Applied Machine Learning

hh3016, Hyeon-Tae “Jack” Hwang
May 08 2025
Github link: https://github.com/Jack-HH6/Applied_ML_Final

1. **Problem and Audience**

The goal of this project is to predict the toxicity of drug compounds using molecular fingerprints to aid in early-stage drug development. Toxicity prediction is critical in drug development to avoid costly failures in clinical trials. The intended audience are chemists focusing on synthesis of drug candidates in pharmaceutical or biotech industries who need scalable machine learning pipelines for chemical safety assessment. This model may help them avoid testing potentially toxic drugs, saving billions of dollars and years of effort that might otherwise be invested in unsuccessful candidates.

2. **Most Impressive Model**

Among all the models tested, the RandomForest model had the highest Precision, indicating that when RandomForest predicts that a drug is toxic, it is more likely to be correct than the other models.

3. **Data**

I used ClinTox provided by MoleculeNet (Wu et.al, 2018), which stores the molecular structure of drugs in SMILES. It also shows drugs approved by the FDA and drugs that failed clinical trials for toxicity reasons. The toxicity of the drugs seem to have been reported through official reports from clinical studies. The dataset doesn't provide detailed mechanistic toxicity measurements such as liver toxicity or cardiotoxicity.

Since the database provides the chemical notation that represents the chemical structure and the status of toxicity, I am planning to generate a feature map that captures the different chemical structure of the molecule and use it to predict the toxicity of the drug. To generate the feature map, Morgan Fingerprints, a commonly used method to create a binary representation of the chemical structure will be used (Sharma et al, 2023). I used 1487 drugs and generated 4096 features per drug.

Methods such as logistic regression, Ridge regression, and Random Forest will be used to predict the toxicity. As mentioned below, the feature matrix shows low covariance and low correlation. Therefore, PCA will not be used.

4. **Method**

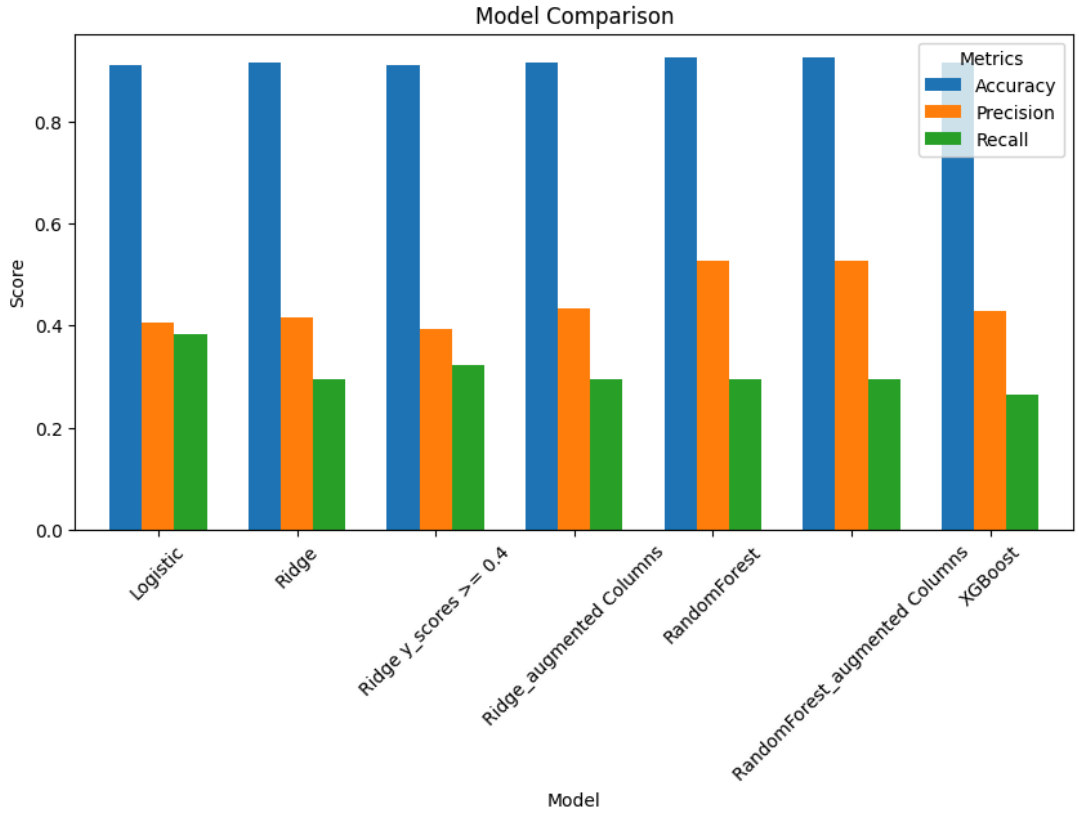
While conducting the data quality check, 4 molecules were excluded from the list because the library function failed to convert the molecule into SMILES. When taking a look at the histogram that shows the frequency of molecules with a certain number of features(substructures), there seemed to be a spike at around 40 features. When analyzing the examples of the molecules that contain these number of features, it didn’t seem like the result was due to a baseline structure, where pharmaceuticals could have used the baseline structure to slightly adjust the structure to have better rates of success. When I googled some of the drugs that are labeled as FDA approved without toxicity concerns, they showed up in the fda report, mentioning that there weren’t toxicity concerns.

