ACENET
Microcredential in Advanced Computing
ISP Report

Project title: Exploring and predicting Post-Partum Depression (PPD) rates across the provinces in Canada

Participant name: Riziel M. Cariño

Date: July 30, 2024

Abstract:

The project aims to explores and predicts Post-Partum Depression (PPD) rates across Canadian provinces using a RandomForestRegressor model, leveraging demographic and medical data from the historical datasets. The analysis includes model optimization through Grid Search and cross-validation, with results visualized to compare actual and predicted PPD rates.

1. Introduction

The project focuses on understanding and predicting Post-Partum Depression (PPD) rates across different provinces in Canada. PPD is a significant maternal mental health issue that can have profound effects on both mothers and their infants.

The motivation behind this project stems from the critical need to identify and support mothers at risk of PPD, thereby improving maternal and infant health outcomes. By accurately predicting PPD prevalence, healthcare providers can better allocate resources and tailor interventions to those most in need.

The primary research question guiding this project is: *Can we accurately predict the prevalence of Post-Partum Depression across Canadian provinces using demographic, medical, and psychological data?* This hypothesis aims to leverage machine learning techniques to uncover patterns and predictors of PPD, ultimately aiding in early detection and prevention efforts.

2. Background

Post-Partum Depression (PPD) is a significant mental health issue affecting new mothers, characterized by prolonged feelings of sadness, anxiety, and fatigue. Unlike the temporary "baby blues," PPD can last for several months and requires professional treatment. This condition impacts not only the well-being of mothers but also child development and family dynamics, making early identification and treatment crucial. Maternal mental health issues, including PPD, are a major public health concern worldwide, affecting societal health and family dynamics.

PPD prevalence varies by region and population, influenced by demographic factors (age, marital status, education), medical history (previous depression or mood disorders), psychological assessments (mental and physical health, life satisfaction), and support systems (social support, parenting programs).

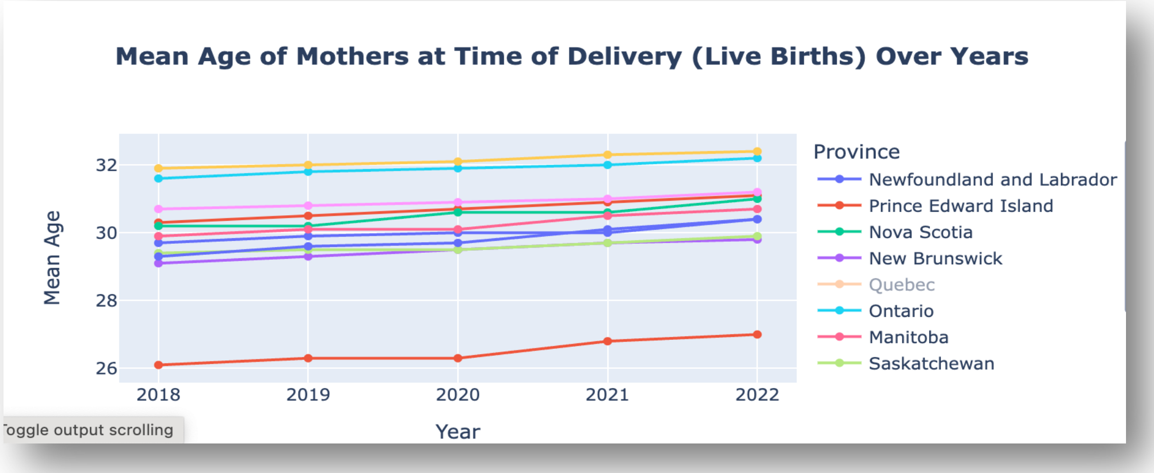
This project uses data from the Survey on Maternal Health (SMH) and the Public Health Agency of Canada, which provide comprehensive demographic, medical, and psychological information on new mothers.

