Data Mining Project

EMANUELE TARTAGLIONE, 61294
MARCEL BUCZKOWSKI, 59426
ALBERTO FALLOCCO, 59378

This project was intended to draw meaningful insight from biological data using data mining techniques: given the nature of the two dataframes we were provided with we tried three approaches. Firstly, cluster analysis resulted in four clusters being detected, two of which featured interesting similarities.Secondly, a number of significantly strong association rules were found. Lastly, a recommendation system was implemented in an attempt to gain a deeper insight.

ACM Reference Format:
Emanuele Tartaglione, Marcel Buczkowski, and Alberto Fallocco. 2023. Data Mining Project. *ACM Trans. Graph.* 37, 4, Article 111 (August 2023), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The provided datasets contain information about 1101 drugs approved by U.S. Food and Drug Administration. The first one stores instances about molecular fingerprints describing presence of 2048 features for each drug, saved in binary format. The second one features known biological targets for each drug, stored in transaction type format. Both datasets were adequately preprocessed in order to make them suitable for the scope of our analysis. When it comes to cluster analysis, after evaluating a range of different methods, KMeans with 4 clusters displayed the best performances. Upon closer inspection of the two most significant clusters - 0 and 3 - we found common features between the two with respect to most common targets. Mining association rules between targets and molecules - using Apriori - resulted in 19 most important rules after dutiful application of thresholds for support, confidence, lift and conviction. More specifically, some of the rules were found to be surprisingly good with confidence around 0.99. Also variables with similar names were found to have meaningful relationships. Lastly, we implemented a recommendation system to check for similarities between molecules.

2 DATA PREPARATION

As a first step of our analysis we imported the data from provided files and stored them in adequate format. The molecular footprint dataframe is a binary dataframe, and was used in every algorithm we performed; the targets dataframe is a string dataset and it was used to check whether or not similar drugs bind to a given similar target. We investigated whether or not our dataframe featured any missing values in order to avoid setbacks during further analysis: since there we did not find any, we were able to carry on with our analysis with no further preliminary action.

3 CLUSTERING

The first method we tried out to get insight from our data was clustering: our goal was to group drugs from our dataset into subgroups of drugs that have similar properties - and therefore, are likely to exhibit similar activity. As a preliminary step we applied Principal Component Analysis with thirty components to the dataframe in order to reduce dimensionality, then we proceeded to select the best clustering algorithm. For agglomerative clustering we tested different linkage-based methods - and ended up selecting single linkage as it turned out to be the one with the highest silhouette score - then we compared agglomerative clustering and KMeans running them multiple times with different number of clusters, as shown in Fig. 1.

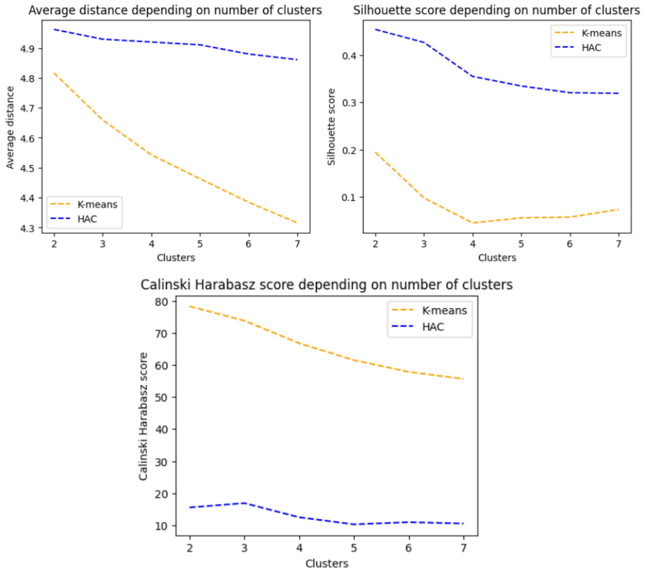


Fig. 1. Scores for Agglomerative Clustering and K-Means

The plots in Fig. 1 are to compare the performances between the hierarchical and KMeans methods with respect to average distance between clusters, silhouette scores and Calinski Harabasz scores. Single linkage seems to be better when it comes to silhouette scores, while KMeans features a better Calinski Harabasz score - and a shorter average distance.

