Scuola universitaria professionale
della Svizzera italiana
SUPSI

University of Applied Sciences and Arts of Southern Switzerland
Department of Innovative Technologies

Applied Case Studies of Machine Learning and Deep Learning in Key Areas II

FINAL PROJECT - DRUG LIKENESS PREDICTION

Davide Gamba
davide.gamba@student.supsi.ch

Nathan Margni
nathan.margni@student.supsi.ch

Federico Weithaler
federico.weithaler@student.supsi.ch

Professor: Dr. Gianvito Grasso
SUPSI, Lugano Switzerland

28/07/2023

Table of Contents

1. Introduction

1.1 Context and objective

Drug likeness prediction plays a crucial role in discovering early, which molecules are more likely to possess desired pharmaceutical properties, in order to discard the other candidates. Pushing a molecule through the approval stages is expensive, if we can limit the total candidates before the first step is taken, we can save a lot of time and resources.

"The drug-likeness has been widely used as a criterion to distinguish drug-like molecules from non-drugs. Developing reliable computational methods to predict the drug-likeness of compounds is crucial to triage unpromising molecules and accelerate the drug discovery process." [1]

Molecules can have different structures, sizes, and compositions, and they can have benefits for humans. Considering the vast diversity of chemical compounds and the possibility of synthesizing new ones quickly, with today's technology, it's understandable how important using machine learning as a support tool is fundamental.

Our goal is to build different ML models, and compare them to available models in literature. We have selected two papers, following the suggestions of Prof. Grasso.

- Nature: Minimal-uncertainty prediction of general drug-likeness based on Bayesian neural networks, 2020 [2]
- Chemical Science: Drug-likeness scoring based on unsupervised learning, 2022 [3]

In the first article they put together a single dataset, that they later on split into a train and test set, formed by different sources, containing positive and negative labeled data.

The second article instead, merges some labeled molecules from two sources (2833 Worlddrug, positive + 2833 ZINC15 negative) and tested the models on four different datasets, each with a different distribution of the target label.

1.2 Data Collection

The dataset that we are using is created by merging the positive data from the "fda_approved" molecules dataset, and the negative data coming from the "non_drug_sample" ZINC15 dataset. We will split this dataset into a train and test set, to comp

Dataset download (Zenodo.org)

2. Features

The features that has been tried widely range from fingerprints to descriptors. Starting from the fingerprints the following has been used:

- **Morgan fingerprints**: very common fingerprint with array representation of 2048 binary values
- **Maccs keys**: another common fingeprint, similar to morgan but with only 167 binary values

And then for the descriptors:

- **Mold2**: calculates a wide range of molecular descriptors for chemical compounds, return array with 777 values.
- **RDKIT**: rdkit descriptors generating 208 features.

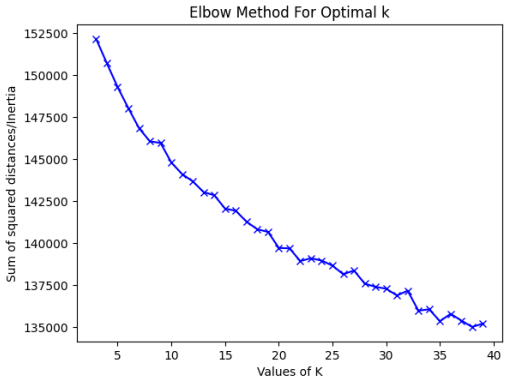

Figure 1: Elbow Method for optimal k

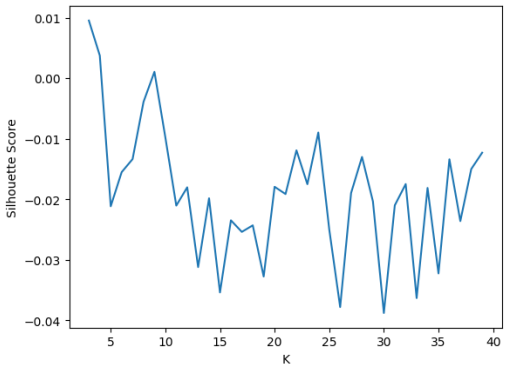

Figure 2: Best K fold silhouette

3. Model buildings

3.1 Unsupervised

Clustering via Kmeans has been performed to check if different cluster could separate the label. Using morgan fingerprints we trained n models to compare the silhouette and intertia score and find the best value for k.

