Environment and
Climate Change Canada
Environnement et
Changement climatique Canada

Project: ML-Ready Dataset Preparation for Environmental Impacts Prediction

CMPT 2400 - Data Preparations and Analytics

Introduction

This assignment accounts for 35% of your final grade in the course. The first phase, Data Project 1, is worth 15%, and Data Project 2 is worth 20%.

Throughout these projects, you will practice exploring, cleaning and merging data.

Throughout these projects, you will work with a group of a maximum of 4 students. To understand the data, your group will employ various tools to detect any issues and patterns in the data, including bad housekeeping, outliers, and missing values, fix those issues, and, in the last project, align and merge multiple datasets. You will also encode features into appropriate representations, and in the case of the third project, do some feature selection and engineering. You will need to explain your process, justify the decisions you make, and deliver machine-learning-ready datasets.

For this project, we are working with Environment Climate Change Canada. The dataset your group is using is the National Pollutant Release Inventory (NPRI) Dataset. The datasets are included data from 2000-2022, separated into three tables: releases, disposals and transfers, and comments.

Please select a suitable machine learning problem which is viable to solve using this dataset. During the data preparation process, the first step is to finalize the specific ML problem you will be addressing. Each group should choose a unique problem from the proposed ML problems provided in this document. Problem assignment priority is determined on a first-come, first-served basis. When submitting your project plan, include your team members' names and the chosen problem. Before making your selection, ensure that no other group has already chosen the same problem.

These projects help you practice the concepts and skills you learned on real-world datasets. This will help you understand the kinds of issues that arise in datasets and how to handle them.

Learning Outcomes

- Practice exploring datasets and using tools that help with interpreting datasets, diagnosing issues and identifying patterns that exist in the data;
- Practice data cleaning techniques including handling outliers, missing values, and aligning and merging datasets (project 2);
- Practice encoding features and feature engineering (only project 2);
- Practice stakeholder engagement by listening to client needs, asking questions, prototyping solutions, and integrating feedback.
- **Develop skills in handling time-series data and addressing their unique challenges**, including trends, seasonality, and temporal dependencies.
- **Gain experience working with datasets that require specific domain knowledge for proper interpretation**, enhancing your ability to collaborate across disciplines and apply contextual understanding to data-driven solutions.

Assignment Instructions

The dataset your group is using is the National Pollutant Release Inventory (NPRI) Dataset, which you are using also as part of the Machine Learning 1 course.

- Explore your data, research the topic of the dataset and explain your findings in a short analytical report;
- Decide on what ML problems you are trying to solve with this dataset
- Detect patterns in the dataset, e.g., correlations between features, possible truncations, distributions of features, etc. and report them;
- Detect issues in the dataset, e.g., mistakes and bad housekeeping, missing values, outliers, scale mismatches, etc. and report them;
- Decide on how to handle the issues. Justify the decisions you took;
- Implement techniques to resolve the identified issues in Python;
- (Project 2 only) Align the datasets and merge them. Explain your method and decision-making process;
- (Project 2 only) Encode features into appropriate numeric representations, e.g., one-hot encoding, feature encoding, etc. Explain what you did;
- (Project 2 only) Normalize the features;

- (Project 2 only) Research feature engineering methods appropriate for the dataset. Then, select and engineer features that may be useful. Report the features you created, the approach that you took and why.
  Note: This is not supposed to be very time-consuming. Some easy and rudimentary feature engineering is what is requested here.
- For the second data project, you will need to locate an additional dataset to combine with the NPRI dataset. This will provide you with the opportunity to address a meaningful and valuable problem. To find a suitable dataset to complement your own, you can explore the following sources:

  1. StatsCan
  2. OpenData
  3. Crude Oil Pricing

Proposed Environmental Impact Prediction Problem Using NPRI Dataset

Predictions using external factors

1. What are predicted releases of nitrogen oxides and carbon monoxide from the oil and gas extraction sector (NAICS code 211110) in 2023 if the price of oil goes up or down?
   a. Crude Oil pricing NRCan
   b. Crude oil prices Stats Can
2. What are the predicted number of facilities reporting to the program in 2023 under different economic growth scenarios.
   a. GDP by industry
3. Has the federal carbon pricing system (started in 2019) decreased emissions of nitrogen oxides and carbon monoxide (substances released from burning fossil fuels)? And if so, what are the predicted decreases in the release of these substances as the carbon pricing system gradually increases in price?
   a. Federal carbon pricing went into effect in 2019 at $20 a Tonnes, increasing by $10 each year to $50 per Tonnes in 2022. From 2023 to 2030, the pricing will increase by $15 per year.
      i. It is important to keep in mind that some provinces already have carbon pricing systems in place, more info can be found here.
4. Depending on provincial population trends, predict the amount of ammonia, nitrate ions and phosphorous released to water from wastewater treatment plants in 2023 (data)