We deemed it important to point out that DBScan and OPTICS were also evaluated, but were not included in the final version of the

Authors' addresses: Emanuele Tartaglione, 61294; Marcel Buczkowski, 59426; Alberto Fallocco, 59378.

0730-0301/2023/8-ART111 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

analysis due to poor performances. Another aspect we thought worthy of mentioning is how PCA affected the models' performances: better results were displayed by KMeans as opposed to all other methods, for which differences turned out to be not significant.
In Table 1 we display the cluster distribution for Kmeans and in Table 2 for HAC.

Table 1. Cluster distribution for KMeans with 4 clusters

cluster	number of drugs falling in that cluster
1	465
2	339
3	227
4	70

Table 2. Cluster distribution for HAC with 4 clusters

cluster	number of drugs falling in that cluster
1	1093
2	3
3	4
4	1

Since HAC resulted in most of the data falling in one cluster, in the end we decided to choose KMeans algorithm with 4 clusters, for which we created the Silhouette plot in figure 2. The plot shows how our clustering analysis did not produce high quality results, even though the scores for cluster 0 and cluster 3 are quite good. Thus, we decided to compare the targets from these two clusters to check for differences.

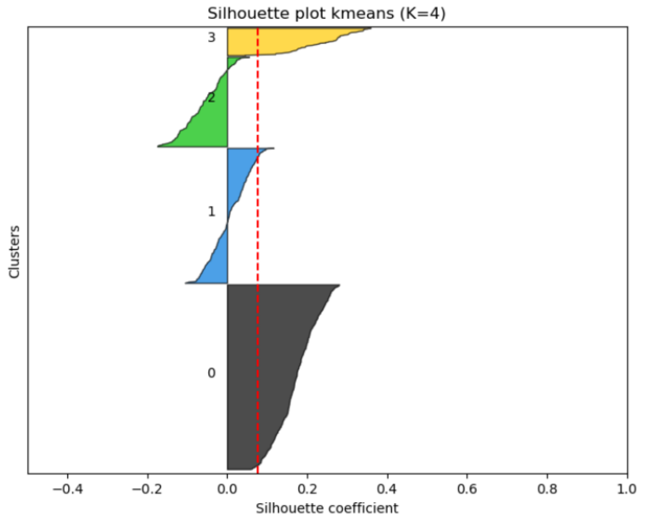
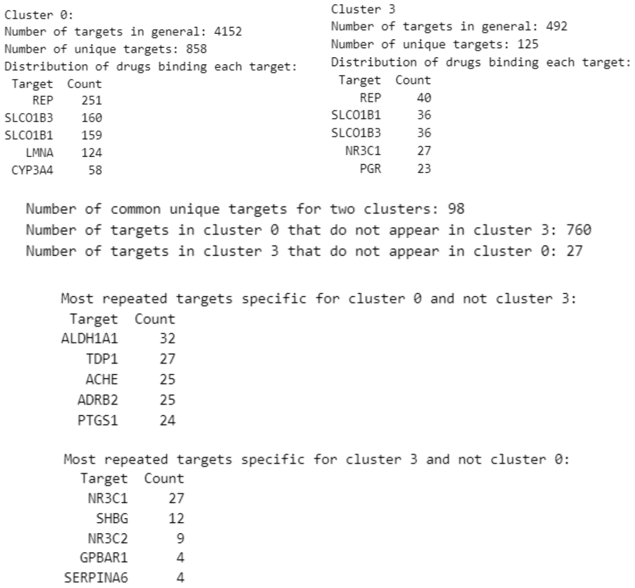


Fig. 2. Silhouette plot for KMeans with 4 clusters



```
Cluster 0:                                        Cluster 3
Number of targets in general: 4152                Number of targets in general: 492
Number of unique targets: 858                     Number of unique targets: 125
Distribution of drugs binding each target:        Distribution of drugs binding each target:
 Target  Count                                     Target  Count
    REP    251                                        REP     40
SLCO1B3    160                                    SLCO1B1     36
SLCO1B1    159                                    SLCO1B3     36
   LMNA    124                                      NR3C1     27
 CYP3A4     58                                        PGR     23

  Number of common unique targets for two clusters: 98
  Number of targets in cluster 0 that do not appear in cluster 3: 760
  Number of targets in cluster 3 that do not appear in cluster 0: 27

      Most repeated targets specific for cluster 0 and not cluster 3:
       Target  Count
      ALDH1A1     32
         TDP1     27
         ACHE     25
        ADRB2     25
        PTGS1     24

      Most repeated targets specific for cluster 3 and not cluster 0:
         Target  Count
          NR3C1     27
           SHBG     12
          NR3C2      9
         GPBAR1      4
       SERPINA6      4
```

