COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Frederick National Laboratory
for Cancer Research
sponsored by the National Cancer Institute

Predicting Kinase Inhibitors Using Small-Molecule Structure Information

Jady Tian
Affiliation: IEOR Department, Columbia University
Dec. 14, 2020
https://github.com/jadytian/ML-predict-kinase-inhibitor

Table of content

Background

- The majority of cancers are related to the malfunctioning of a group of enzyme proteins called kinases
- Kinase Inhibitors: certain drugs binding to kinase proteins and inhibit their activity
- They serve an important and unique role such as metabolism, cell-growth/differentiation, etc
- Side effects happen when drugs bind to non-intended receptors/kinases

Resources

- DeepChem Tutorials
- PubChem
- DrugBank
- Ravi's Github
  - https://github.com/ravichas/ML-predict-drugclass

Kinase Inhibitor Selectivity Project Overview

Objective

- Build a machine learning model for kinase target selectivity
  - Identify if a drug is a kinase inhibitor
  - Identify if a drug is a particular kinase inhibitor
  - Able to predicting ki (how strongly does a drug bind to a target)

Process

- Generate molecular properties
  - Molecular Fingerprints
  - Mordred Descriptors
- Supervised Learning using Random Forest Classification
  - Random Forest
  - Hyperparameter Optimization
  - Regular and KFold Random Forest Models

Data Collection

- Source from PubChem
  - Drugs category
  - Kinase inhibitor tags
- 130480 data entries consisting of kinase inhibitors(203) and other drugs(130277)
  - Down-sample the imbalanced dataset into 203 data entries for each category
  - **Columns:** drug name, SMILES string, whether or not the drug is a kinase inhibitor

	SMILES	Inhibitor
0	CC(C)C[C@@H](C(=O)N[C@@H](CC1CCNC1=O)C(O)S(=O)...	No
1	CCOC1=CC=C(C=C1)N=NC2=CC(=C(C=C2)/C=C/C3=C(C=C...	No
2	CN(C1=CC(=CC=C1)OCF)C(=NC2=C(C=CC(=C2)SC)Cl)N	No
3	CCN1C2=C(C=C(C=C2)S(=O)(=O)O)C(C1=CC=CC3=[N+](...	No
4	CC(C)C[C@@H](C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](C...	No
5	CCCC(C(=O)CC)C(=O)SCCNC(=O)CCNC(=O)[C@@H](C(C)...	No

Modelling

- Generate molecular properties
  - **Molecular Fingerprints:** series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule
  - **Mordred descriptors:** another way of representing molecular characteristics

```
[[0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
 [0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0]
 [1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]]
```

	ABC	ABCGG	nAcid	nBase	...	Zagreb1	Zagreb2	mZagreb1	mZagreb2
0	31.697605	23.233843	0	0	...	194.0	202.0	18.7917	10.250000
1	18.372846	14.479004	0	0	...	122.0	142.0	7.27778	5.500000
2	42.135164	31.576535	2	0	...	294.0	353.0	20.5972	11.083333
3	13.980375	10.324630	0	2	...	98.0	120.0	3.41667	3.611111
4	10.744501	10.170579	0	0	...	66.0	72.0	5.58333	3.638889
..	...	...	...	...	...	...	...	...	...
401	9.289847	8.749080	0	0	...	62.0	73.0	3.94444	2.666667
402	26.208040	19.318438	0	1	...	178.0	212.0	8.97222	7.194444
403	23.148902	19.019093	0	1	...	162.0	200.0	7.97222	6.361111
404	10.094413	9.759293	0	0	...	70.0	86.0	5.41667	2.833333
405	28.749097	21.532616	0	1	...	190.0	216.0	13.9236	8.625000

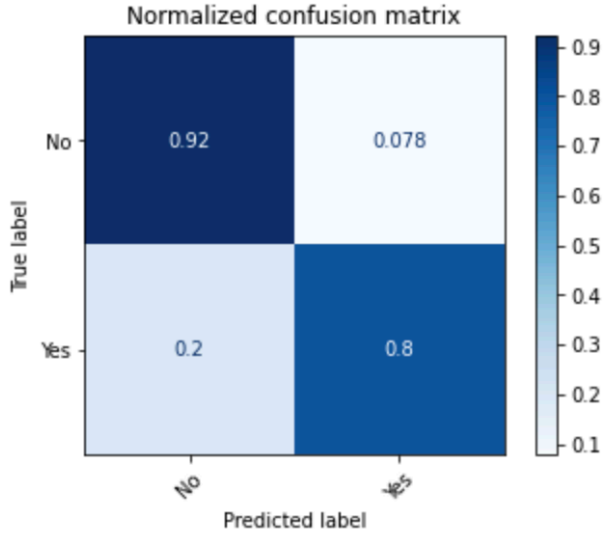
Modelling

- Supervised Learning using Random Forest Classification
  - **Random Forest:** a classification algorithm that consists of many decisions trees and makes predictions based on an uncorrelated forest of trees
  - Use **hyperparameter optimization** to find the most optimal set of parameters
  - Run **regular and KFold** random forest models properties using molecular properties molecular
    - Regular – training, testing versus KFold – training, testing, validation

