:20,
- 70:30 and
- 60:40

The support vector machine classifier was chosen as the baseline classifier model since couple of papers [4] and [5] have cited good accuracy using SVM for their particular dataset. The following baseline was established for this project. The baseline accuracy using SVM on the whole dataset with all features is represented by Figure 6 measuring a maximum of 97%.

```
Test set accuracy: 0.065
Test set accuracy: 0.08
Test set accuracy: 0.19
Test set accuracy: 0.655
Test set accuracy: 0.05
Test set accuracy: 0.22
Test set accuracy: 0.015
Test set accuracy: 0.345
Test set accuracy: 0.035
Test set accuracy: 0.965
Test set accuracy: 0.965
Test set accuracy: 0.965
Test set accuracy: 0.965
Test set accuracy: 0.965
Test set accuracy: 0.97
Test set accuracy: 0.975
Test set accuracy: 0.955
Test set accuracy: 0.955
Test set accuracy: 0.95
```

Figure 6: Accuracy obtained for baseline model.

5.2 Feature Dimensionality Reduction

Feature reduction or dimensionality reduction is the process of reducing the number of features without losing important information. For this drug repositioning dataset, we find a total of 1648 features which needs to be reduced. Reducing the number of features means the number of variables is reduced making the computation faster. For this project the following two techniques are adopted to get three cases (200 features, 100 features and 20 features). Additionally, a case with just 3 feature is included.

5.2.1 Recursive Feature Elimination

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the model's coefficient or feature importance attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model. Here we use logistic regression model for RFE.

RFE requires a specified number of features to keep, however it is often not known in advance how many features are valid. To find the optimal number of features cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features.

5.2.2 Principal Component Analysis

Principal component analysis (PCA) is used to reduce the dimensionality of a data set consisting of many variables correlated with each other. It is done by transforming the variables to a new set of variables, which are known as the principal components. The 1st principal component retains maximum variation that was present in the original components. The principal components are the eigenvectors of a covariance matrix.

5.3 Classifier Models

The following classifier models are used to perform binary classification on drug repositioning dataset.

5.3.1 Logistic Regression Classifier

The logistic function commonly called the sigmoid function represented as

$$1/(1 + e^{-val})$$

where e is the base of the natural logarithms and val is the numerical value to be transformed. Since this is a binary classification problem, logistic regression is a good classifier to be used here as it is a linear model that predicts based on the logistic function.

5.3.2 Support Vector Machine

We use SVM to perform both classification and regression. But for this project. we will perform the classification task.

We determine the vector length using:

$$||x|| = \sqrt{\sum_{n}^{i=1} x_i^2}$$

Next determine the hyperplane to divide the two sets of classes using the equation:

$$W.X + b = 0$$

Linear Kernel is used for SVM.

5.3.3 Random Forest

Random Forest classifier is an ensemble tree-based learning algorithm. It is a set of decision trees from randomly selected subset of training set. It works by aggregating the votes from different decision trees to decide the final class of the test object. Each classifier in RF $h_k(x) = h(x|\Theta_k)$ is a predictor of n such that $y = \pm 1$ = outcome associated with the input x, where: Θ = parameter of tree that determines a random subset D_Θ and $x = \{x_1, x_2..x_d\}$.

5.4 Evaluation Criteria

For this project, we model the drug repositioning task as a binary classification problem where each drug either treats or does not treat a particular disease. We measure the final classification performance using three criteria: accuracy, specificity, and area under the ROC curve. In order to provide the definitions of these three criteria, we first define the classification confusion table for binary classification problems where the two classes are indicated as positive and negative, which is constructed by comparing the actual data labels and predicted outcomes. Then we can define the classification evaluation metrics as:

$$TruePositiveRate = TP/(TP + FN),$$

$$FalsePositiveRate = FP/(FP + TN),$$

$$Precision = TP/(TP + FP),$$

$$Recall = TP/(TP + FN).$$

6 Timetable

Here is a timetable that will be followed to complete the project Figure 7.

