The Data Mine

Nicholas Lenfestey, Corporate Partner Advisor

Jessica Jud, Director of Partnerships

5/7/2025

PURDUE UNIVERSITY®

The Data Mine Overview

- Living, learning community (many students live in Hillenbrand, McCutcheon, or Harrison Hall)
- All backgrounds and majors welcome: **Data Science for All!** (student go from *no background* to *career readiness*)
- Supportive, active learning environment
- Learn data science skills in Python, R, SQL, HPC; no lectures; all project based
- Data sets at modern complexity and size, from all disciplines
- 160+ unique majors represented by students (not just CS, DS, or Engineering)

CONVERGENCE @ Discover Park District

McCutcheon Hall

Hillenbrand Hall

Harrison Hall

PURDUE UNIVERSITY

The Data Mine Growth

Beyond West Lafayette

West Lafayette

- 1576 students
- 521 in Corporate Partners
- 57 partners
- 69 projects
- 21 Indiana based companies

Indianapolis

- Started January 2024, with 17 students & 1 TA
- 135 students
- 8 Corporate Partners
- 12 projects
- In-person mentor collaboration

Rockies

- Started August 2024 through an OEDIT grant
- 48 students
- 6 Corporate Partners
- 8 projects
- Focused on needs for US DoD's Strategy for Operations

PURDUE UNIVERSITY

The Data Mine Staff

Data Science for All

- Open to students with no Data Science background
- Orthogonal to traditional university recruiting
- Interdisciplinary: everyone works together on solutions
- 9 months (2 sems.) experience: much longer/deeper than co-op
- Engagement across industry sectors, e.g.,
  - aerospace engineering,
  - agriculture,
  - life sciences,
  - manufacturing,
  - pharmaceutical science and computational drug development, etc.
- Students learn first-hand what is impactful at interview time:
  - soft skills
  - cross-disciplinary
  - tech skills and domain-specific skills

Students in seminar at Purdue University in West Lafayette

PURDUE UNIVERSITY

Impact & Innovation

Developing Tech Talent (Brain Gain)

- Geared toward all students; fast onramp and no background required
- Working with interdisciplinary students for 9 months (compared to 10 week summer internships)
- Companies provide innovative, domain specific challenges
- Technical communication to professional and nonprofessional experts

Pipeline/Pathways

- High-quality industry pathway that is welcoming
- Emphasis on regional economies
- Students become TAs then mentors, companies have access to Purdue students for hiring
- Augmenting the industry workforce
- Increasing Indiana talent

Experiential Learning at scale

- No blueprint, student-led data-driven innovative solutions to industry challenges
- Early-career training in modern data tools: Python, R, SQL, AWS, Azure, containers, predictive analytics, ML/AI, etc.
- Developing soft skills & professional development
- Access to High Performance Computing Cluster -ANVIL

PURDUE UNIVERSITY

Indiana Data Mine

- The Data Mine continues to leverage our presence at three Purdue campuses (Fort Wayne, Indianapolis, West Lafayette)
- Continued outreach to build partnerships with universities throughout the state and beyond
- Expand opportunities for DS training for first-generation and UGs from minority groups; also community colleges and liberal arts

2024-2025 Academic Year

- 96 total Indiana students (up from 9 students in 2022!)
- 15 Indiana institutions (9 new for AY 24-25)
- 7 additional Indiana institutions expressing interest for 2025
- The Data Mine team works with faculty from each participating institution to implement a program that meets their students' and region's needs

Ball State University
Indiana Wesleyan University*
Indiana University – Kokomo
Ivy Tech (5 campuses)*
Purdue Fort Wayne
Purdue Northwest*
Saint Mary's College*
Trine University*
University of Evansville*
University of Indianapolis*
University of Notre Dame
University of Southern Indiana*
Valparaiso University
Wabash College*

