Abstract

Humans have been battling the age-old battle against illness for centuries. Only recently have humans made significant progress in averting illness through the creation of drugs and the field of drug discovery. To create a drug and deem it usable requires a rigorous process that examines one attribute in particular: safety. Unfortunately, ninety percent of drug development fails despite the many successful strategies used. Current chemical methodologies rely on a hit-and-miss approach where large amounts of drugs are analyzed for their properties by hand. This process is expensive, time-consuming, and often inaccurate. This project aims to address this through the implementation of various machine-learning algorithms that can predict the IC50 and pIC50 values of possible drug candidates. These values are essential in determining the quantity of a drug needed to inhibit a biological process by half but also are important in determining the toxicity of a drug and how it impacts patients. Using regression-based machine learning models and bioactivity data of compounds and target proteins these values can be predicted and outputted. Upon completion of this project, we developed multiple regression models for the target protein of the SARS coronavirus. After statistical analysis, the best model was chosen: DecisionTreeRegressor. This model had a root mean squared error score of 0.34 and an R-squared score of 0.90. This implies that this model fits the situation and makes accurate predictions. We conclude that drug discovery can become quicker, more accurate, and cost-effective through the implementation of machine learning algorithms.