	Term Project Timeline	Duration	Start	ETA
1	Computational Drug Repositioning using Machine Learnin	43 days	19.03.20	30.04.20
2	Identify suitable data/ research papers/ algorithms.	5 days	19.03.20	23.03.20
3	Label data	2 days	24.03.20	25.03.20
4	Re-order data	1 day	26.03.20	26.03.20
5	Features Dimensionality Reduction	5 days	27.03.20	31.04.20
6	Feature Elimination	3 days	01.04.20	03.04.20
7	Train classifier or models	8 days	04.04.20	11.04.20
8	Data/result visualization	2 days	12.04.20	13.04.20
9	Report and documentation	22 days	27.03.20	17.04.20
10	Presentation preparation	2 days	18.04.20	19.04.20
11	Buffer	10 days	20.04.20	30.04.20

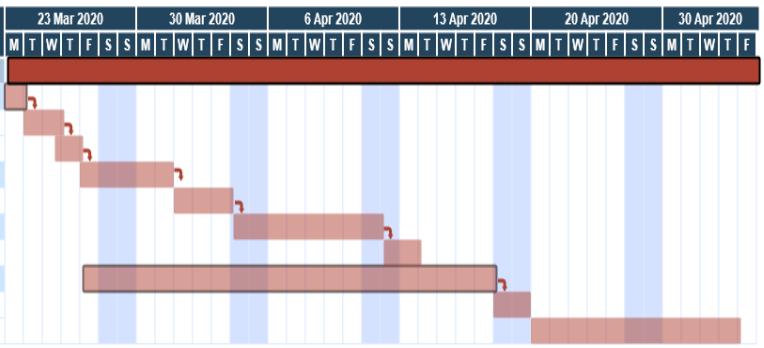


Figure 7: Gantt chart timetable for term project

6.1 Alternatives if failure of primary methods

Working on dimensionality reduction would matter, depends how many features to keep and how many to reduce and how many to compare. Principal component analysis (PCA) can be included. Different machine learning models can be improvised if dimensionality is as issue.

7 Results

We discuss and analyse the performance of each classification model based on the feature elimination performed on the drug repositioning dataset. https://www.overleaf.com/project/5e275bd5f1
This section is divided into three based on the cases. The feature elimination techniques applied to achieve these features and their performance on both the models will be discussed individually.

7.1 Case 1- 200 features

To begin with, the case 1 with 200 features obtained using RFE and PCA. The following Figure 8 represents the distribution of features after using RFE as feature elimination technique, features obtained using PCA and in order to identify the relation between these features, Pearson correlation is found.


```
#Merge features to extract in CSV format
extractPCACSV200 = trainData[['Disease-Drug']]
dfPCA = pd.DataFrame(featPCA200)
extractPCACSV200 = pd.merge(extractPCACSV200, dfPCA, left_index = True, right_index=True)
extractPCACSV200.loc[:, 'Class'] = trainData.loc[:, 'Class']
extractPCACSV200.to_csv('trainData_PCA_200.csv', index = False)
extractPCACSV200[:5]
```

	Disease-Drug	0	1	2	3	4	5	6	7	8	...	191	192	193
0	acute lymphoblastic leukemia (all) (precursor ...	6.693368	-1.658225	9.362689	0.190428	-4.016489	-1.765267	0.381426	-0.041606	-0.747267	...	6.336077e-16	-3.638582e-16	1.873501e-16	-2.026
1	acute lymphoblastic leukemia (all) (precursor ...	2.295896	0.810448	-6.167722	-2.030408	-4.156993	0.223584	5.124359	-1.778741	-3.750664	...	2.168404e-18	2.190088e-16	1.318390e-16	-2.199
2	acute lymphoblastic leukemia (all) (precursor ...	4.662952	0.021688	4.578135	-1.424623	1.357322	0.205435	-1.470713	0.142984	-0.007939	...	3.404395e-17	-7.068998e-17	-4.198031e-16	-3.022
3	acute myeloid leukemia (aml)-imatinib	4.658048	0.006389	4.569564	-1.406955	1.371682	0.206565	-1.464812	0.166243	0.026837	...	3.751340e-17	-5.941428e-17	-4.163336e-16	-3.080
4	acute myeloid leukemia (aml)-imatinib	4.670829	0.027438	4.557820	-1.466503	1.402681	0.308281	-1.214811	0.293733	0.336807	...	1.734723e-17	-6.938894e-17	-4.163336e-16	-3.035

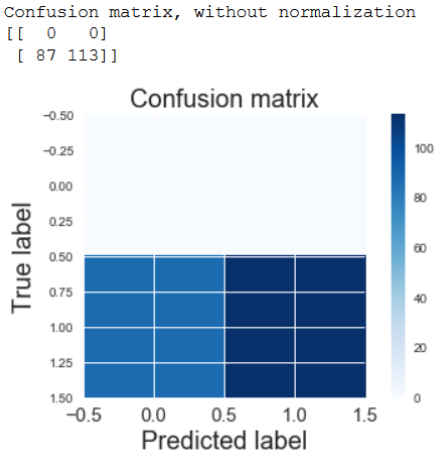
5 rows × 202 columns

Figure 8: Recursive Feature Elimination, PCA and Pearson correlation for 200 features.