3. Analysis

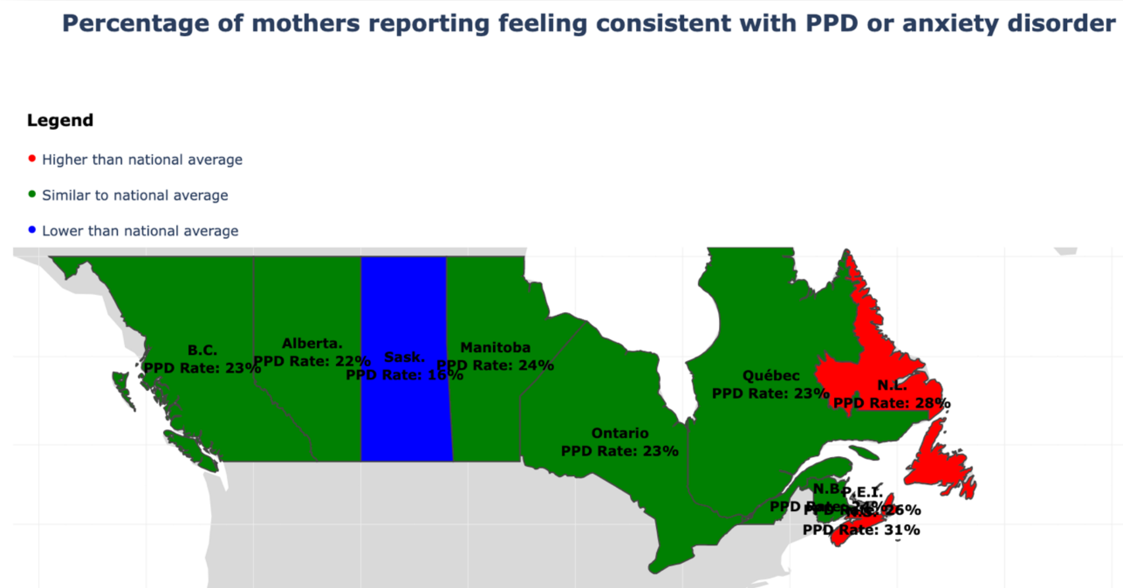
a. Dataset

The dataset was sourced from the Survey on Maternal Health (SMH) and the Public Health Agency of Canada. This dataset includes detailed demographic, medical, and psychological data on new mothers across Canadian provinces.

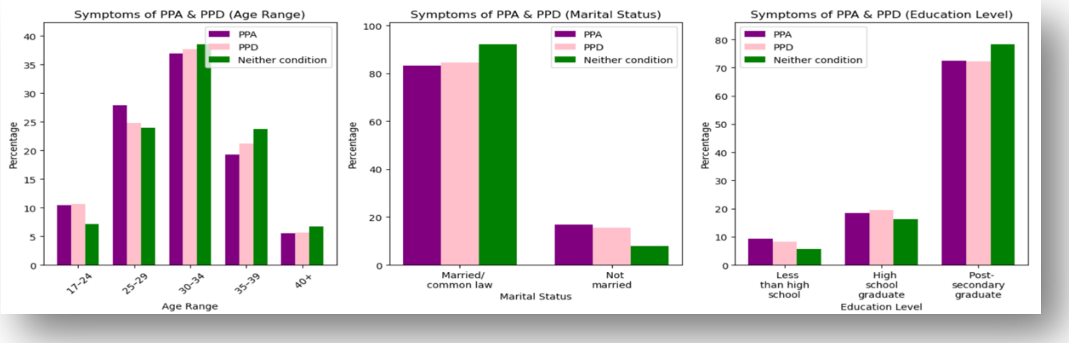
- https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310041701 - Mean age of mothers at Time of Delivery (Live Births)

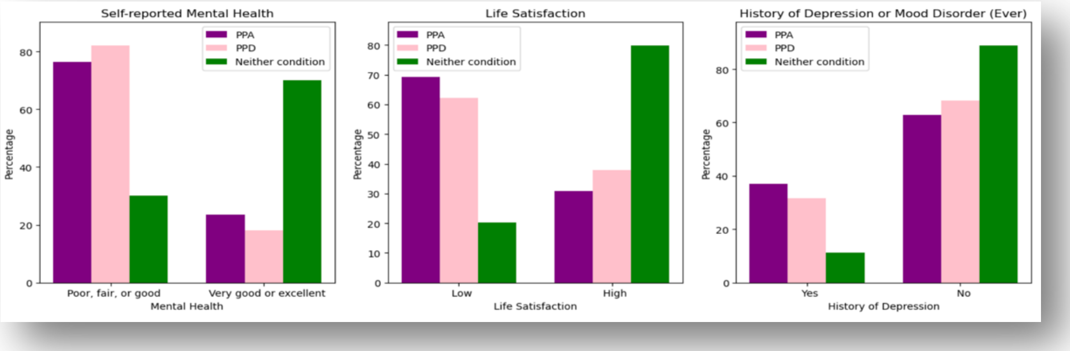


- https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2019041-eng.pdf - Maternal Mental Health (Outcome Variable: Percentage of mothers reporting PPD symptoms in each province)



