**The Toxicity Prediction Challenge II**

**CSCI 555 course – Prof. Othman Soufan**

**Report by Laxman Uppal (202003782)**

Introduction

Before starting the competition on Kaggle platform, there are few steps to go through to form a reproducible code as the steps below :

1. Understand the problem statement: Read and understood the problem statement of the competition to identify the key objectives and standards that need to be achieved.
2. Analyze the data: Analyzing the data provided to depict the features and target variables. Explored the data and the constraints related to it by going through the documentation.
3. Preprocess the data: Preprocess the data to make it suitable for training the machine learning models. This includes handling missing values to normalize the data.
4. Train the models: Train multiple machine learning models on the preprocessed data. Experimented with different algorithms like RandomForest classifier, Deep Neural Networks, GradientBoost classifier and Catboost classifier. Along with the algorithms, I have been studying about the hyperparameters to achieve the best performance through Google and Chat-GPT in understanding the best parameters to go with to achieve maximum toxicity prediction.
5. Evaluate the models: Evaluate the performance of the models using the different algorithms and measurers provided by the competition to identify the models that perform the best on the test set.
6. Submit the predictions: Generate predictions on the test set using the best-performing model. Submit the predictions to the competition and obtain the score.
7. Best choice of submissions: Talking about achieving maximum score, it was tricky with the private leaderboard hidden. So, I had to choose different submissions as my best choices among which one of them was my highest achievable score on public leaderboard.

Pre-processing and Feature extraction:

I have pre-processed the data by dividing the whole data into two columns assay\_id and chemical\_id. Then, removed the null values from the data and have tried using different combinations of features with the help of Chat-GPT which has been recommending me multiple descriptors among the 208 which are said to be effective in predicting the toxicity as per few research papers.

I have refined the best features by trying out different combinations and based on the research through Chat-GPT and other chemical prediction projects I was able to see Chi carrying the characteristics of molecular shape and topology of the chemical compound which proved to be effective in producing better toxicity prediction by testing out on Kaggle after trying out using the internal evaluation.

Based on the classifier, I have gone through the parameters by reading the significance of each parameter and got to know the key values that contribute different factors of prediction, balancing of the data, reduce overfitting and many more and then choose the best ones with the help of internal evaluation using K-fold cross validation.

Features:

I have considered assay\_id, MolLogP, TPSA, Chi2v, MolMR, Chi1v, Chi3v, fr\_benzene, Chi4v among which Chi topologies seem to be the major components to contribute my prediction. I have used very less number of features as I have included assay\_id under my belt as per the suggestions in the class. MolWt has been giving good results too but MolLogP, TPSA, MolMR and fr\_benzene seem to have much better co-relation as their combination seemed to produce better results in terms of internal evaluation.

Results:

My best submission results me to give a prediction of 82.302% in private leaderboard. But as per the competition my highest submission landing me in Rank 22 was 81.800%.

Graphical user interface, website

Description automatically generated

Conclusion :

I have attached the link to my GitHub profile where I have attached all my files related to the execution of the code to predict the best outcome over the competition. Based on my experience, after multiple trials I can conclude that assay\_id and Chi Topology features of every chemical compound seem to predict the toxicity much better than molecular weight and other key constraints.

Private Leaderboard screenshot :

Graphical user interface, application

Description automatically generated

Accessibility to my Project code for execution:

<https://github.com/Laxmanlee/Toxicity_pred_DM-ML>