Below are the results obtained after running SVM on the RFE obtained feature. The figures are in the following order: Figure 9 shows the confusion matrix and accuracy.

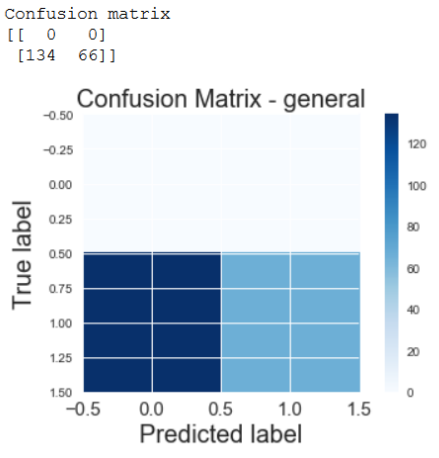
The following results are obtained after performing SVM classifier on PCA features. Figure 10 shows the confusion matrix and obtained accuracy.

After applying LR on RFE, the achieved score is 0.64 and the score on PCA features is 0.642.




```
hinge 1
Test set accuracy: 0.055
squared_hinge 1
Test set accuracy: 0.2
hinge 11
Test set accuracy: 0.08
squared_hinge 11
Test set accuracy: 0.325
hinge 21
Test set accuracy: 0.07
squared_hinge 21
Test set accuracy: 0.455
hinge 31
Test set accuracy: 0.16
squared_hinge 31
Test set accuracy: 0.51
hinge 41
Test set accuracy: 0.505
squared_hinge 41
Test set accuracy: 0.545
hinge 51
Test set accuracy: 0.275
squared_hinge 51
Test set accuracy: 0.56
hinge 61
Test set accuracy: 0.365
squared_hinge 61
Test set accuracy: 0.565
```

Figure 9: Confusion matrix and test accuracy for case 1 with 200 features using SVM on RFE.




```
hinge 1
Test set accuracy: 0.065
squared_hinge 1
Test set accuracy: 0.215
hinge 11
Test set accuracy: 0.285
squared_hinge 11
Test set accuracy: 0.325
hinge 21
Test set accuracy: 0.495
squared_hinge 21
Test set accuracy: 0.325
hinge 31
Test set accuracy: 0.495
squared_hinge 31
Test set accuracy: 0.33
hinge 41
Test set accuracy: 0.495
squared_hinge 41
Test set accuracy: 0.33
hinge 51
Test set accuracy: 0.495
squared_hinge 51
Test set accuracy: 0.33
```

Figure 10: Confusion matrix and test accuracy for case 1 with 200 features using SVM on PCA.

7.2 Case 2- 100 features

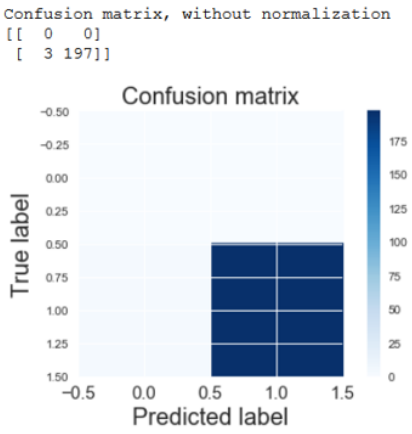
Case 2 has 100 features obtained using RFE and PCA. Figure 11 represents the distribution of features after using RFE as feature elimination technique, features obtained using PCA and correlation graph.

Figure 11: Recursive Feature Elimination, PCA and Pearson correlation for 100 features.

Below are the results obtained after running SVM on the RFE obtained feature. The figures are in the following order: Figure 12 shows the confusion matrix and accuracy.

