The Toxicity Prediction Challenge

Report by

Ruthvik Sairam Jaini

Student ID: 202002268

Email: x2020dmu@stfx.ca

Department of Computer Science

Data-Preparation

- In Train and Test Datasets the data was split in ‘chemical_Id’ and ‘assay_Id’ by using split function that was delimited by “;”.
- In ‘V15’ column the infinite values were replaced by mean of the column.
- For Data Parsing the train and test dataset files were merged into feamat dataset file.
- In Feature Scaling used Min-Max scaler and each element in the features was transformed to specific range.
- The scaled data was fitted into the feature Matrix.
- For Feature Selection used Select-K-best feature selection method to select the top features for the model and used top 1000 features.
- Graphical representation of top features using Matplotlib.
- In Train data dropped ‘chemical_id', 'Id','V2', 'Expected'.
- In Test Data dropped 'chemical_id', 'x', 'V2'.

Model training/testing

- The model used was XGBoost.
- The parameters used were XGBoost “n_estimators” and “max_depth”.
- For Validation stratified k fold method is used to split the data.
- The number of splits used was 10. (K = 10)
- For evaluation of the score, Avg Accuracy score and F1_score was used.
- The best accuarcy was given by XGBoost with n_estimators = 250, Max_depth = 12 using top 1000 columns.

Tried Methods

- For feature selection Principal Component Analysis (PCA) and feature importance methods were used but it did not give a better accuracy when compared to Select K best features method.
- Tried Using Grid SearchCV for selecting best parameters but it was taking a lot of computational time.

- For preprocessing tried using Standard Scaler and robust Scaler but Min-Max scaler methods gave a better result.

Challenges

- Since the dataset was too huge, feature Selection was one of the biggest challenges I faced in this project.
- Choosing the right parameters for the models to get better accuracy using GridSearch CV.

Leaderboard Score

Comparison of Models with their respective scores.

Models	Top features (Select K best method)	Parameters	F1_score	Public score	Private score
Bagging classifier (Kfold)	400 columns.	n_estimators = 200, Max_depth = 6	0.77050	0.76803	0.76293
Gradient Boosting (Kfold).	400 columns	n_estimators = 200, Max_depth = 6	0.75920	0.76558	0.76361
XBoosting (Skfold)	400 columns.	n_estimators = 200, Max_depth = 6	0.77319	0.78494	0.76634
XGBoosting (Skfold)	1000columns.	n_estimators = 250, Max_depth = 12	0.78112	0.79007	0.76523
XGBoosting (Skfold)	1000 columns.	n_estimators =200, Max_depth = 6	0.78424	0.78759	0.77063

Private Leaderboard Score

16	▲1	x2020dmu		0.77063	42	9d

Public Leaderboard Score

17		x2020dmu		0.79007	42	9d