The choosen k is 24, since it have low inertia and a peak of silhouette value.

Plotting clusters with the proportion of labels it can be saw that the cluster are not separating well the label (except cluster 0, 4, 6 and 8) and therefore we do not use this approach to make predictions.

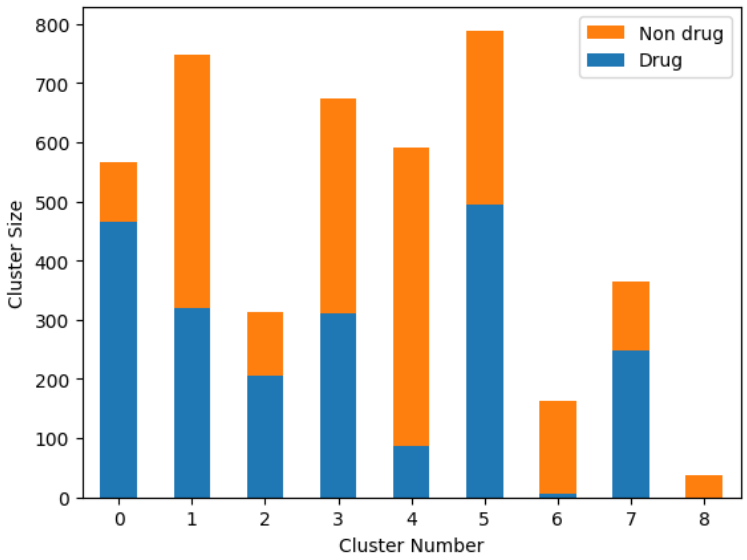

Figure 3: Clusters

3.2 Supervised

We tried a lot of different supervised algorithms, starting from random forest to deep models such as Autoencoders (to generate encoded sequences) and GCNN. For all models many trials with different features as input has been performed, evaluated appropriately with cross validation and tested on a left out test set.

The main metric computed is accuracy, as on the "Minimal-uncertainty prediction of general drug-likeness based on Bayesian neural networks"[2] paper for easier comparison, also confusion matrices, precision, recall and f1 scores are obtained for a deeper evaluation but not included in this paper for simplicity.

3.2.1 Random Forest

Some forest models has been trained with all features, rdkit and mold2 descriptors to visualize feature importance that we could obtain from the model instance:

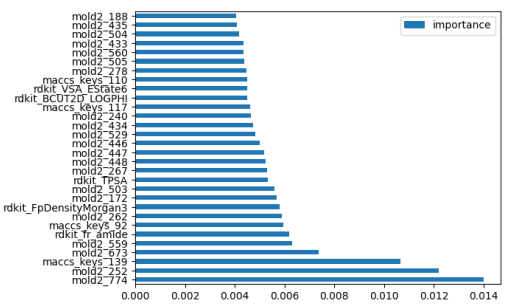

Figure 4: Feature importance among all features

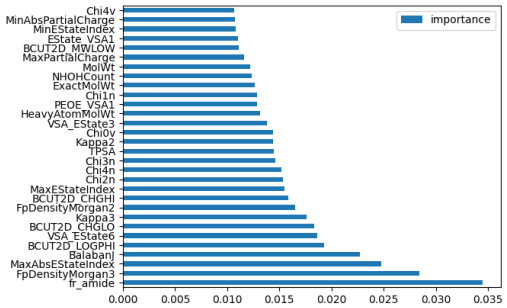

Figure 5: Feature importance among rdkit features

Those inshigts are useful for an expert that know the meaning of the single rdkit feature, for example it shows that fr amide is the most important feature.