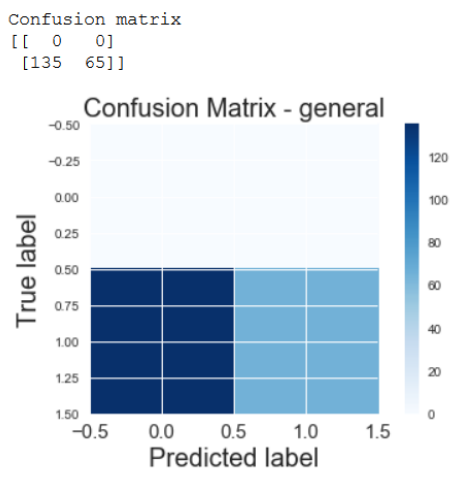
The following results are obtained after performing SVM classifier on PCA features. Figure 13 shows the confusion matrix and obtained accuracy.

After applying LR on RFE, the achieved score is 0.642 and the score on PCA features is 0.638.




```
hinge 1
Test set accuracy: 0.02
squared_hinge 1
Test set accuracy: 0.2
hinge 11
Test set accuracy: 0.02
squared_hinge 11
Test set accuracy: 0.125
hinge 21
Test set accuracy: 0.02
squared_hinge 21
Test set accuracy: 0.72
hinge 31
Test set accuracy: 0.02
squared_hinge 31
Test set accuracy: 0.925
hinge 41
Test set accuracy: 0.065
squared_hinge 41
Test set accuracy: 0.975
hinge 51
Test set accuracy: 0.11
squared_hinge 51
Test set accuracy: 0.975
hinge 61
Test set accuracy: 0.14
squared_hinge 61
Test set accuracy: 0.985
```

Figure 12: Confusion matrix and test accuracy for case 2 with 100 features using SVM on RFE.




```
hinge 1
Test set accuracy: 0.045
squared_hinge 1
Test set accuracy: 0.15
hinge 11
Test set accuracy: 0.175
squared_hinge 11
Test set accuracy: 0.32
hinge 21
Test set accuracy: 0.325
squared_hinge 21
Test set accuracy: 0.325
hinge 31
Test set accuracy: 0.325
squared_hinge 31
Test set accuracy: 0.325
hinge 41
Test set accuracy: 0.325
squared_hinge 41
Test set accuracy: 0.325
hinge 51
Test set accuracy: 0.325
squared_hinge 51
Test set accuracy: 0.325
```

Figure 13: Confusion matrix and test accuracy for case 2 with 100 features using SVM on PCA.

7.3 Case 3- 20 features

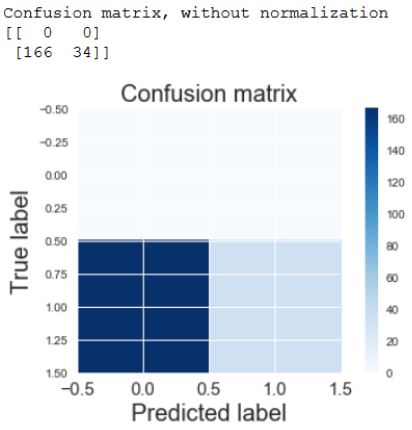
Case 3 has 20 features obtained using RFE and PCA. Figure 14 represents the distribution of features after using RFE as feature elimination technique, features obtained using PCA and correlation graph.

Figure 14: Recursive Feature Elimination, PCA and Pearson correlation for 20 features.

Below are the results obtained after running SVM on the RFE obtained feature. The figures are in the following order: Figure 15 shows the confusion matrix and the obtained accuracy.

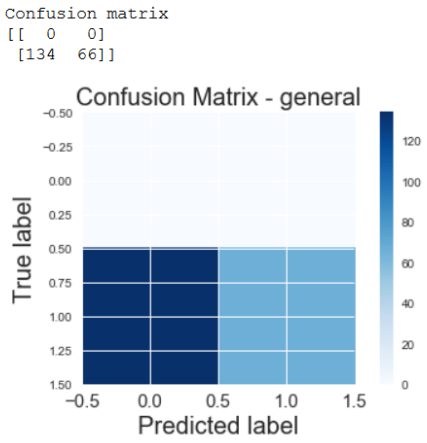
The following results are obtained after performing SVM classifier on PCA features. Figure 16 shows the confusion matrix and the obtained accuracy.

