Comparison of Different Machine Learning Algorithms for Solubility Prediction

Authors : Muhammet Baldan, Emel Timuçin

Abstract : Molecular solubility prediction plays a crucial role in various fields, such as drug discovery, environmental science, and material science. In this study, we compare the performance of five machine learning algorithms—linear regression, support vector machines (SVM), random forests, gradient boosting machines (GBM), and neural networks—for predicting molecular solubility using the AqSolDB dataset. The dataset consists of 9981 data points with their corresponding solubility values. MACCS keys (166 bits), RDKit properties (20 properties), and structural properties(3) features are extracted for every smile representation in the dataset. A total of 189 features were used for training and testing for every molecule. Each algorithm is trained on a subset of the dataset and evaluated using metrics accuracy scores. Additionally, computational time for training and testing is recorded to assess the efficiency of each algorithm. Our results demonstrate that random forest model outperformed other algorithms in terms of predictive accuracy, achieving an 0.93 accuracy score. Gradient boosting machines and neural networks also exhibit strong performance, closely followed by support vector machines. Linear regression, while simpler in nature, demonstrates competitive performance but with slightly higher errors compared to ensemble methods. Overall, this study provides valuable insights into the performance of machine learning algorithms for molecular solubility prediction, highlighting the importance of algorithm selection in achieving accurate and efficient predictions in practical applications.

Keywords : random forest, machine learning, comparison, feature extraction
Conference Title : ICBBE 2024 : International Conference on Bioinformatics and Biomedical Engineering
Conference Location : Toronto, Canada
Conference Dates : June 13-14, 2024

Open Science Index, Biomedical and Biological Engineering Vol:18, No:06, 2024 publications.waset.org/abstracts/186745.pdf