OPTION 1
GROUP MEMBER: Aubrie Pressley
GROUP MEMBER: Lisette Kamper-Hinson

The goal of this assignment was to develop a machine learning model pipeline that predicts the inhibition values of molecules. Our pipeline will handle cleaning of the data, anomaly detection and the best regression model chosen from our test regression models. It was decided that this was a regression task because the target variable, “Inhibition” is a continuous and numerical variable.

In Phase 3, all columns were first converted to numeric format, then missing values were imputed, and extreme infinite values were capped last. The data had mixed types therefore converting all to numeric ensured uniformity. Additionally, the models can only handle numeric data. Values that could not be converted to numeric, were given a NaN value, which were later imputed. NaN/missing values were chosen to be imputed by using the column mean. The quality report showed that all of the rows in the published data had missing values, therefore, rows could not simply be removed. The quality report also showed that an insignificant amount of values in each column were missing. Therefore, the missing values transformer checks if the column has a certain percentage of missing values. If it does, then the column will be removed. However, if it does not, then the values are imputed because there is enough information for imputation to be trusted. Finally infinite values were capped at 10^9 to avoid computational issues. The columns “gmin”, “CID”, and “Name” were removed from the training data as the first step in data preprocessing. “Name” and “CID” were chosen to be removed because these should not be features that affect the “Inhibition” prediction. Additionally, “Name” and “CID” are not present in the new molecules data. The column “gmin” was removed simply because even after imputing and capping values, the “gmin” column was making the entire model fail. Multiple approaches were taken to try and solve this problem, however, removing the column was the only solution that worked.

To improve data quality, anomaly detection was employed. In this assignment we chose Isolation Forest because it uses random decision trees to detect anomalies. Reading in this method was provided as well as an example therefore was the most familiar technique for us to use. A transformer to remove anomalies is only applied to the training data, to help improve performance for later predictions. In the prediction data and the new molecules data, the transformer was not used, however, a function was applied to simply identify which out of the 30, 000 molecules are outliers. When generating the ranked molecules report, the molecules that were identified as outliers were given a “UNK” value instead of their predicted value because their predictions may not be accurate.

For model selection, three models were used. Linear Regression was used as a baseline and then compared to the performance of Random Forest and Gradient Boosting. Random Forest and Gradient Boosting handle data with non-linear relationships well therefore was a good option when comparing performance of Linear Regression. The best model was selected based on the resulting MSE from the dev set. The lowest MSE determined our best model. MSE was chosen as

the evaluation metric because MSE is a good choice when wanting to penalize larger errors more than smaller errors, and when the target variable is normally distributed.

Linear Regression was included as a baseline, while Random Forest and Gradient Boosting were chosen due to their ability to handle complex, nonlinear relationships. A pipeline was created that combined the preprocessing steps and the regression model to ensure consistency during both training and testing. The best model was selected based on its MSE, minimizing prediction errors while avoiding overfitting. After the best regression model was chosen, the pipeline was retrained with the entire published data. The best-performing model, along with its full preprocessing pipeline, was then saved using `joblib`, ensuring that it could be applied directly to unseen data.

The next step of this assignment was to generate the ranked_predictions.csv. To complete this the model pipeline was loaded and the new_molecules.csv was processed through the pipeline following the same transformations as the training and dev set did (except for the anomaly detection transformer explained above). The final predictions were saved to a CSV file with their corresponding inhibition scores.

The final step of the assignment was to find which features, if any, differ significantly among the top 100 and bottom 100 ranked molecules. This analysis helps in understanding which molecular properties are most influential in determining the inhibition scores. To achieve this, we performed a statistical comparison between the features of the top and bottom ranked molecules. First the ranked predictions were merged with the original feature data from new_molecules.csv based on the CID column. The same preprocessing transformations used during training were applied to ensure consistency in feature representation. For each feature, a two sample t test was conducted to compare the feature values between the top and bottom groups. A p value threshold of 0.05 was used to determine statistical significance. Features with p values below 0.05 were considered significantly different between the top and bottom molecules. These significant features were saved in a CSV file (significant_features.csv).