3.2.2 MLP

We tried many MLP (feed forward neural networks) models varying the input features and model structure:

- - Trained the first model with all the features we extracted (fingerprints and descriptors). We also performed PCA to extract only 500 principal components in the data and reduce the dimensionality of the dataset, that also improved the model performance, scaled the data and the with the following 3 hidden layers structure: 1000 - 500 - 250 and dropout layers between them.
- Only with the features we computed, 60 - 30 - 15 structure
- With morgan fingerprints, 2000 - 1000 - 500 "
- With maccs keys fingerprint, 100 - 50 "
- With rdkit descriptors, 100 - 50 - 25 "
- With mold2 descriptors, 1500 - 700 - 300

  For the two descriptors we had to cap outliers and perform a robust normalization (Robust scaler) because strong outliers were present, for fingerprints this was not needed because the values were already on a suitable scale.

3.2.3 SVC

We trained a support vector machine model on the morgan fingeprints, since it should a very good algorithm for classification.

3.2.4 AE model

Another approach we tried is to use an Autoencoder model, training encoder and decoder on all features and then use the encoder to generate encoded vectors of size 2000 used as input for a MLP model. Also we trained it with only morgan fingerprints with encoded vectors of size 200.

3.2.5 GCNN with smiles featurized

The last model we trained our data on, it's a Graph Convolution Neural Network, which is particularly well suited for this kind of data. Since it works with nodes and edges, it's easy to understand how atoms can be seen as nodes and chemical bonds as edges. Additional information can also be carried, adding features to as nodes or edges.

We couldn't perform Cross Validation on this model, due to the high computational power needed, since every try was taking more than 50 minutes. This is the reason we just tested this model on the test set. Still, the score is high enough to be good even without performing cross validation.

3.3 Explanability

We used the MLP model with morgan fingerprint to obtain some explanations, for example we tried to create a sample space with exmol library based on the ibuprofen molecule, that the model is very confident in predicting it is as drug. Then with this library we could check for very little changes in the ibuprofen molecule that change the model prediction, that is very interesting:

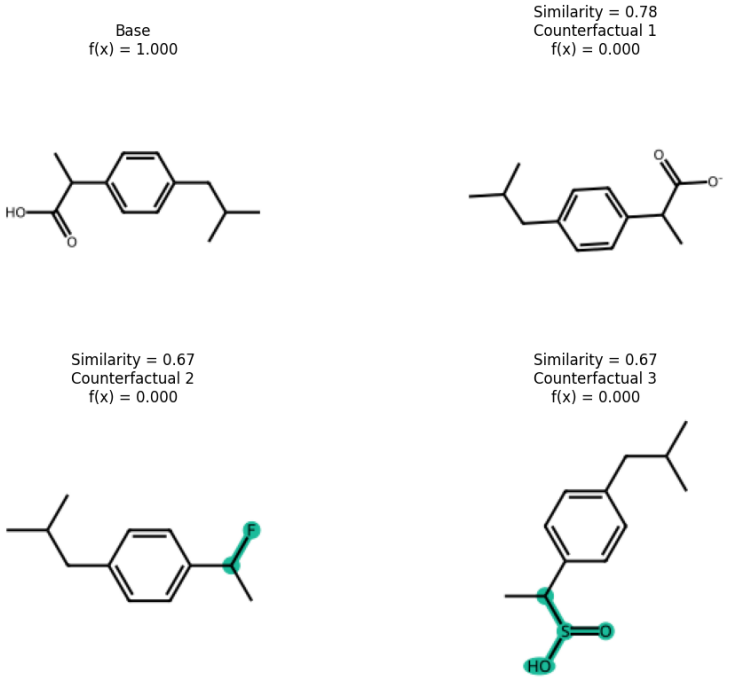

Figure 6: Little changes in the molecule that change the prediction

Also we could plot the entire generated space and its noted that there is not a very clear separation on the predicted labels in 2d components.