After applying LR on RFE, the achieved score is 0.63 and the score on PCA features is 0.618.




```
hinge 1
Test set accuracy: 0.015
squared_hinge 1
Test set accuracy: 0.05
hinge 11
Test set accuracy: 0.02
squared_hinge 11
Test set accuracy: 0.05
hinge 21
Test set accuracy: 0.02
squared_hinge 21
Test set accuracy: 0.05
hinge 31
Test set accuracy: 0.02
squared_hinge 31
Test set accuracy: 0.05
hinge 41
Test set accuracy: 0.02
squared_hinge 41
Test set accuracy: 0.05
hinge 51
Test set accuracy: 0.02
squared_hinge 51
Test set accuracy: 0.14
hinge 61
Test set accuracy: 0.02
squared_hinge 61
Test set accuracy: 0.17
```

Figure 15: Confusion matrix and test accuracy for case 3 with 20 features using SVM on RFE.

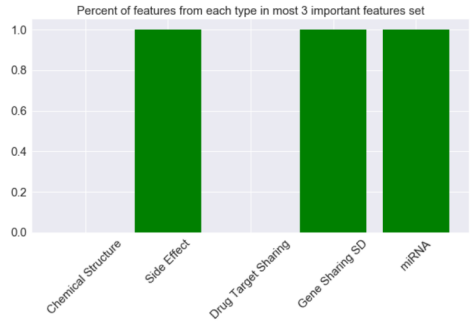



```
hinge 1
Test set accuracy: 0.0
squared_hinge 1
Test set accuracy: 0.015
hinge 11
Test set accuracy: 0.1
squared_hinge 11
Test set accuracy: 0.29
hinge 21
Test set accuracy: 0.295
squared_hinge 21
Test set accuracy: 0.3
hinge 31
Test set accuracy: 0.445
squared_hinge 31
Test set accuracy: 0.33
hinge 41
Test set accuracy: 0.425
squared_hinge 41
Test set accuracy: 0.33
hinge 51
Test set accuracy: 0.425
squared_hinge 51
Test set accuracy: 0.33
```

Figure 16: Confusion matrix and test accuracy for case 3 with 20 features using SVM on PCA.

7.4 Additional- 3 features + 70:30 + 60:40 data split

Additional case has just 3 features obtained using RFE and PCA. Figure 11 represents the distribution of features after using RFE as feature elimination technique and features obtained using PCA. Below are the results obtained after running SVM on the RFE



	Disease-Drug	0	1	2	Class
0	acute lymphoblastic leukemia (all) (precursor ...	6.693368	-1.658225	9.362689	1
1	acute lymphoblastic leukemia (all) (precursor ...	2.295896	0.810448	-6.167723	1
2	acute lymphoblastic leukemia (all) (precursor ...	4.662952	0.021688	4.578135	1
3	acute myeloid leukemia (aml)-imatinib	4.658048	0.006389	4.569565	1
4	acute myeloid leukemia (aml)-imatinib	4.670829	0.027438	4.557820	1

Figure 17: Recursive Feature Elimination and PCA for 3features.

obtained feature. Figure 18 shows the obtained accuracy. Since the accuracy (1%) is very low, proceeding with PCA, LR and Random Forest classifier was waste of time.

The following results are obtained after performing SVM classifier on PCA features. Figure 16 shows the confusion matrix and the obtained accuracy.

Further more, the results from other data splits, that is 70:30 and 60:40 are analysed in this section. Following is the accuracy, precision and recall of using SVM on 70:30 data split along with the confusion matrix and ROC curve displayed in Figure 19.

Followed by the results for LR and RF in Figure 20 and Figure 21 respectively.

Similarly the following results are established for 60:40 data split. Following is the accuracy, precision and recall of using SVM, LR and RF on 70:30 data split along with the confusion matrix and ROC curve displayed in Figure 19, Figure 23 and Figure 24 .

```
Test set accuracy: 0.01
hinge 31
Test set accuracy: 0.01
squared_hinge 31
Test set accuracy: 0.01
hinge 41
Test set accuracy: 0.01
squared_hinge 41
Test set accuracy: 0.01
hinge 51
Test set accuracy: 0.01
squared_hinge 51
Test set accuracy: 0.01
hinge 61
Test set accuracy: 0.01
squared_hinge 61
Test set accuracy: 0.01
```

Figure 18: Test accuracy 3 features using SVM on RFE.

```
Accuracy: 0.6
Precision: 0.309523809524
Recall: 0.125
```