NPRI data predictions

1. Which industry reported the most spills and predict the number of spills that will occur in 2023 and the total amount spilled.
2. Based on NPRI data, which industry is predicted to have the highest growth of releases in 2023? Which will have the largest decline?
3. Based on NPRI data, what is the predicted proportion of releases to disposals in 2023.
4. What are the predicted trends for criteria air contaminants (sulphur dioxide, nitrogen oxides, volatile organic matter, particulate matter and carbon monoxide) for 2023.
5. Based on NPRI data, which province is predicted to have the largest decrease of substance releases (air, water and/or land) in the next five years.
6. What are the expected proportions of substance sent to landfills compared to substances sent for treatment or recycling in 2023
7. What are the predicted trends for methanol and ethanol releases across Canada and predict their releases in 2023?
8. What are the predicted proportions of nitrate ions to ammonia releases from wastewater treatment plants in 2023?

Related Resources:

1. Carbon pollution pricing systems across Canada
2. Factsheet of facility fuel type distinction
3. GDP by Industry

Project Deliverables

This is a multi-phase project that you will work on throughout the term. The following stages and submissions will be required.

Date	Deliverable
September 17	Project introduction by instructor & team formation
September 23 **12:30PM – 14:00PM in 4-013**	Meet, greet, and project introduction by Environment Climate Change Canada (ECCC)
September 26 **Project Plan DUE**	Project Plan Submission
October 30 **12:30PM – 14:00PM in 4-013** **Feedback Session 1**	Feedback Session - Well documented Jupyter notebook. Teams demonstrate the data project 1 delivery codes and documentation to instructor.
November 05 **12:30AM – 14:00PM in 4-013** **DEMO 1**	Teams demonstrate the output of data project 1, their classification model's result and their findings to ECCC team.
December 04 **12:30 – 14:00PM in 4-013** **Feedback Session 2**	Well documented Jupyter notebook. Teams demonstrate the data project 2 delivery codes and documentation to instructor.
December 11 **12:30PM – 14:00PM in 4-013** **FINAL DEMO**	Teams demonstrate the final project output (regression model) and their findings to ECCC team

Engaged Learning Assessments (1 Submissions)

You are required to prepare at least two questions to ask our clients, the ECCC, during the project introduction and meet and greet on September 23th. These questions need to be submitted on the Moodle Forum by midnight the night before the presentation. Your questions should not duplicate those of other students, so post early!

Project Plan (2%)– DUE Sep 26, 2024

After the project introduction, your team will meet to discuss how you want to approach the project. You should start by organizing yourselves for success- introduce yourselves to each other and discuss the best way to communicate as a group. You can use a group chat, email, and set regular in-person meeting times.

Your project plan should have the following components:

1. A team name.
2. All group members' names and contact information.
3. Your communication method and meeting times for the entire term.
4. The ML prediction problem you choose or propose and a brief description about what you understand from your ML problem and what you need to predict to achieve the goal of the project
5. An outline of the milestones for the project- both those set by your instructor, but also those you want to meet as a team (such as practicing together before a client demo).

Deliverables

- Write all the above information in a professionally formatted document and submit it to the Moodle drop-box before the deadline.

As this is a group submission, only one group member is required to submit. **All group members are responsible for the submission's content.**

Grading

Criteria	Exemplary (9-10 Points)	Proficient (7-8 Points)	Basic (5-6 Points)	Needs Improvement (1-4 Points)
Group Organization	A creative team name and all group members' details are included. Content demonstrates collaboration and thought, and planning has gone into team formation and on-going communication.	Group name lacks creativity and/or group member details are missing. Team formation is started, and some planning done, lacking details for communication planning.	Group lacks cohesive naming and details. Team formation is disjointed, and no solid communication plan is included.	Team details and plans are minimal to non-existent.
Timeline and Milestones	Provides a detailed and realistic timeline with clearly defined milestones that go above and beyond those provided by the instructor.	Provides a timeline with milestones, including some outside of those set by the instructor. Some details may be lacking or slightly unrealistic.	Provides a basic timeline with milestones as defined by the instructor, lacking details and may not be realistic.	Timeline and milestones are unclear, incomplete, or missing.
Clarity and Organization	The project plan is exceptionally clear, well-organized, and easy to follow. Sections are logically structured with a strong flow	The project plan is clear and well-organized with minor issues in flow or structure.	The project plan is somewhat organized but may lack clarity or have disjointed sections.	The project plan is unclear or poorly organized, making it difficult to follow.
Total Points				**/30 points**

