ACM.

3. Flavian C, Gurrea R & Orus C. Web Design: A Key Factor for the Website Success. Journal of Systems and Information Technology, 11 (2) pp. 168-184, 2009.

COMPARISON OF CLASSIFICATION QSAR MODELS TO CALCULATE BLOOD-BRAIN BARRIER (BBB) PROPERTIES RELATED TO ADME/T

Serhiienko O., student

Gubaryeva O., PhD, Associate Professor

Kharkiv National University of Radioelectronics

Many experimental methods in drug development are costly and time consuming.

Methods of ligand based drug design can help to overcome a variety of problems. These methods include the prediction of such properties as absorption, distribution, metabolism, excretion and toxicity, also called ADMET. It makes it possible to filter molecules and select those that satisfy a specific criteria of the research.

Therefore, in order to facilitate the rapid and inexpensive (low-cost) profiling tool, various in silico tools have been developed. Among them are QSAR models that have been established to help predict a number of ADME/T properties for new chemicals.

However, despite the fact that there is a diverse range of simulation algorithms, each approach has its advantages and disadvantages, and it’s imperative to find a balance that satisfies the requirements of performance and reliability.

During this work it is necessary to complete the following tasks:

– implement and compare the classification QSAR models for Blood-Brain Barrier prediction;

– make a conclusion about the QSAR models that provide better performance and efficiency.

Implementation of the model consists of the following steps:

– Data preprocessing.

We have formed a dataset that contains 2 columns for SMILES (Simplified molecular-input line-entry system) and BBB activity value which takes the value equal to positive or negative. Second step implies the dataset cleaning. The NAN (not a number) values have been pulled out. Kekulization, normalization, reionization, neutralization and tautomerization also have been performed for achieving better results. As a feature vector, Morgan Fingerprints (Circular Fingerprints) have been chosen and calculated using a Python library named rdkit.

– Model.

For categorical data, in our case for BBB activity value, it is desirable to perform One-Hot Label Encoding so as to convert information into a format that may be fed into machine learning algorithms to improve prediction accuracy. After that, we performed 10 fold cross validation. Among the advantages of such an approach are the use of all data given, opportunity to obtain more metrics, models stacking and parameters fine tuning. Random Forest Classifier and Support Vector Machine algorithms have been taken for the survey.

In order to evaluate the results basic metrics such as f1-score, precision, recall, ROC AUC score were calculated. Received ROC curves are presented below (fig. 1).

The ROC AUC score makes it possible to understand how efficient the model is. The higher the AUC, the better the model's performance at distinguishing between the negative and positive values. Based on this, we are coming to the conclusion that a model built with the use of the Support Vector Machine algorithm is better.

References:

1. Dhiraj K. Advantages and Disadvantages of K fold cross-validation [Електронний ресурс] / Dhiraj. – 2021. – Режим доступу до ресурсу: https://dhirajkumarblog.medium.com/advantages-and-disadvantages-of-k-fold-cross-validation-5e833009ddb1.

2. Druglikeness analysis [Електронний ресурс] – Режим доступу до ресурсу: https://admet.scbdd.com/home/interpretation/.