# Final Report - DM & ML (CSCI - 555)

*St Francis Xavier University, Antigonish, NS*

*Department of Computer Sciences*

*Master's in Applied Computer Sciences*

*Team: X-Men (202006214, 202004299, 202006239)*

## Abstract

The toxicity prediction challenge (Kaggle) predicts the chemicals listed in the CSV which of them are toxic. The count of the chemicals produced overtime is increasing day by day and the testing of the chemicals over the bacteria, human cells and / or animal cells is becoming a tedious task. As a machine learning and data mining student, the primary aspect of the test is to predict the best data set from the given dataset of which all chemicals are toxic in nature and submitting the correct list of chemicals from over 9000 of those assay IDs is the challenge that the toxicity prediction challenge deals in.

## 1 Introduction

Machine learning and its concepts are used by the team X-men and its members to reproduce the best of results from the given sub-set of data. The applied concepts are derived from the class syllabus, internal team discussions, meetings with the professor and testing the methodologies over the time by making timely submission on Kaggle. The practice of machine learning concepts on the challenge has strengthened the base for performing machine learning related queries between the team members and has devised the practical side of the machine learning and its concepts being used in the real-life scenarios.

The methods disclosed in the forthcoming sections of the report hereby state the prior-act of the way discussions were carried out, plans were made, score was achieved and more improvements to the approaches were done. The best score achieved by the team is also mentioned in the document including the technique which led to the best prediction sets offered by X-men team members over the time and course duration. The novelty of the code and the methods which were used the latest in the code are disclosed in the document. The novelty here does not signify to the creation of a new method from scratch but for the concepts that the team has used and dug into overall during the project's research area.

## 2 Acknowledgement

The team is very thankful to Dr Soufan and the discussions carried out by the fellow classmates over the slack channel that helped build the program, reach the score that the team has acquired and keep faith in all the times of despair. Special thanks to Dr Soufan for guiding the team and suggesting some of the best possible solution to the problems faced by the team. His mentorship really assisted the team in reaching the goal and the score of Kaggle.

TEAM X-MEN

MINHAJ SHEIKH (SID: 202006214)

MUKUL NANDA (SID: 202004299)

NAVDEEP SINGH (SID: 202006239)

# 3 Methods used and novelty disclosed:

## 1. Data preparation technique

The data preparation is the primary step used in processing information from the raw data in order to generate meaningful predictions out of the data used. The primary step included in data selection / data preparation technique was to remove the columns which had redundant data and use the relevant features which remained after the feature selection process.

Initially we removed 1075 columns from the end since they seemed redundant and were increasing the overall execution time. We also used Interquartile range (IQR) to remove outliers from our dataset but it was a no go since it reduced our score by 10%.

We used left join on **chemical id** to combine our test and training datasets with feature matrix.

Later, we decided to change our approach and applied filters to our entire dataset. We made use of these filters:

- Removed constant features: Constant features are the one that contains single unique value. They provide no relevant information that can help in classification of the record at hand.
- Removed Quasi-constants: These features are almost same as constant ones. But in this, we provided a **VarianceThreshold** of 0.01 to remove the features with more than 99% similar values.
- Removed Duplicate Features: Duplicated features are the one that have similar values. Since they do not contribute to the problem in hand and add overhead and unnecessary delay to the training time. Therefore, we decided to remove them.
- Removed correlated features: If two of the features are highly correlated, they convey redundant information to the model, hence only one of them should be kept in dataset. We used correlation threshold of 90% between two columns for removal.

## 2. Model Training / Testing

Model training is referred to as the data feeding process which is done after the data preparation. In the data feeding process, the main targeted area is to provide the models with the right set of data so that accurate results are produced out of the combination of the data preparation and the training techniques. Generally, there are 2 types of trainings to look for – supervised and unsupervised. The model training performed in the toxicity prediction challenge consists of the following training and testing modules:

| Model | Changes | Accuracy | F1 Score | Leaderboard score |
|---|---|---|---|---|
| **Decision Trees** | Removed outliers using IQR | 0.86 | 0.76 | 0.72 |
| **XGBoost** | Tuned parameters | 0.90 | 0.78 | 0.77 |
| **Gradient Boosting** | SMOTE for Sampling | 0.85 | 0.73 | 0.74 |
| **Gradient Boosting** | Tuned parameters, SelectKBest for feature selection and Upsampling | 0.90 | 0.84 | 0.79 |

| | | | | |
|---|---|---|---|---|
| **Gradient boosting + XGB + LightGBM** | Parameter tuning with optuna, smote for sampling and SelectKBest for feature selection | 0.94 | 0.94 | 0.80 |