Logistic regression was used as a baseline model, since the model has to predict whether a drug has toxicity or not. To tune the hyperparameter, ridge was used. Also, Random Forest and XGBoost were used. To assess how robust the model would be on a different dataset, I performed stratified 3-fold cross-validation on the training data. The correlation between the features was low, so PCA wasn't conducted.

Feature engineering through generating interactive features (multiplying correlations or feature columns that seem important) in ridge and random forest was conducted. After re-training the RandomForest with the augmented feature, one of the features (1226) became more influential (from rank 5 to 2), and one of the augmented features (4100) showed up on the top 20 most important feature box plot. The feature importance distribution became flatter, and it seemed like the features that are less important then 4100 can be discarded. Feature 2763 showed up as a more important feature after adding the augmented features. A prediction of what 2763 feature might represent was conducted in the 'Random_Forest_Feature_Generation,_XGBoost' file. Further study on feature 2763, 378 (seemed important before and after using augmented features), and 1226 was conducted in the 'Comparing_Performance,_Discussion' file.

5. Results

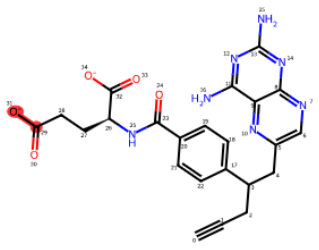
In terms of the Precision, Random Forest performed the best. Comparing it with the logistic regression models, it performed well, so further analysis was focused on the RandomForest model. Considering the limited number of toxic molecules, as shown in the limitations section below, the model performed well. It doesn't seem like the results are due to chance, considering the bias of non toxic molecules in the dataset. The augmentation of the features didn't help the models improve dramatically.



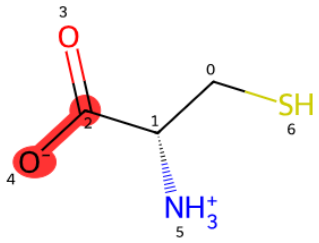
```
=== Bit 2763 ===
Fraction of molecules with this bit by class:
label
0    0.269818
1    0.026786
Name: bit, dtype: float64
Ridge coefficient: -0.1427

=== Bit 378 ===
Fraction of molecules with this bit by class:
label
0    0.234909
1    0.517857
Name: bit, dtype: float64
Ridge coefficient: 0.0467

=== Bit 1226 ===
Fraction of molecules with this bit by class:
label
0    0.193455
1    0.017857
Name: bit, dtype: float64
Ridge coefficient: -0.0849
```

Some of the features that showed strong importance from the RandomForest were analyzed in detail. Based on the analysis conducted above, looking at whether toxic molecules had a tendency of having that certain substructure and the value of Ridge coefficient of that feature, it seems like the O- (feature 2763) and COO- group (feature 1226) are more common in non-toxic groups, while N-heterocycle (feature 378) is associated with toxicity. An example of this analysis is shown below:

Molecule 17 with Bit Index: 1226



Molecule 21 with Bit Index: 1226



6. Limitations

The following are the result for the number of drugs with labels of toxicity:

FDA approved without toxicity concern report: 1375
FDA approved with toxicity concern report: 18
FDA not approved without toxicity concern report: 0
FDA not approved due to toxicity concern report: 94

As shown here, the number of drugs with toxicity is extremely low. For future research, the performance of the model may significantly improve if the number of toxic molecules are included in the dataset.

Other research groups have been using DeepLearning, so alternative approaches may include using deep learning models with more datasets.

7. Citations

Sharma, B., Chenthamarakshan, V., Dhurandhar, A. et al. Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. Sci Rep 13, 4908 (2023). https://doi.org/10.1038/s41598-023-31169-8

Wu, Zhenqin et al. “MoleculeNet: a benchmark for molecular machine learning.” Chemical science vol. 9,2 513-530. 31 Oct. 2017, doi:10.1039/c7sc02664a