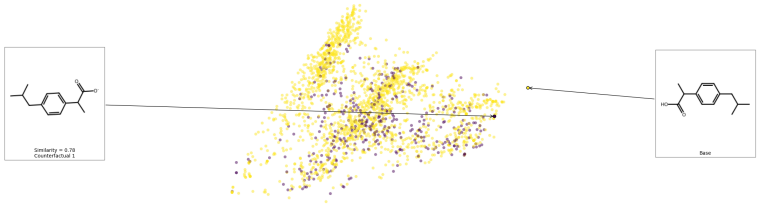

Figure 7: Generated spaced with predicted label hue

Finally we also plotted the descriptor t-statistics values using maccs or ecfp descriptors to obtain the main factors that influence the model in the ibuprofene prediction:

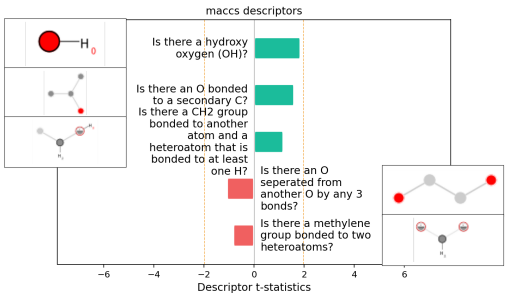

Figure 8: Explanation with maccs descriptors

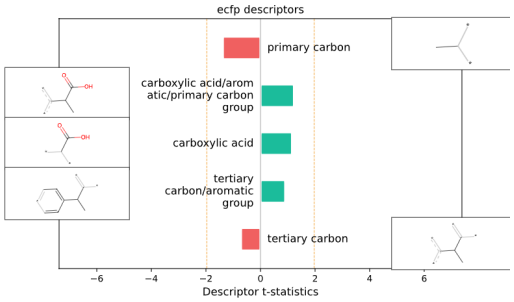

Figure 9: Explanation with ecfp descriptors (Morgan)

This method can be used to analyze why the mode mistakenly predict drugs as non-drugs and viceversa, or to give the researcher interesting information about why a molecule is a candidate to be a drug for the model.

4. Results and Conclusions

4.1 Evaluation and result

4.1.1 Cross validation

For each model we performed Cross Validation with 5 folds. The results are comparable using the same metric score, hence the accuracy on the Cross Validation, or the test set.

4.1.2 Result's table

Table 1: Models Performance

Model	Features	CV Accuracy	Test Accuracy
Random Forest	mold2	85.47	86.72
Random Forest	rdkit	85.97	87.43
MLP	all features and pca	88.76	88.95
MLP	features	70.62	75.56
MLP	morgan fingerprint	86.50	87.43
MLP	mold2	85.65	87.19
MLP	maccs keys	81.74	85.90
MLP	rdkit	85.26	87.07
SVC	morgan fingerprint	86.79	89.19
AE model	all features (encoded)	**88.85**	88.01
AE model	morgan fing (encoded)	85.12	85.78
GCNN	graph features	-	**90.01**

The model that we assume to be the "best" one it the one with highest cross validation accuracy, therefore the AE model with all features encoded, we show below more metrics and confusion matrix of this model:

4.1.3 Comparison Benchmarks

This table shows the original results present in the paper "Minimal-uncertainty prediction of general drug-likeness based on Bayesian neural networks" [2]. We can only compare the results superficially,

```
Mean accuracy in folds:  88.64705920219421

Evaluation on test set:
27/27 [==============================] - 0s 4ms/step
              precision    recall  f1-score   support

         0.0       0.88      0.88      0.88       458
         1.0       0.86      0.87      0.86       393

    accuracy                           0.87       851
   macro avg       0.87      0.87      0.87       851
weighted avg       0.87      0.87      0.87       851

Accuracy on test set:  0.8707403055229143
27/27 [==============================] - 0s 4ms/step
```

Figure 10: Evaluation AE model

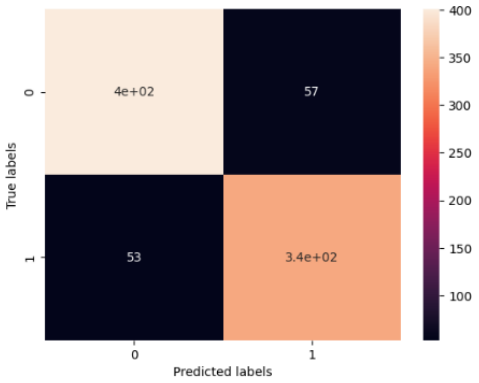

Figure 11: Cunfusion matrix of AE model

as the features used and the models trained may have differences. We also trained some new models, in respect to the ones showcased in the papers, hence, we just provided our results without comparing them.

