TIME-SENSITIVE CHEMICAL IDENTIFICATION TOOL

TA: Mohammed Abdul Aman; Member Names: Rishit Agrawal, Dinesh Burigala, Ethan Garcia, Tristan Li, Brian Morgan, Joey Mushrush, Nikita Rao, Rhushath Panyala, Ananya Uppal, Andrew Weiland, Christine Yuan

The Data Mine

Introduction

Time-sensitive/peroxide forming chemicals (PFCs) used in chemical laboratories can create major safety hazards, such as spontaneous explosions, if not stored and handled properly. However, the process of peroxide formation in some chemicals is not well-understood.

Our objective was to develop a user application for Dow Chemical Company that identifies chemicals as peroxide-forming or not and provides relevant safety information.

- Retrieve data for a chemical existing in the database
- Train a machine learning model to classify unfamiliar chemicals

PFC Classes

Class A: Forms explosive peroxides when exposed to air.
Class B: Forms explosive peroxides when concentrated.
Class C: Forms peroxides that react to heat/shocks.
Class D: May form peroxides; miscellaneous.

A B
C D

Data Preparation

Given: Data for 250 chemicals (both PFCs and non-PFCs) from a public spreadsheet; later, data for 272 more chemicals was obtained.

For all chemicals, extracted:
- Chemical names
- CAS numbers

For PFCs, extracted:
- PFC classes
- Incident histories

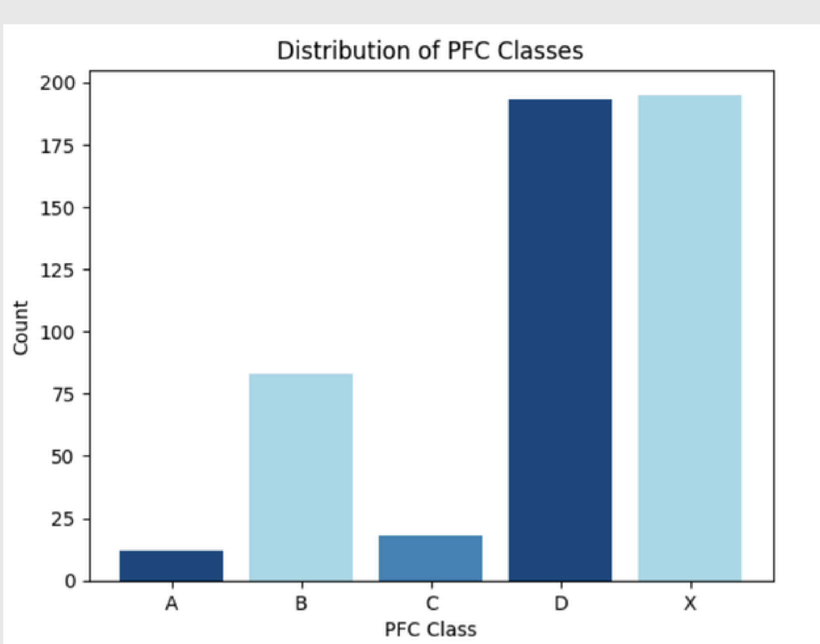

Figure 1: Bar diagram showing the distribution of chemical data by class (A, B, C, D, or non-PFC).

For the PFCs, binary data was generated by representing the presence/absence of certain functional groups with 1s and 0s.

All the chemical data was stored in an **SQL database** along with molecular weights and various other chemical properties. The distribution is shown in Figure 1.

Machine Learning Model

The team trained two models to answer two different questions:
1. **Does the chemical form peroxides?**
2. **If the chemical is a PFC, what class is it?**

Various models used for testing were:
- **Random forest:** Random decision-making trees; see Figure 2
- **XGBoost:** Similar to random forest, but model learns from each decision
- **Support vector machine:** Map data in space and determine physical boundaries for classification

Our team obtained the best results from **random forest models with hyperparameter tuning**, using functional groups and molecular weight as parameters.
Peroxide formation was predicted with **81% accuracy**, and PFC class with **69% accuracy**.