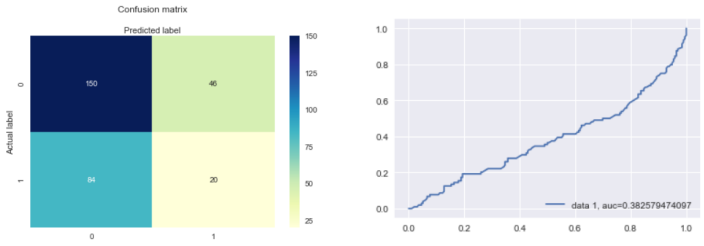


Figure 19: Analysis results for 70:30 data split using SVM.

```
Accuracy: 0.566666666667
Precision: 0.30303030303
Recall: 0.192307692308
```

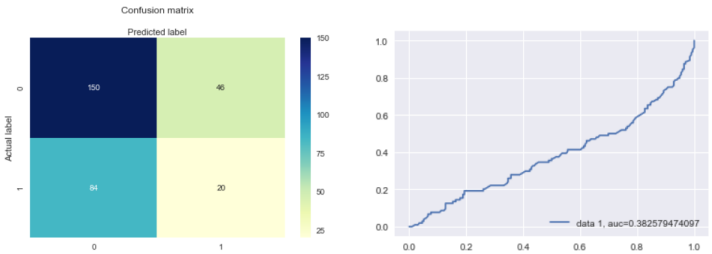


Figure 20: Analysis results for 70:30 data split using LR.

```
Accuracy: 0.453333333333
Precision: 0.211538461538
Recall: 0.211538461538
```

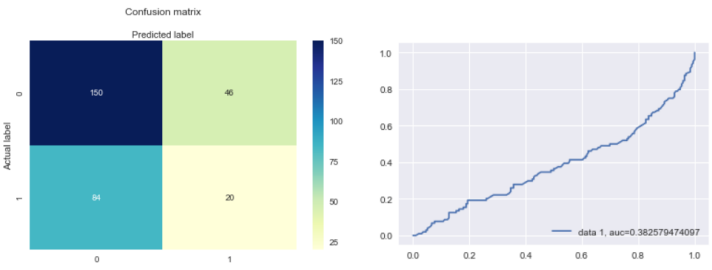


Figure 21: Analysis results for 70:30 data split using RF.

Accuracy: 0.56
Precision: 0.266666666667
Recall: 0.0774193548387

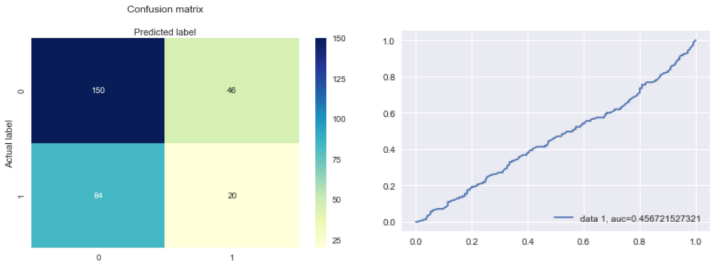


Figure 22: Analysis results for 60:40 data split using SVM.

Accuracy: 0.5275
Precision: 0.297619047619
Recall: 0.161290322581

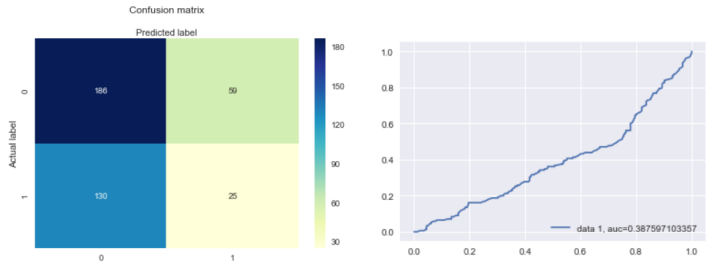


Figure 23: Analysis results for 60:40 data split using LR.

Accuracy: 0.5025
Precision: 0.260869565217
Recall: 0.154838709677

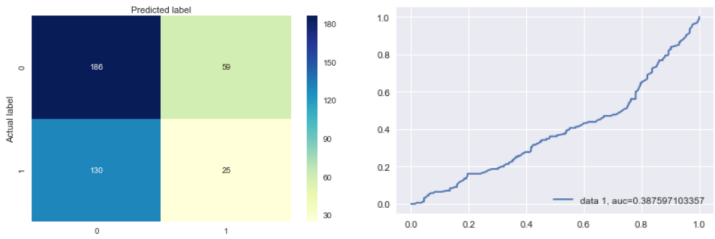


Figure 24: Analysis results for 60:40 data split using RF.