Fig. 3. Analysis of clusters' drugs

Afterwards, we find out how many targets there were for the drugs of cluster 0, for cluster 3, how many of them were unique to one cluster, how many they were in common and how many they were not (Fig. 3). This gave us an idea of the degree to which drugs within one cluster are actually similar, and how different these are from drugs in different clusters. Moreover, we were able to draw some conclusions about the distributions of each cluster's targets: we observed that for both clusters the targets of its drugs are often repeated, targets as REP, SLCO1B1 and SLCO1B3 being the most common ones. This seems to suggest that such targets are universal. The figure also shows how unique clusters' targets are with respect to each other, and how distributed are the targets that belong to drugs of one cluster only.
Furthermore, had we been able to validate our analysis with external information, we could have compared new molecules that we would have wanted to examine with the rest of the molecules in our clustering space, to see in which cluster it would fall. This could have given us an idea of what kind of molecules we were working with, and what its suspected properties and activity could have been.

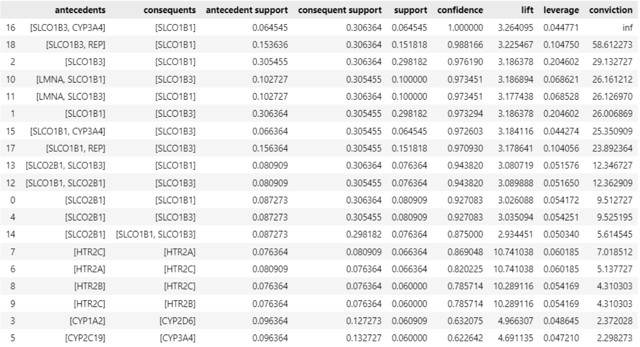
4 ASSOCIATION RULES

We carried out with our analysis by mining association rules: we tried to find association rules between drugs' targets. In order to do so, we first created a binary database using *TransactionEndocder*. In doing so, we were able to run *apriori algorithm* with different support thresholds in order to select the most adequate one. This step is visualised below in Table 3.

We chose a minimum support threshold of 0.06 which resulted in 48 itemsets. Four of them were of maximal length three. We then proceeded to generate association rules, setting a lift and conviction thresholds of 2, in order to select the very best ones. Such constraints resulted in 19 rules being generated. Rules are shown in Fig. 4.

Table 3. Apriori Testing

minimum support	number of itemsets	apriori run time
0.10	15	0.024539
0.09	18	0.008233
0.08	24	0.017475
0.07	34	0.007861
0.06	48	0.016479
0.05	95	0.016447