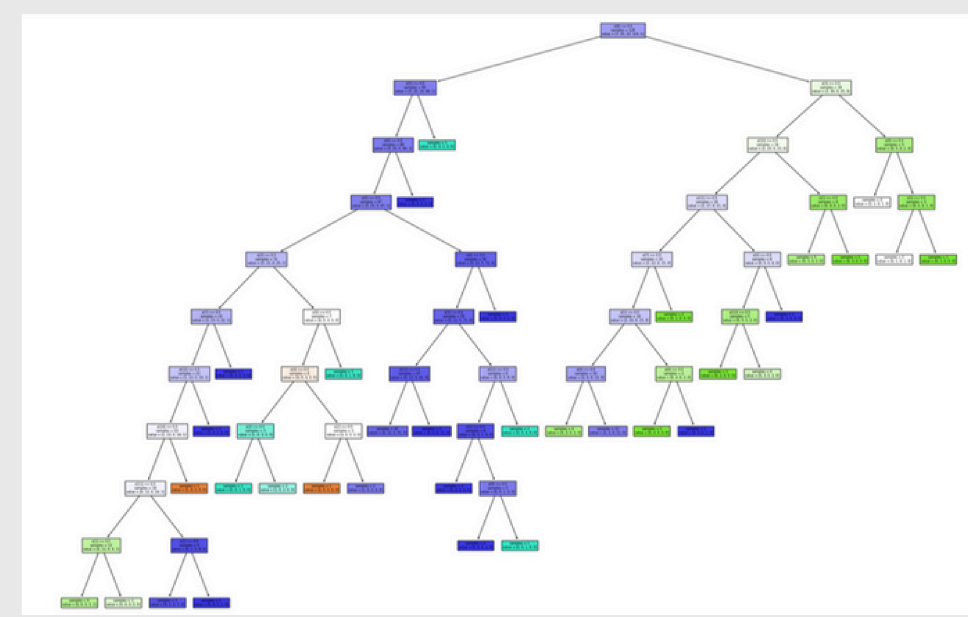

Figure 2: Diagram of decision trees for predicting PFC class.

Model Evaluation

Confusion matrices were used to compare performance between models. They represent the number of correct and incorrect predictions that the model makes during testing. Figures 3 and 4 show the confusion matrices for the final random forest models.

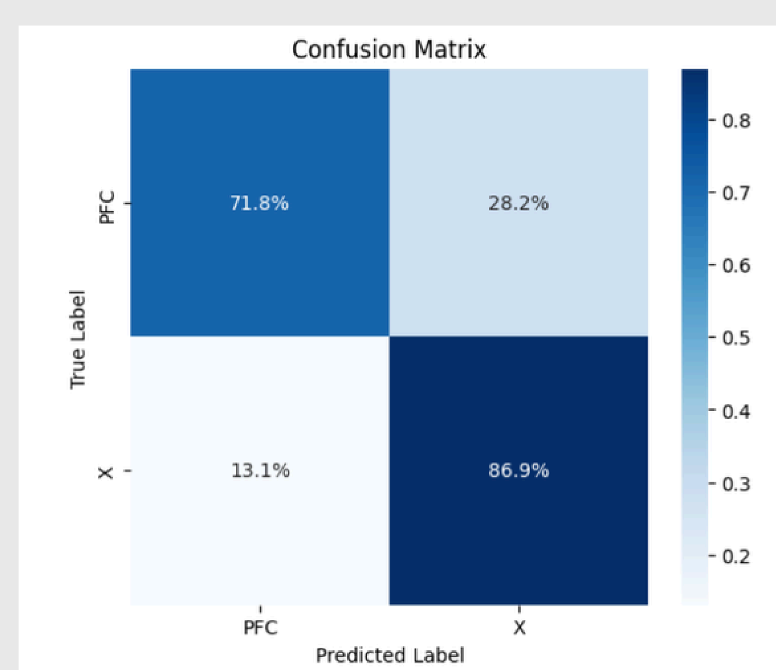

Figure 3: Confusion matrix for binary classification model (PFC or non-PFC).