Primary reasons for using the methodologies listed in the table:

**1. Decision Trees:** Decision trees have been a part of the code and strategies used by x-men which were made during the earlier submissions too. This time while using decision trees, it was a contemplated decision to remove the outliers from the data created / fetched by the decision trees by using IQR (Interquartile Range). This technique helps divide the data into 4 quadrants and uses median of the data altogether to predict the best set for the toxicity problem of the chemicals in the given CSVs. The F1 score at this stage was 0.76 which obtained the accuracy of 0.86 altogether.

**2. XGBoost:** We decided to use XGBoost to train our model since it has been dominating Kaggle competitions. XGBoost is a decision tree-based algorithm that uses a gradient boosting framework. We tuned the parameters on XGBoost by using RandomsSearchCV, from this point onwards we decided to scale our data using StandardScaler. The score achieved by using XGBoost was 0.78 with the accuracy of 0.90.

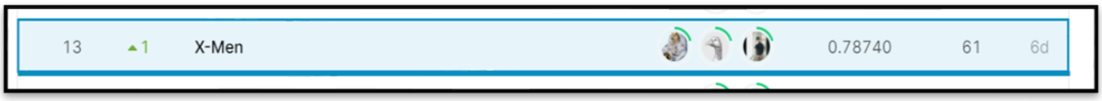
**3. Gradient Boosting with sampling:** The discussion of gradient boosting came up after studying the concept of boosting in the lecture of CSCI-555. The problem that later came up with the gradient boosting was to address the issue of oversampling and under-sampling of the training data set. SMOTE or Synthetic Minority Oversampling Technique was used to address the issue as using this sampling technique for imbalanced classification was helpful in combining together the Random oversampling (where machine tends to create more data for the minority of the sets in the training dataset) and the Under-sampling of the data (where the machine would delete, merge examples in the majority class set). **Novelty:** SMOTE came into the notice of the team X-men when the research for balancing the sampling techniques was suggested by one of the team members as there was a visible difference between the accuracy of the score achieved i.e., the internal validation score vs the Kaggle leaderboard score. The score was later taken care of by using the sampling technique with SMOTE and the score of 0.73 was achieved with the accuracy score of 0.85.

**4. Gradient Boosting with tuned parameters for sampling:** The parameter tuning was decided to be performed as it was done during the XGBoost which led to increasing the score of the team on Kaggle. Hence, during parameter tuning for gradient boosting, the team decided to use feature selection while tuning the parameters. The methods used for feature selection were SelectKBest (f_classif), RFE & Sequential wrapper methods. The score of the team improved to 0.84 with the accuracy of 0.90 where the idea of scaling was dropped as it was affecting the score with ± 10%.

**5. Ensembled Model:** We decided to ensemble three of our best models using a voting classifier. We combined three models: Gradient Boosting, XGBoost and LightGBM with tuned parameters using Optuna Framework. Smote for sampling the data and SelectKbest (f_classif) for feature selection. This gave us our best leaderboard score of 0.80.

## 3. Results / Leaderboard score

The Kaggle leaderboard has placed the team X-men at the 13th rank on the public leaderboard. The score of the team from the last submission stands at 0.78740. The team has utilised and exhausted all the above-mentioned techniques as part of the last / final submission.



| | | | | | |
|---|---|---|---|---|---|
| 13 | ▲1 | X-Men | 0.78740 | 61 | 6d |

# 4 Conclusion

The overall result of the project task was to get the team members familiarised with the concepts of machine learning and their practical use in the real-time world. The toxicity prediction challenge has met its requirements on teaching the learners about devising the best possible set of data for the prediction of the toxic chemicals and has not failed at being tough throughout the time.

With improvement in the score, the team got to learn about concepts which could only had been learnt the best using the practical knowledge of the concepts. The challenge demanded analytical thinking, a programming language’s knowledge, brain for mathematics and logic and tested the patience of the team members.

# 5 Appendix

1. Code Regeneration:

The other part of the documentation supports the code which can be easily reproduced by going through the instructions mentioned in the readme file of the zip-folder.