**CSCI-555.20-Data Mining & Machine Learning**

**Final Report (The Toxicity Prediction Challenge)**

|  |  |
| --- | --- |
| **TEAM** | **NextGen** |
| Milan Mangukiya (Leader) | x2020flg@stfx.ca |
| Harsh Jasani | x2020fli@stfx.ca |
| Harshkumar Patel | x2020fle@stfx.ca |

1. **Data Preparation**
2. Handling data in feature matrix(feamat.csv)

**Infinite Data**: Specifically, the feature ‘V15’ (‘Kier1’) contained the infinite data.

Removed infinite data, and replaced with the mean value of the feature containing infinite data.

**Elimination of features**: Removed all feature having **Zero variance features.**

Reason**:** Feature havingconstant value will not improve the performance of the model.

Around 128 features got eliminated from the feature matrix(feamat.csv) having Zero variance features (all 0’s).

1. Mapping features from the feature matrix(feamat.csv) to Train and Test dataset.

Firstly, performed decomposition of ‘Id’ into Assay ID and Chemical ID from Train and Test dataset. (for e.g., ‘88-60-8;1682’ into Chemical Id: ‘88-60-8’, Assay ID: ‘1682’)

Mapped features from feature matrix to Train and Test dataset and written newly mapped data into csv’s.

1. Handling data in Train and Test dataset.

After obtaining data from the feature matrix into Train and Test dataset. We observed that there were around 68 Features in Train data set that had Zero variance features.

**Elimination of features**: Acquired features having Zero variance in the Train dataset and removed those features from both Train and Test dataset.

1. Data preparation for multiple models on the basis of “assay\_id”.

DATA-SET SPLIT: In order to generate individual classifier to each assay\_id, we first divided test and train data into 150 individual train and test set according to the assay\_id.

CLASSIFICATION: Classifier return and store prediction for each assay\_id.

MERGE OUTPUTS: Gather all the prediction result and merge into required single file for final submission.

1. **Classification model**
2. Feature selection

//explain all technique explored

//Which featured were used in top score model

//Why we have selected only top 50? Or any number (how to explain?)

Table 2.1: Classification model with parameters and internal evaluation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Model** | **Accuracy score** | **F1score**  **(Macro)** | **Leaderboard**  **Score** | **Features** | **Parameters** |
| 1 |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |

1. Training classification model

// Explain top 2 model, why this model was ensembled? How parameter was chosen (GridSearchCV)? internal score.

1. Internal evaluation

For internal validation of the model, we computed accuracy and f1 score. Particularly F1 macro score: mean of label-wise F1-scores.

For evaluating estimator performance, we use Cross-Validation with “number of folds” = 5.

Also, for controlling randomization in our predictions we are using ‘random\_state=1’ for Train dataset split and, in all algorithm used for building models.

1. **Leader-board Rank: 1st**

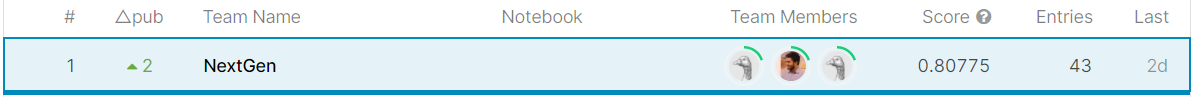


Fig. 3.1 Leader-board Rank for NextGen

References:

<https://scikit-learn.org/stable/index.html>