The vitamin B transport inhibition data mining task was a binary classification one. The output label, “Inhibition,” has two discrete values: 0 for no inhibition, 1 for inhibition. The input features used for training the pipeline, all of which were numerical, were preprocessed by dropping columns with 60 percent or more null values, filling in positive and negative infinite values and null values with 0 (imputation), clipping very large and small values to fit float32 and int32 data types for consistency with custom transformers’ requirements, and scaled for compatibility with selected models. Moreover, input features were selected using recursive features elimination using cross validation (RFECV) with random forest as its model.

Model selection was rooted in data transformations, such as scaling numerical input features. The best performing model was chosen based on two evaluation metrics and the baseline model: accuracy and area under the receiver-operating characteristic curve (AUC); dummy classifier with its classification algorithm being accuracy.

There are three major limitations in the data mining process for vitamin B data. The training data size is relatively small, about 1,300, meaning that there may not be enough data to generate a high-quality model. Similarly, because of the data imbalance—there is a majority class which consists of about 80 percent of the training data. Models may overfit to one label and the predictions may not be effective. Using data augmentation, however, could potentially improve model quantity and quality and thus, predictions. Finally, the approach taken for this task was binary classification. Using an alternative approach, such as anomaly detection, may lead to better performance and predictions given the problem domain.