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
16	[SLCO1B3, CYP3A4]	[SLCO1B1]	0.064545	0.306364	0.064545	1.000000	3.264095	0.044771	inf
18	[SLCO1B3, REP]	[SLCO1B1]	0.153636	0.306364	0.151818	0.988166	3.225467	0.104750	58.612273
2	[SLCO1B3]	[SLCO1B1]	0.305455	0.306364	0.298182	0.976190	3.186378	0.204602	29.132727
10	[LMNA, SLCO1B1]	[SLCO1B3]	0.102727	0.305455	0.100000	0.973451	3.186894	0.068621	26.161212
11	[LMNA, SLCO1B3]	[SLCO1B1]	0.102727	0.306364	0.100000	0.973451	3.177438	0.068528	26.126970
1	[SLCO1B1]	[SLCO1B3]	0.306364	0.305455	0.298182	0.973294	3.186378	0.204602	26.006869
15	[SLCO1B1, CYP3A4]	[SLCO1B3]	0.066364	0.305455	0.064545	0.972603	3.184116	0.044274	25.350909
17	[SLCO1B1, REP]	[SLCO1B3]	0.156364	0.305455	0.151818	0.970930	3.178641	0.104056	23.892364
13	[SLCO2B1, SLCO1B3]	[SLCO1B1]	0.080909	0.306364	0.076364	0.943820	3.080719	0.051576	12.346727
12	[SLCO1B1, SLCO2B1]	[SLCO1B3]	0.080909	0.305455	0.076364	0.943820	3.089888	0.051650	12.362909
0	[SLCO2B1]	[SLCO1B1]	0.087273	0.306364	0.080909	0.927083	3.026088	0.054172	9.512727
4	[SLCO2B1]	[SLCO1B3]	0.087273	0.305455	0.080909	0.927083	3.035094	0.054251	9.525195
14	[SLCO2B1]	[SLCO1B1, SLCO1B3]	0.087273	0.298182	0.076364	0.875000	2.934451	0.050340	5.614545
7	[HTR2C]	[HTR2A]	0.076364	0.080909	0.066364	0.869048	10.741038	0.060185	7.018512
6	[HTR2A]	[HTR2C]	0.080909	0.076364	0.066364	0.820225	10.741038	0.060185	5.137727
8	[HTR2B]	[HTR2C]	0.076364	0.076364	0.060000	0.785714	10.289116	0.054169	4.310303
9	[HTR2C]	[HTR2B]	0.076364	0.076364	0.060000	0.785714	10.289116	0.054169	4.310303
3	[CYP1A2]	[CYP2D6]	0.096364	0.127273	0.060909	0.632075	4.966307	0.048645	2.372028
5	[CYP2C19]	[CYP3A4]	0.096364	0.132727	0.060000	0.622642	4.691135	0.047210	2.298273

Fig. 4. Association rules generated from our dataset

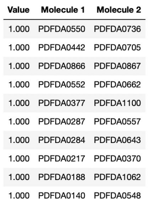
Support measures how often items from antecedents and consequents combined appear in data. Confidence of each rule indicates how often a given rule is observed in data with respect to the antecedents -> consequents relationship. The higher the confidence the more important the rule is. Lift is similar to confidence but this measure controlls also how popular consequents are. Values bigger than 1.0 means that the occurrence of the antecedent has a positive effect on the consequent. Conviction assesses how much the presence of the antecedents influences the absence of the consequents. When less than 1 the presence of the antecedents reduces the likelihood of the consequents. On the other hand when more than 1, the antecedents have a positive influence on the consequents. When conviction equals 1 there is no relationship or dependency between the antecedents and consequents.

From the generated rules we can state out few exemplary conclusions:

- whenever drug binds to SLCO1B3 and LMNA it will also bind to SLCO1B1 with 97% probability.
- there is 100% chance that if drug binds to CYP3A4 and SLCO1B3 it will also bind to SLCO1B1.
- binding to HTR2C increases the chance to bind to HTR2A by 1070%.

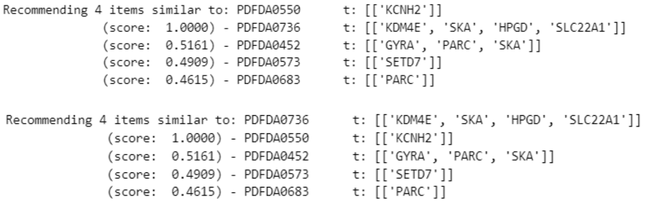
5 RECOMMENDATION SYSTEMS

Lastly, we tried to build a recommendation system that would find the most similar drugs based on their molecular properties. We used our molecular fingerprint dataframe to compute Cosine similarities between drugs; in doing so we were able to find the drugs that are most similar to each other - which are shown in Fig. 5 - and most importantly we were able to build a recommendation system that returns drugs that are most similar to the drug of our interest, and also prints its targets. This allowed us to notice how some of the molecules are identical under the point of view of the features. For instance, molecule PDFDA0550 is very similar to molecule PDFDA0736, and PDFDA0442 is similar to PDFDA0705. We then exploit this information to run recommendation system with a drug of interest to check the targets of similar drugs (Fig. 6). As en examplary analysis with recommendation system we searched for most similar drugs (and its targets) to drugs that were defined in Fig. 5. Because these two drugs - PDFDA0550 and PDFDA0736 - are firmly similar, they also bind to the same drugs having the same similarities. This not only proves the accuracy of our recommendation system but also shows that even though these two drugs do not share any targets in our dataset, maybe some additional molecular research should be held to discover if each drug could bind to its similar drug's targets. In Fig. 6 it is clear to observe that higly similar drugs PDFDA0550 and PDFDA0736 do not share any targets, but maybe it would be worthy to examine if in fact drug PDFDA0736 would not bind to 'KCNH2' target.



