

W



Technology Review

Formulation group

Zichen Zhu; Chenyi Mao; Lin Zhang; Dawei Gu; Jing Xu

1. Background

- Dataset: 927 drugs includes m
- Purpose: Build a model can pr
- pharmaceutical indigent(API).
- Properties of Drugs: solubility weight, etc...
- Significance: simulate the form
- Also, this model could potentia



Pharmaceutical
engineering

ve

molecular

situation.

ng value

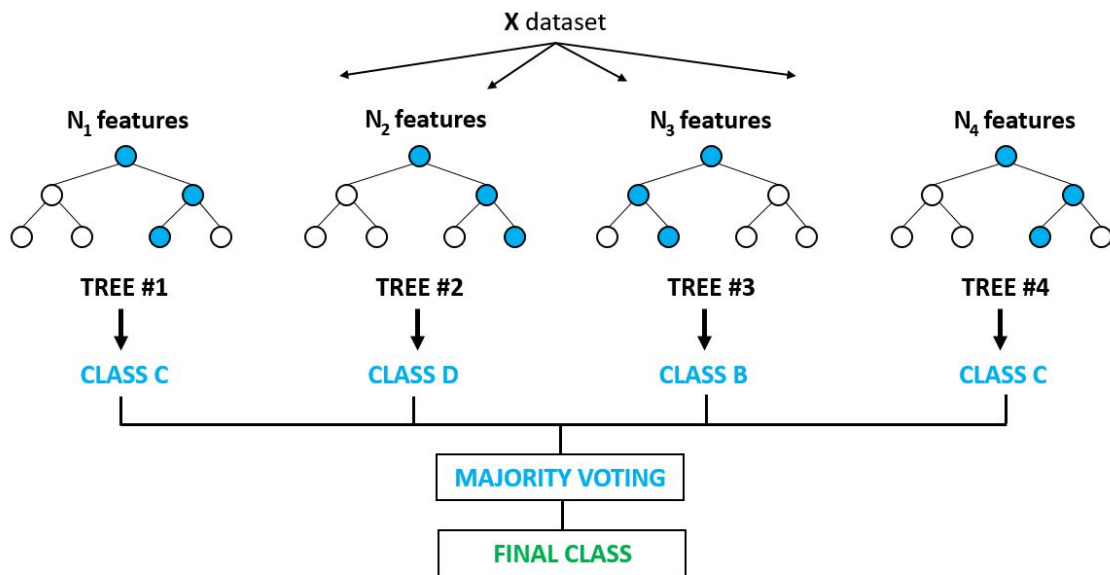
2. Lists of available technologies

1. K Nearest Neighbor (value of parameter K)
2. Support Vector Machines (two-class, sparse data)
3. Random Forest (multiclass, heterogeneous features)
4. Neural Networks (fancier but more complicated)

	Generic Name	Formulation	Measured Solubility (mg/mL)	MW Drug (g/mol)	Measured LogP	HBA	HBD	PSA (\AA^2)
1	Abacavir Sulfate	tablets	77	286.34	1.20	6	3	95.80
2	Acarbose	tablets		645.62	-8.83	19	14	351.80
3	Acebutolol Hydrochloride	capsules		336.43	1.71	5	3	97.05
4	Acetaminophen; Paracetamol	tablets	23.7	151.17	0.20	2	2	55.41
5	Acetohexamide	tablets	3.43	324.40	2.44	4	2	103.22
6	Acetylsalicylic Acid; Aspirin	tablets	10	180.16	1.19	3	1	68.21
7	Adenosine	solution	5	267.25	-0.98	8	4	135.44



3. Intro to Random Forest Classifier



- ❖ Aggregating a number of binary decision trees.
- ❖ Each tree is fitted to a bootstrapped training set.
- ❖ When fitting each tree, at each split, choose a sample of m predictors as split candidates.
- ❖ To aggregate the prediction results, take the majority.
- ❖ Sklearn.ensemble.
RandomForestClassifier



4. Appeal of Random Forest Classifier (RFC)

Why do we choose RFC:

- It provides high accuracy.
- It is very simple to get started.
- It has the power to handle a large data set with higher dimensionality.

Why do we choose Scikit-learn package:

- Because that's the only package that I could find has built-in RFC module.
- Almost every other ML package is doing NN only.



5. Drawbacks of Random Forest Classifier

- Random forest models are not that interpretable, more like black boxes
- When dealing with large data sets, the huge size of the trees would take up a lot of memory
- Random forest models can tend to overfit



Backup



W