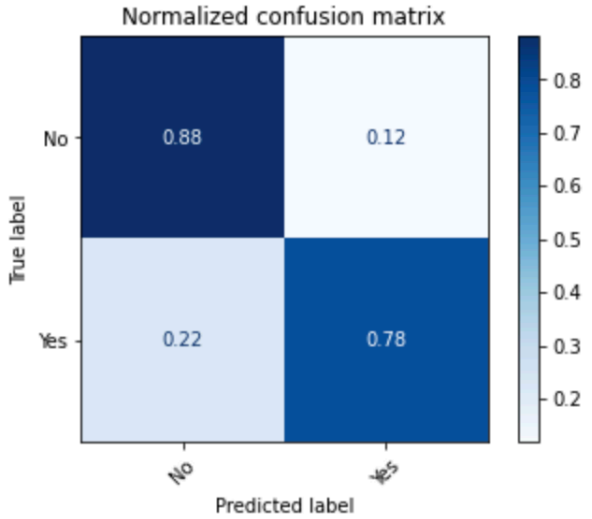
Parameter	Values
n_estimators	250
max_features	sqrt
min_samples_leaf	1
class_weight	None

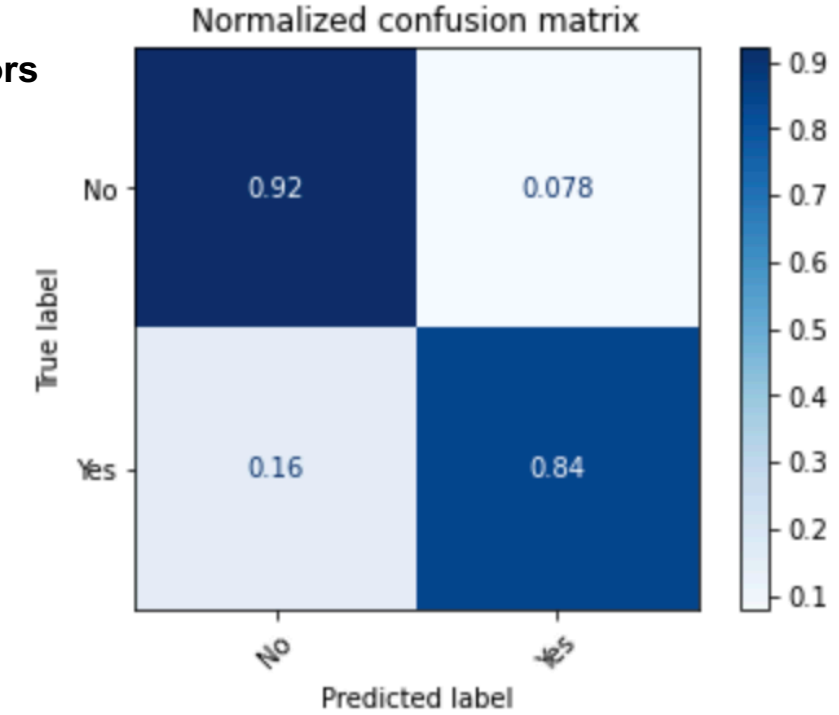
Results

Regular Random Forest with Molecular Fingerprint

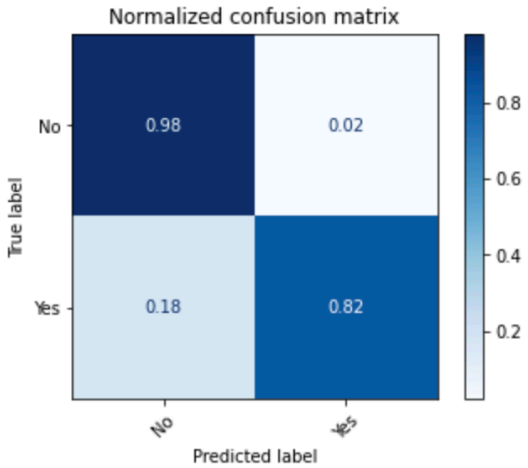


KFold Random Forest with Mordred Descriptors





KFold Random Forest with Molecular Fingerprint



Regular Random Forest with Mordred Descriptors

Next Steps

- Perform the best Random Forest model on several specific kinase inhibitors
- Construct similar model predicting ki (how close a kinase inhibitor binds to its target) of kinase inhibitors

THANK YOU!

Thank you for this opportunity, especially to Ravi and Naomi. We learned about FNL and about collaborating to solve real machine learning problems in medical science. We've learned a great deal and hope to be able to work with FNL again.

Any questions?