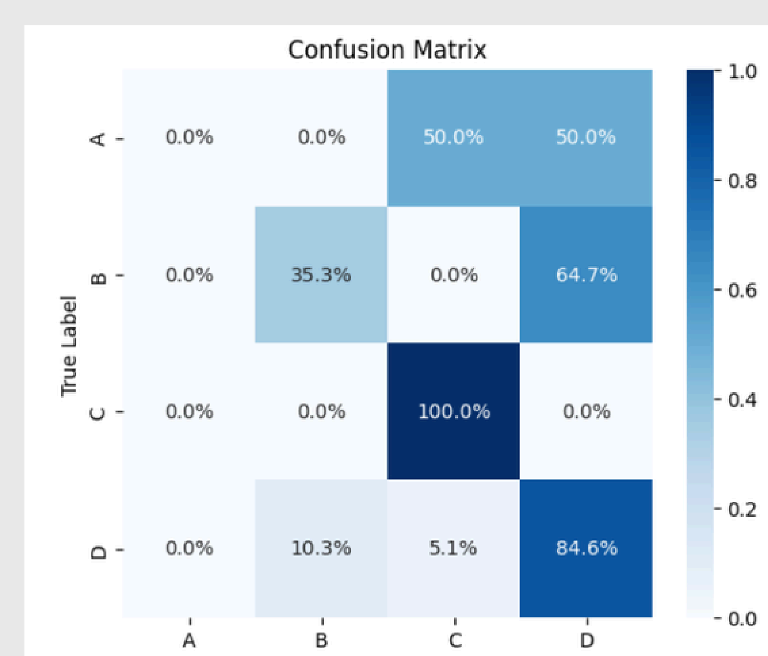

Figure 4: Confusion matrix for PFC class classification model (A, B, C, or D)

Dashboard/User Interface

The team used **Python Shiny**, which supports seamless integration of interactive components and machine learning models within a web-based application, to develop the final application.

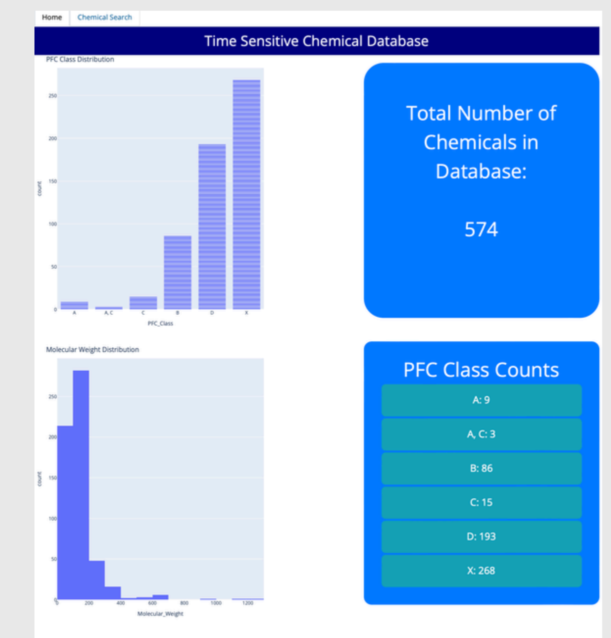

Figure 5: Application homepage made in Shiny

Homepage: Overview of database with distribution of PFC classes and warning levels; see Figure 5.

Interactive database: View PFC data; filter chemicals by class, molecular weight, functional groups, etc.

Predictive model: Input a chemical to predict its probability of peroxide formation.

Conclusion

Peroxide formation by chemicals can be predicted reasonably well (81% accuracy), but predicting the specific class of PFC likely requires knowledge of the peroxide formation reaction (69% accuracy).

The **random forest model** was found to identify PFCs with the highest accuracy. Confusion matrices show that the model could identify most non-PFCs, and that PFC classes A and B were most difficult to identify. Increasing the size of the dataset and including hyperparameters improved model accuracy.

Future Plans

Expand the application's functionality and provide more user-customization:
- More detailed chemical information (e.g. storage information, potential hazards)
- Chemical inventory for users to monitor PFCs used in laboratory and receive time-sensitivity warnings
- Export predictive model results for future reference

Acknowledgements

Kelly, Richard J. (1996) "Review of Safety Guidelines for Peroxidizable Organic Chemicals". *Chemical Health & Safety, 3*(5), 28-36.

University of Minnesota. (2019) *Peroxide Forming Chemical Data* [Spreadsheet].

We would like to thank our TA Mohammed Abdul Aman, our Dow mentors Luis Briceno-Mena and Bart Dervaux, and the Data Mine staff for all their support during this project