Machine Learning in Drug Discovery and Design

Predicting the Blood-Brain Barrier Penetration of Drugs

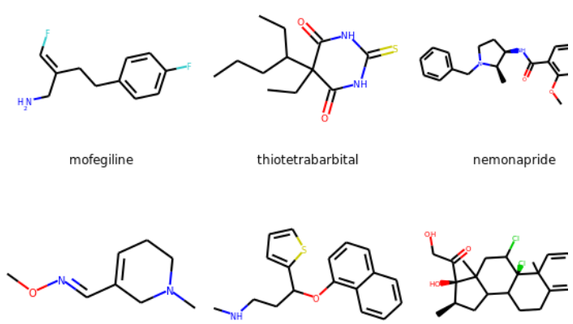
Ganiyu Ibraheem, Bsc(Hons) Software Engineering

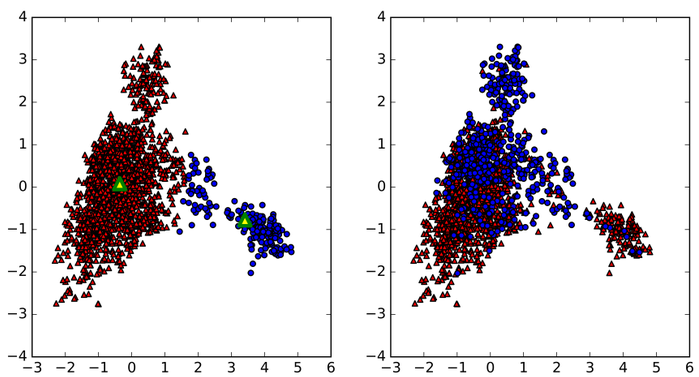
Abstract

Drug discovery and design is a very expensive process and lots of new compounds are being developed rapidly. Only roughly 2% of Central Nervous System (CNS) drugs can pass through the blood-brain barrier, this presents a problem in CNS drug development.

This Project presents a Machine Learning Classifier that can predict with high accuracy the probability of a drug passing through the blood-brain barrier.

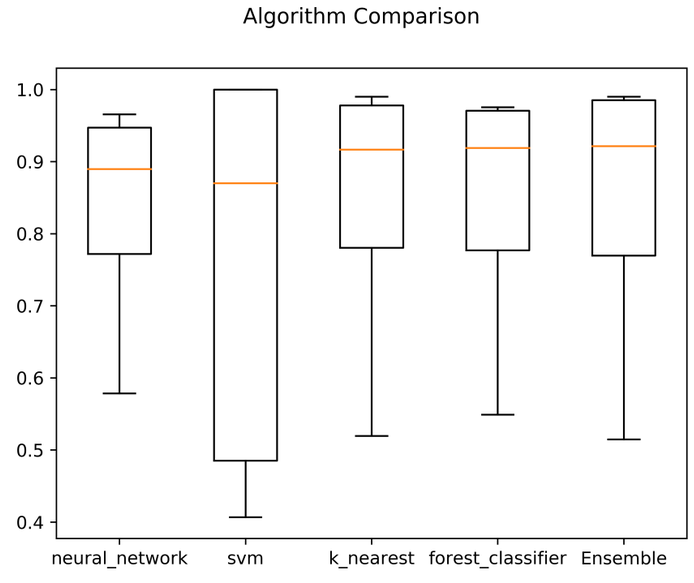
The blood-brain barrier dataset is loaded, parsed and processed for use

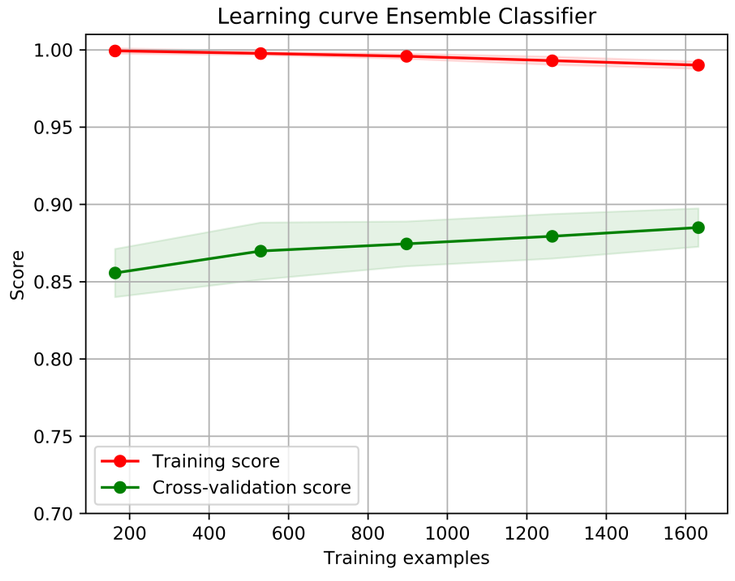




Unsupervised learning techniques are then applied to cluster the molecules extracted from the dataset.

The goal is to gain insights into the dataset to determine if its feasible to separate between the molecules that pass through the barrier and the ones that do not.





The end product is a REST API endpoint where prediction results for a given molecule can be requested. A sample curl call is shown below and the prediction result returned as a JSON is also shown.

```
curl -H "Content-Type: application/json" -X POST -d
'{"smile":"Cn1c2CCC(Cn3ccnc3C)C(=O)c2c4ccccc14"}'
http://localhost:5000/api/prediction
```


```
{
  "category": "p",
  "probability": {
    "n": 0.07729955064888563,
    "p": 0.9227004493511144
  },
  "smile": "Cn1c2CCC(Cn3ccnc3C)C(=O)c2c4ccccc14"
}
```


A neural network, support vector machine, k-nearest neighbour classifier and a Random forest classifier and also an Ensemble classifier consisting of all the previously mentioned models are trained with the blood-brain barrier dataset. The training results showing their prediction accuracies are presented above and also the performance of the Ensemble classifier as the number of training examples increases is also shown above.