Project Ideas — Ranked by Acceptance Likelihood

Scoring: Each factor rated 1-5. Weighted composite determines rank.

Ranking Summary

Rank	Idea	Novelty	Data	Canadian Focus	ML Fit	Feasibility	Equity	**Score**
1	Indigenous Water Quality (DWA Prediction)	5	5	5	5	4	5	**4.8**
2	Food Desert Mapping	5	5	5	4	4	5	**4.7**
3	Opioid Overdose Risk Prediction	5	4	5	5	3	4	**4.3**
4	Wildfire Smoke Health Impact	4	5	5	5	4	3	**4.3**
5	Housing Affordability Risk	4	5	5	4	4	3	**4.0**
6	Transit Accessibility for Seniors	5	3	4	3	3	4	**3.7**

---

1. Predicting Drinking Water Advisories on First Nations Reserves

Problem: ~30 long-term drinking water advisories still active on First Nations reserves. Indigenous communities disproportionately affected by unsafe water — a uniquely Canadian crisis.

ML Approach: XGBoost/Random Forest classification to predict which water systems are at risk of issuing advisories based on system characteristics (age, source type, operator certification, province, population).

Key Data:

- ISC Long-term DWA dataset on Open Canada Portal
- StatCan environmental sustainability indicators
- XGBoost study achieved 86% accuracy on this exact problem

Why #1: Perfectly aligned with RBC Borealis equity mission. Uniquely Canadian. Government dataset readily available. Published ML research validates feasibility. Never done in past cohorts.

Risk: Sensitivity around Indigenous data sovereignty — proposal should acknowledge this respectfully.

Deep dive: `02-idea-indigenous-water-quality.md`

---

2. Mapping and Predicting Food Deserts in Canadian Cities

Problem: Low-income neighborhoods with poor access to grocery stores. Disproportionately affects seniors, immigrants, and disabled persons. 62% of families in subsidized housing are food insecure.

ML Approach: Clustering (k-medians) to identify food desert areas + classification (decision trees/CHAID) to predict emerging food deserts from socioeconomic features.

Key Data:

- StatCan "Indices of food desert and accessibility to food shops" on Open Canada
- Can-FED pan-Canadian dataset (56,589 dissemination areas)
- StatCan Business Register food environment measures

Why #2: Government dataset purpose-built for this problem. Novel for this program. Clear community impact — results could inform city planning. Proven ML methodology in literature.

Risk: Mostly analysis/mapping rather than prediction — need to frame a forward-looking ML task.

Deep dive: `03-idea-food-desert-accessibility.md`

---

3. Predicting Opioid Overdose Risk at Community Level

Problem: Canada's opioid crisis: 40,000+ deaths since 2016. BC hardest hit. Overdose rates surged during COVID and remain elevated.

ML Approach: Gradient boosting / logistic regression to predict community-level overdose risk from demographic, prescription, and social determinants. Literature shows 83.7% balanced accuracy.

Key Data:

- BC Opioid Overdose Analytical File (BCOOAF) — 13,318 records linked to StatCan
- StatCan health and social data
- Alberta administrative health data (428K+ patients)

Why #3: Extremely urgent Canadian issue. Strong published ML precedent. High impact — predictions could direct harm reduction resources.

Risk: Health data access may be restricted. May need to work with aggregated/public data rather than individual records. Sensitive topic requires careful framing.

Deep dive: `04-idea-opioid-overdose-prediction.md`

---

4. Predicting Wildfire Smoke Health Impact on Communities

Problem: 2023 Canadian wildfires caused 49-5,400 premature deaths from smoke exposure. PM2.5 spikes cause respiratory hospitalizations. Affects entire population during fire season.

ML Approach: Random Forest / LSTM to predict PM2.5 exposure from weather + fire data, then correlate with health outcomes (ER visits, respiratory admissions).

Key Data:

- NAPS (260 stations, data since 1969) on Open Canada
- AQHI data with JSON/CSV API
- CIHI hospitalization data

Why #4: Different from past wildfire projects (health impact, not fire detection). Extremely timely after 2023. Excellent government data with API access. Clear actionable outcome (public health warnings).

Risk: Similar enough to past wildfire projects that reviewers might conflate them. Need to strongly differentiate the health angle.

Deep dive: `05-idea-wildfire-smoke-health.md`

---

5. Predicting Housing Affordability Risk for Canadian Neighborhoods

Problem: 45% of Canadians very concerned about housing. Shelter costs up 25% since 2021. Rent up 24%. Nearly half of renters spend >30% of income on housing.

ML Approach: Regression to predict future affordability from economic indicators + classification to flag at-risk neighborhoods. Time-series forecasting of shelter-cost-to-income ratios.

Key Data:

- StatCan Canadian Housing Survey (annual)
- New Housing Price Index (monthly since 1981) on Open Canada
- Kaggle "Housing Affordability in Canada" competition dataset

Why #5: Universally relatable, strong data going back decades, clear ML task. Kaggle competition validates this as an ML-friendly problem.

Risk: More economics than social good. Less clear who the "community" beneficiary is. Weaker equity angle compared to top ideas.

Deep dive: `06-idea-housing-affordability.md`

---

6. Predicting Transit Accessibility Gaps for Seniors

Problem: Elderly increasingly dependent on public transit. Declining physical capabilities + retirement income = transport poverty. Canada's population aging rapidly.

ML Approach: Accessibility scoring model combining transit data + demographics + facility locations. Predict underserved areas.

Key Data:

- StatCan 28 spatial accessibility measures (7 destinations x 3 transport modes x peak/off-peak)
- Metrolinx assessment data (GTHA)
- Municipal GTFS transit feeds

Why #6: Novel angle never attempted. Aging population is growing concern. Government data exists.

Risk: Hardest to scope tightly. Accessibility is multidimensional. May struggle to define a clean ML prediction task. Data integration across sources is complex.

Deep dive: `07-idea-transit-accessibility-seniors.md`

Recommendation

Primary proposal: #1 (Indigenous Water Quality) or #2 (Food Deserts) — both have the highest combination of novelty, data quality, equity alignment, and ML feasibility.

Secondary proposal: #4 (Wildfire Smoke Health) — strong data, timely, different enough from primary to offer the mentor a real choice.