Demo 1 (2%) and Demo 2(2%) – DUE nov 05, 2024 and Dec 11, 2024

ML development application requires an iterative, prototyping process to be most successful. The client communicates expectations and goals to your group, you interpret them and present an initial solution. Then the client provides feedback on what you have created, and you get to know if you are heading in the right direction. If there is a misalignment between your interpretation and their expectations, you adjust. This is an expected process. Often clients don't have a full picture of what they want, and their expectations and goals may change as they see it through your solutions. Your job is to continue to communicate back and forth and adjust until you reach a satisfactory solution for everyone. Misunderstandings and adjustments should not be viewed as failure. These are learnings and being able to receive feedback and adjust is an important skill in this field.

This demonstration is your group's first opportunity to show the client the direction you have started in and get to know if it aligns with their desired outcomes. It should be a two-way conversation between your group and the client, you should ask questions and clarify any areas of confusion. The goal is to leave the demo with a better understanding of what you will work on next. The goal of the demo is NOT to present your work and walk off the stage.

Anytime you do a presentation, you should keep your audience in mind. What do they already know? No need to repeat that. Do they know how to read code? Does the code matter to them? Let's leave out back-end details. What are they most interested in? Let's focus on that. What do we need to know from the client? Let's focus on that.

Each group will have 10 minutes with the client. Your group should use these 10 minutes wisely. Remember that you want to hear from the client as well. Show them what you have so far, and what you are thinking about doing next. Ask them what they think about it and if they have any concerns. Clarify details where your group had disagreements or got stuck. Ask them what they want to see next and what their priorities are.

You are required to observe all the other group's presentations.

Deliverables

For this submission, your group should submit the following files in Moodle:

- Presentation files from client demo.

Grading

Your demo for client will be graded on the following criteria:

Criteria	Exemplary (9-10 Points)	Proficient (7-8 Points)	Basic (5-6 Points)	Needs Improvement (1-4 Points)
Presentation Content	Content focuses on complete and operational features and functions. Areas of future improvement are thoughtful. Content demonstrates an understanding of the audience and was displayed in a professional, well-organized manner.	Content focuses on complete and operational features and functions. Areas of future improvement are thoughtful. Content demonstrated minor misunderstandings of the audience, but was displayed in a professional, organized manner.	Content focus was on incomplete or inoperable functions. Minimal or no future improvements mentioned. Content demonstrated major misunderstandings of the audience, and/or was disorganized, inconsistent, or hard to follow.	The content focus was on code details and not application use. No space was made for client feedback. Content was disorganized, inconsistent, or hard to follow.
Presentation Delivery	Group members were all well-prepared and spoke clearly and concisely. Members engaged the client. The group completed their presentation within the given time slot.	Most group members were well-prepared and spoke clearly and concisely. Members engaged the client. The group completed their presentation within the given time slot or within +/-2 mins.	One or two group members were well-prepared and/or some spoke quietly or caused confusion. Members engaged minimally with the client. The group had to be stopped on time or was less than 5 mins.	The group was unprepared and spoke quietly or caused confusion. The client was not included in the conversation.
Group Synergy	All group members attended and engaged. Members worked together to complete the presentation and ask and answer questions.	Most group members attended and engaged. Members worked together to complete the presentation and ask and answer questions.	One or two group members attended and engaged. Other members were absent or did not participate in asking and answering questions.	The group members demonstrated disorganization and tension. A lack of team communication and synergy was apparent.
Total Points				**/30 points**

FEEDBACK SESSIONS 1 & 2– DUE OCT 30 AND DEC 04:

The feedback sessions in this course are designed to simulate key industry practices, particularly the roles of both a **tech lead** and a **product owner** in machine learning (ML) and data science teams. In professional environments, tech leads, or product owners review each developer's code to ensure it meets the project's technical and business requirements. Similarly, in these feedback sessions, you submit well-documented Jupyter notebooks that include **justifications for their decisions and explanations of their methodologies**. The instructor then reviews the code, engages with students individually, asks relevant questions, and provides tailored feedback to help them refine their work.