**New for AY 24-25*

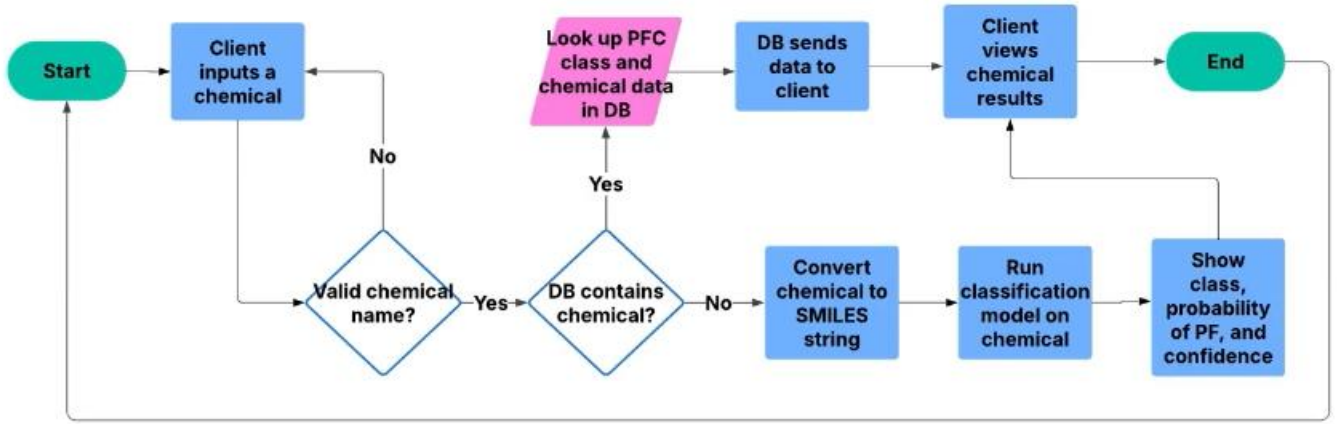
Figure 1: IDM Participating Colleges and Universities

PURDUE UNIVERSITY | The Data Mine

2024 – 2025 Corporate Partners

The Data Mine – Dow Project (Spring 2025)

Application Flowsheet



The Data Mine – Dow Project (Spring 2025)

PROJECT: TIME SENSITIVE CHEMICAL IDENTIFICATION TOOL

- **Description:** This project aims to develop a database and user-friendly application for identifying and classifying time-sensitive chemicals, particularly those that can form peroxides.
- **Keywords:** Peroxides, time sensitive chemicals, safety, data science, machine learning, representation learning, classification.
- **Tools/skills to be used:** Python, SQL Server, Azure SQL Server, Posit, Shiny, Data Science, Machine Learning, Chemistry, Cheminfomatics, Web App Development, Version Control

DOW General Business

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

DOW

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

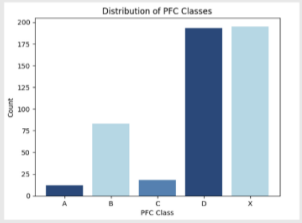


Figure 1: Bar diagram showing the distribution of chemical data by class (A, B, C, D, or non-PFC).

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**.

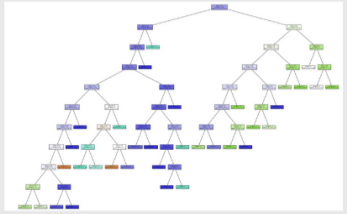


Figure 2: Diagram of decision trees for predicting PFC class.

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.

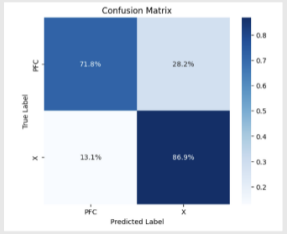


Figure 3: Confusion matrix for binary classification model (PFC or non-PFC).

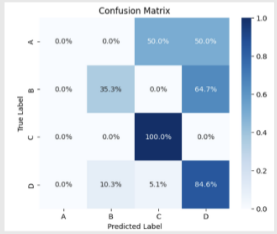


Figure 4: Confusion matrix for PFC class classification model (A, B, C, or D)

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.

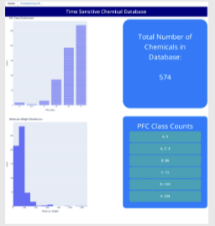


Figure 5: Application homepage made in Shiny

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

Thank You

Jessica Jud – Director of Partnerships – jljud@purdue.edu
Nicholas Lenfestey – Corporate Parner Advisor – nlenfest@purdue.edu

PURDUE UNIVERSITY®