Value	Molecule 1	Molecule 2
1.000	PDFDA0550	PDFDA0736
1.000	PDFDA0442	PDFDA0705
1.000	PDFDA0866	PDFDA0867
1.000	PDFDA0552	PDFDA0662
1.000	PDFDA0377	PDFDA1100
1.000	PDFDA0287	PDFDA0557
1.000	PDFDA0284	PDFDA0643
1.000	PDFDA0217	PDFDA0370
1.000	PDFDA0188	PDFDA1062
1.000	PDFDA0140	PDFDA0548

Fig. 5. Most similar molecules



```
Recommending 4 items similar to: PDFDA0550      t: [['KCNH2']]
            (score:  1.0000) - PDFDA0736      t: [['KDM4E', 'SKA', 'HPGD', 'SLC22A1']]
            (score:  0.5161) - PDFDA0452      t: [['GYRA', 'PARC', 'SKA']]
            (score:  0.4909) - PDFDA0573      t: [['SETD7']]
            (score:  0.4615) - PDFDA0683      t: [['PARC']]

Recommending 4 items similar to: PDFDA0736      t: [['KDM4E', 'SKA', 'HPGD', 'SLC22A1']]
            (score:  1.0000) - PDFDA0550      t: [['KCNH2']]
            (score:  0.5161) - PDFDA0452      t: [['GYRA', 'PARC', 'SKA']]
            (score:  0.4909) - PDFDA0573      t: [['SETD7']]
            (score:  0.4615) - PDFDA0683      t: [['PARC']]
```

Fig. 6. Recommendation System

6 LIMITATIONS

This recommendation system can be also used for first hand prediction about the drug of interest that is not stored in our database. Examinating this drug's molecular properties and featuring it into our data space we could find most similar drugs to it and predict the targets it might bind to.

When it comes to exact predicting wheter or not the drug of interest will bind to specific target, it is complicated to implement because we do not have any classes provided, and approches like Naive Bayes Classifier or Decision Trees require that since they are supervised learning methods. What could be done is to create pseudoclasses based on cluster association of each drug, but since clustering quality is rather poor this approach would mostlikely result in inaccurate predictions. Moreover, it is quite brief to assume classes based on clustering, even if in this way drugs will be divided into classes, we cannot assume that falling in specific class (cluster) will be equal with biding to specific set of targets by every member of this class. Thus this analysis would not be very informative when it comes to target prediction.

7 CONCLUSION

To sum up, we applied three data mining techniques to the two datasets: clustering, association rules and recommendation systems. All the methods gave us interesting results for a further analysis when it comes to predict the data. However, when it comes to exact predicting wheter or not the drug of interest will bind to specific target, it is complicated to implement because we do not have any classes provided, and approches like Naive Bayes Classifier or Decision Trees require that since they are supervised learning methods. What could be done is to create pseudoclasses based on cluster association of each drug, but since clustering quality is rather poor this approach would mostlikely result in inaccurate predictions. Moreover, it is quite brief to assume classes based on clustering, even if in this way drugs will be divided into classes, we cannot assume that falling in specific class (cluster) will be equal with biding to specific set of targets by every member of this class. Thus this analysis would not be very informative when it comes to target prediction. The association rules gave us interesting finding about the targets, with the most interesting one that is: every drug that binds to CYP3A4 and SLCO1B3 it will also bind to SLCO1B1. It also seems that the variables with similar "name" have meaningful relationships, since some of the rules are: [SLCO1B3] $\rightarrow [SLCO1B1]$, $[SLCO2B1] \rightarrow [SLCO1B3]$, $[HTR2C] \rightarrow [HTR2A]$, $[CYP1A2] \rightarrow [CYP2D6]$.

The recommendation system can be also used for first hand prediction about the drug of interest that is not stored in our database. Examinating this drug's molecular properties and featuring it into our data space we could find most similar drugs to it and predict the targets it might bind to.