Additionally, like a product owner, the instructor evaluates whether the students' approach aligns with the project's objectives and customer requirements.

This process reflects real-world ML workflows, where continuous testing and code reviews ensure code quality, alignment with project goals, and readiness for deployment. By engaging in this practice, students not only improve their current projects but also develop the critical skills needed for professional collaboration and feedback in industry settings. These sessions help students prepare for situations where tech leads and product owners provide feedback to optimize both the technical and business aspects of a solution.

Deliverables

For this submission, your group should submit the following files in Moodle:

- A well-documented Jupyter notebook detailing your understanding of the data accompanied by visualizations and comments, an explanation of the decisions made, and coding steps taken to create a machine-learning ready dataset.

Grading

Your Jupyter notebook in feedback session will be graded on the following criteria:

Component	Marks	Beginning (0–40%)	Novice (40–60%)	Proficient (60-80%)	Exemplary (80–100%)
Data Project 1: EDA & Data Cleaning					
Understanding	6	No understanding of the data is given.	A very basic understanding is provided. E.g., A description (not explanation) of features and the task	The data is explained reasonably well.	The explanations show that the topic of the data has been well researched and is connected to other ideas.
Issue Detection	2	No issue detection is done.	A list of problems is provided but is not complete.	All problems are pointed out and reported.	All problems are reported and discussed. The likely causes for problems and strategies on how to handle them are well thought out.
Pattern Discovery	2	No report of patterns observed in the data.	The patterns are listed, but they are incomplete.	Patterns for every feature and every pairwise combination of features are reported.	The possible causes for observed patterns are elaborated upon.
Visualizations and Statistics	2	No visualizations and statistics.	Incomplete visualizations and statistics.	Sufficient visualization and statistics to demonstrate the patterns and problems in the data.	Excellent visualizations and statistics and annotated with comments.
Fixing Bad Housekeeping	4	No fixing of bad housekeeping problems.	Some bad housekeeping problems are addressed. The decisions are not explained well.	Bad housekeeping is fixed, and the choices are reported.	All bad housekeeping issues are fixed. The choices are discussed, and the likely causes elaborated. .
Handling Outliers	5	Nothing done for outliers.	Outliers are removed with little explanation.	Outliers are pointed out, and the likely reason they happened is discussed.	The causes for outliers are explained exceptionally well.
Handling Missing Values	6	Nothing was done about missing values.	Missing values are removed or are imputed without	Missing values are handled with an explanation of choices made how to handle	There is a discussion on how to handle different missing values, and the strategies chosen are

			much explanation.	them.	justified.
Data Project 2: Consolidating Data & Feature Engineering					
Aligning Datasets	3	No alignment.	The datasets are rudimentarily aligned. The scales of features may be incorrect, or they may be duplicates.	The alignment is complete and sound.	The alignment is done based on research, and the choices are well explained.
Merging Datasets	2	No merging.	Basic merging is done, but some issues are present.	Problem-free merging of data.	The merging process and choices are well explained.
Feature Engineering					
Feature Encoding	4	No feature encoding.	Incomplete feature encoding.	Every feature has become a number in the correct way.	The decisions are explained.
Normalization	5	No normalization.	Some normalization is done.	The features are normalized.	There is discussion around the impact of features, and the features are weighted based on importance derived from research on the data.
Feature Engineering (Project 3 Only)	3	No feature engineering.	No feature engineering. Some ideas are given about the data and good features.	Some feature engineering.	Good feature engineering based on research. The engineered features are well explained and connected to data research.
Feature Selection (Project 3 Only)	6	No feature selection.	Feature selection is not correct.	Good feature selection.	The process and reasons for selection are explained.
Individual					
Individual participation	70%	The individual did not participate.	The individual minimally participated in the group work.	Active participation in the project.	Active participation in group work and bringing ideas.
Total Grade: _____ /50					

Optional Task: Interactive Dashboard with Plotly (5%)

As an additional, fun, and engaging activity, we propose that you create an interactive dashboard using Plotly to visualize the NPRI (National Pollutant Release Inventory) dataset. The NPRI dataset contains valuable information about pollutant releases across various regions of Canada. This optional task will help you better understand the data while practicing your visualization skills, creating a more intuitive and narrative-driven way to explore environmental impacts across different provinces.

Task Description:

- **Objective**: Build an interactive dashboard using Plotly Dash to visualize pollutant releases across Canada. Your dashboard should allow users to select different provinces, regions, or pollutant types and display how emissions vary geographically. Use storytelling techniques to guide users through the data, helping them uncover insights and understand the broader context of environmental impacts.