Table 1 | Comparison of different classifiers

Model	Negative dataset	Input features	Fivefold cross-validation (accuracy, %)	Independent test set
Logistic regression	NOC	QED	48.7 ± 0.6	42.3
	ZINC		60.1 ± 1.2	57.0
	PDB		59.0 ± 0.7	43.5
Three-layer MLP (60 hidden neurons)	NOC	Mold2	80.1 ± 0.3	79.1
	ZINC		90.1 ± 0.2	85.6
	PDB		79.8 ± 0.6	78.2
	NOC	RDKit	79.2 ± 0.5	78.2
	ZINC		87.9 ± 0.5	85.7
	PDB		81.2 ± 0.8	78.9
	NOC	MCS	77.7 ± 0.9	75.6
	ZINC		87.0 ± 0.4	80.8
	PDB		79.6 ± 0.7	73.7
	NOC	ECFP4	78.1 ± 0.9	77.7
	ZINC		88.1 ± 0.4	86.4
	PDB		78.5 ± 0.7	78.2
	NOC	Mol2vec	79.2 ± 0.7	77.2
	ZINC		87.7 ± 0.4	87.1
	PDB		80.3 ± 0.6	72.1
GCNN	ZINC	Atom features from Supplementary Table 2	87.6 ± 0.4	88.5
AE classifier[13]	ZINC	RDKit	88.9 ± 0.6	87.6
AE classifier[13]	ZINC	Mold2	89.9 ± 0.3	87.6

Figure 12: Reference benchmarks for model accuracy

4.2 Conclusions

Their best models AE classifier and MLP with Mold2 descriptors have similar scores in respect to our best models, and they only differ by an additional 1% compared to our scores. Also some models using the same features as our have the same score, meaning that our project reached good results.

Therefore we are satisfied with our work, we managed to test many different models with promising results, we were able to reach accuracy close to the ones of reference. Given the enormous amount of resources available on the internet to gather molecules dataset, we have chosen to use the data already available in the reference paper, to be able to compare the results.

A possible future implementation could be, adding an AutoEncoder model trained on mold2 descriptors, that we didn't manage to add due to time/technical constraints.

With some more attention on the feature selection and on understanding which features are most prone to give good results, we could tweak and tune the parameters to extract even more value from this dataset.

Moreover, the project has provided practical implications for real-world applications, and consolidated our knowledge in the field.

References

[1] J. Sun, M. Wen, H. Wang, *et al.*, "Prediction of drug-likeness using graph convolutional attention network," *Bioinformatics*, vol. 38, no. 23, pp. 5262–5269, Oct. 2022, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac676. eprint: https://academic.oup.com/bioinformatics/article-pdf/38/23/5262/47465922/btac676.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btac676.

[2] W. Beker, A. Wołos, S. Szymkuć, and B. A. Grzybowski, "Minimal-uncertainty prediction of general drug-likeness based on bayesian neural networks," *Nature Machine Intelligence*, vol. 2, no. 8, pp. 457–465, Aug. 2020, ISSN: 2522-5839. DOI: 10.1038/s42256-020-0209-y. [Online]. Available: https://doi.org/10.1038/s42256-020-0209-y.

[3] K. Lee, J. Jang, S. Seo, J. Lim, and W. Y. Kim, "Drug-likeness scoring based on unsupervised learning," *Chem. Sci.*, vol. 13, pp. 554–565, 2 2022. DOI: 10.1039/D1SC05248A. [Online]. Available: http://dx.doi.org/10.1039/D1SC05248A.