8 Conclusion

This project helped analyse and reason how different feature extraction and binary classifiers effect the accuracy, recall and other evaluation criterias. SVM classifier model produced the best accuracy compared to Logistic Regression and Random Forest in general. Considering the cases, case 2 with 100 features gave the best accuracy of 98.5% where as case 3 with 20 features gave an accuracy of 17%. Additional case of 3 feature gave the least accuracy. This is because case 3 does not have enough features to give a good prediction. The exact count of required features for the best accuracy is a hit and trial case. Evidently, comparing to the baseline mode, the acquired accuracy of SVM is higher which is a success.

On the other hand, case 1 with 200 features have a mediocre accuracy signifying that considering quite a lot of features is not helpful. Logistic Regression classifier gave more of steady accuracy of 60% to 70% for both cases and for both the feature elimination techniques and Random Forest produces a bit lesser accuracy between range 55% to 65%. The provided accuracy account as the average of three data splits, 80:20, 70:30 and 80:20 respectively.

Another aspect of having confusion matrix as evaluation criteria provided the best intuition regarding the success of prediction by the classifiers. SVM for case 2 provided 197 of True Negatives which is good and a little to False Negatives whereas for case 3 it shows 166 towards False Negatives and merely 34 towards True Negatives. The TN says that drug cannot treat the disease and the model has predicted no. The FN says the otherwise. This is very important in identifying the performance apart from accuracy for this drug repositioning problem.

9 Future Work

Considering the fact that this project sustains only for this particular dataset with the provided features, the research experiment could be expanded to perform drug repositioning for various datasets with different features.

More work can be done performing complex analysis such as giving weights to different features based on their importance for various drug-disease combinations. Further more, biological validation is required to for the extracted features for model deployment.

References

[1] Pushpakom, S., Iorio, F., Eyers, P. et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 18, 41–58 (2019). https://doi.org/10.1038/nrd.2018.168 3

[2] Xue H, Li J, Xie H, Wang Y. Review of Drug Repositioning Approaches and Resources. Int J Biol Sci. 2018;14(10):1232–1244. Published 2018 Jul 13. doi:10.7150/ijbs.24612 3

[3] Yella JK, Yaddanapudi S, Wang Y, Jegga AG. Changing Trends in Computational Drug Repositioning. Pharmaceuticals (Basel). 2018;11(2):57. Published 2018 Jun 5. doi:10.3390/ph11020057 4

[4] Napolitano, Francesco Zhao, Yan Moreira, Vânia Tagliaferri, Roberto Kere, Juha D'Amato, Mauro Greco, Dario. (2013). Drug Repositioning: A Machine-Learning Approach through Data Integration. Journal of cheminformatics. 5. 30. 10.1186/1758-2946-5-30. 4, 9

[5] Zhao, Kai So, Hon-Cheong. (2018). Drug Repositioning for Schizophrenia and Depression/Anxiety Disorders: A Machine Learning Approach Leveraging Expression Data. IEEE Journal of Biomedical and Health Informatics. PP. 1-1. 10.1109/JBHI.2018.2856535. 4, 9

[6] Zhang, Ping Agarwal, Pankaj Obradovic, Zoran. (2013). Computational Drug Repositioning by Ranking and Integrating Multiple Data Sources. 8190. 579-594. 10.1007/978-3-642-40994-337. 5

[7] Nagaraj, A., Wang, Q., Joseph, P. et al. Using a novel computational drug-repositioning approach (DrugPredict) to rapidly identify potent drug candidates for cancer treatment. Oncogene 37, 403–414 (2018). https://doi.org/10.1038/onc.2017.328 6

[8] Xue H, Li J, Xie H, Wang Y. Review of Drug Repositioning Approaches and Resources. Int J Biol Sci. 2018;14(10):1232–1244. Published 2018 Jul 13. doi:10.7150/ijbs.24612 5

[9] Talevi, Alan Bellera, Carolina. (2019). Challenges and opportunities with drug repurposing: finding strategies to find alternative uses of therapeutics. Expert Opinion on Drug Discovery. 10.1080/17460441.2020.1704729. 5