- **Requirements**:
  - Visualize data for different **pollutants** (e.g., Nitrogen Oxides, Carbon Monoxide) across Canadian regions.
  - Allow users to filter by **province/territory** and visualize trends over time, such as the change in pollutant releases from 2010 to present.
  - Provide interactive **maps** that highlight specific regions and display emission hotspots.
  - Include additional features like **bar charts** or **line plots** to compare pollutant levels between different industries (e.g., oil and gas, electricity generation).
  - Make the dashboard user-friendly with a clean and simple interface that displays key insights clearly.

This activity encourages you to explore the **geographical spread** and **trends in pollution** across Canada, helping to better understand the **environmental impact** of different industrial activities. It will also allow you to dive deeper into the **NPRI dataset** and spot patterns related to specific provinces or industries, making your analysis more meaningful and interactive. You'll also have the chance to practice key industry-relevant skills such as data visualization and dashboard creation, both valuable in ML and data science careers. Plus, your dashboard can serve as a tool to communicate findings in a visually compelling way to non-technical audiences, like how environmental data is shared with stakeholders in real-world projects.

Resources You May Need:

- *Plotly Documentation*
- *YouTube Tutorial on Plotly Dash*

Grading:

Criteria	**Exemplary (9-10 Points)**	**Proficient (7-8 Points)**	**Basic (5-6 Points)**	**Needs Improvement (1-4 Points)**
Dashboard Design & Usability	The dashboard has an intuitive, user-friendly design with a clean layout and clear navigation. All interactive elements (filters, buttons, etc.) are responsive and easy to use. The color scheme and layout enhance readability and accessibility.	The dashboard is generally user-friendly, with minor issues with layout or navigation. Most interactive elements work as intended, but some may need improvement.	The dashboard has a basic design with limited usability. Some interactive elements are not fully functional or are confusing to use.	The dashboard lacks a cohesive design and is difficult to navigate. Many interactive elements are broken or not user-friendly.
Data Visualization Quality	Visualizations are highly effective, accurately represent data, and include clear labels, legends, and titles. Interactive maps and plots provide deep insights and allow users to explore the data meaningfully.	Visualizations are effective but may lack some clarity in labeling or titles. Interactive elements provide some insights but could be more engaging.	Visualizations are simplistic and lack depth. Limited interactivity, and visual elements may be confusing or poorly labeled.	Visualizations are inaccurate, unclear, or misleading. Little to no interactivity is provided, and visual elements are poorly executed.
Functionality & Features	All required features (e.g., filtering by province, pollutant type, industry comparison) are fully functional. Additional advanced features (e.g., drill-down capabilities, animations)	Most required features are functional, with minor issues. Some advanced features may be implemented but not fully optimized.	Basic required features are functional, but with several limitations or bugs. Few to no advanced features are included.	Required features are missing or poorly implemented. The dashboard lacks key functionalities and advanced features.

	are implemented effectively.			
Creativity & Storytelling	Demonstrates exceptional creativity in visual design and storytelling. The dashboard uses narrative techniques (e.g., annotations, guided story panels, interactive timelines) to present findings in a compelling way that engages users and provides clear insights.	Shows good creativity in design and some use of storytelling elements. The narrative helps guide the user through the data, but it could be more refined or impactful.	Limited creativity in design and minimal use of storytelling. The dashboard is functional but lacks a coherent narrative to engage users.	Lacks creativity and storytelling elements. The dashboard is basic, with no effort to create a compelling narrative or engage users.
Technical Execution	The code is well-structured, clean, and well-documented. Efficient use of libraries (e.g., Plotly, Dash) and optimization techniques are applied to ensure smooth performance.	The code is mostly well-structured and documented, with minor issues. Some optimization is applied, but the performance could be improved.	The code is functional but poorly organized or documented. Performance issues may be present due to a lack of optimization.	The code is poorly written, disorganized, or lacks comments. The dashboard performs poorly due to technical flaws.
Explanation & Justification	The decisions behind visualizations, data choices, and features are clearly explained and justified in a separate document or within the dashboard.	Most decisions are explained and justified, showing a good understanding of data visualization principles.	Some decisions are explained, but justifications lack depth or clarity. Understanding data visualization principles is basic.	Little to no explanation or justification of decisions. The rationale for data choices and visualization techniques is unclear.
Total				/60 points