- https://www150.statcan.gc.ca/n1/pub/82m0021x/82m0021x2024001-eng.htm - Mental Health and Access to Care Survey
  Survey data variables:
  - Demographics: Age groups (17-24, 25-29, 30-34, 35-39, 40+), marital status (Married/Common-Law, Not Married), education levels (Less than High School, High School Graduate, Post-Secondary Graduate).
  - Medical History: History of depression or mood disorders (Yes, No).
  - Psychological Assessments: Self-reported mental and physical health status.
  - Support Systems: Availability of social support, participation in parenting support programs.





b. Data Preparation:

- **Data Cleaning:** Ensured consistency and completeness of the dataset by handling missing values and correcting any discrepancies.
- **Feature Encoding:** Converted categorical variables to numerical values to be used in the machine learning model.
- **Normalization:** Standardized numerical features to ensure uniformity and improve model performance.

c. Analysis Method:

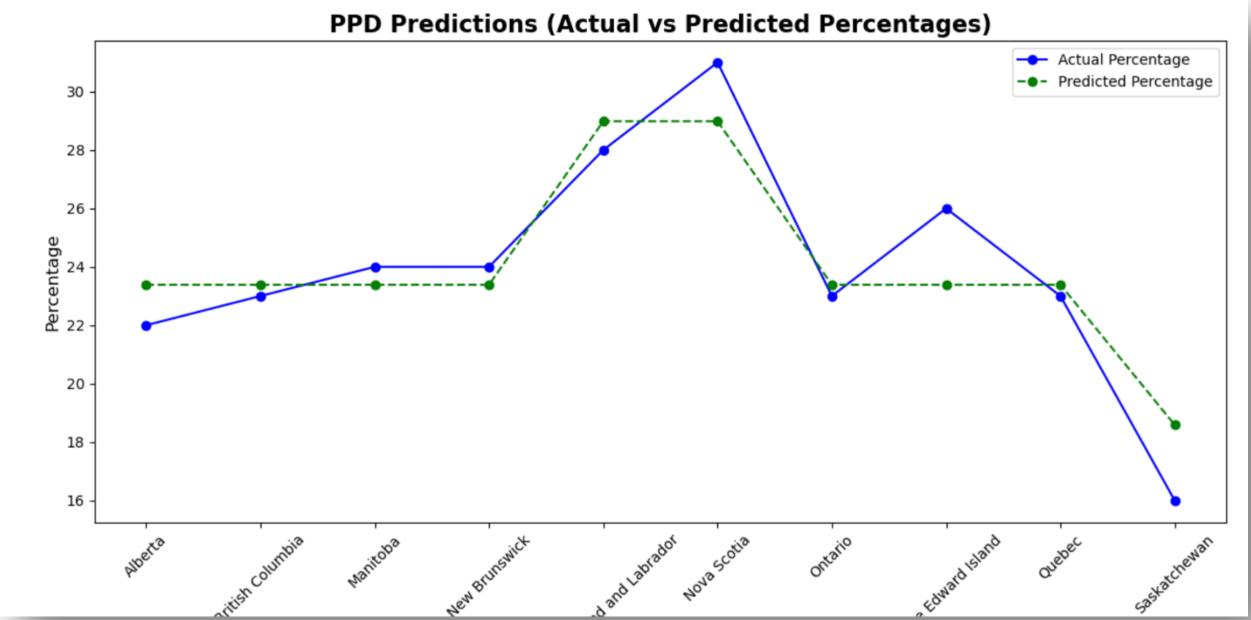
- **Model Selection:** RandomForestRegressor model due to its ability to handle complex, non-linear relationships and its robustness in dealing with various feature types and missing values.
- **Hyperparameter Tuning:** Utilized Grid Search with cross-validation to optimize the model's hyperparameters (n_estimators, max_depth, min_samples_split).

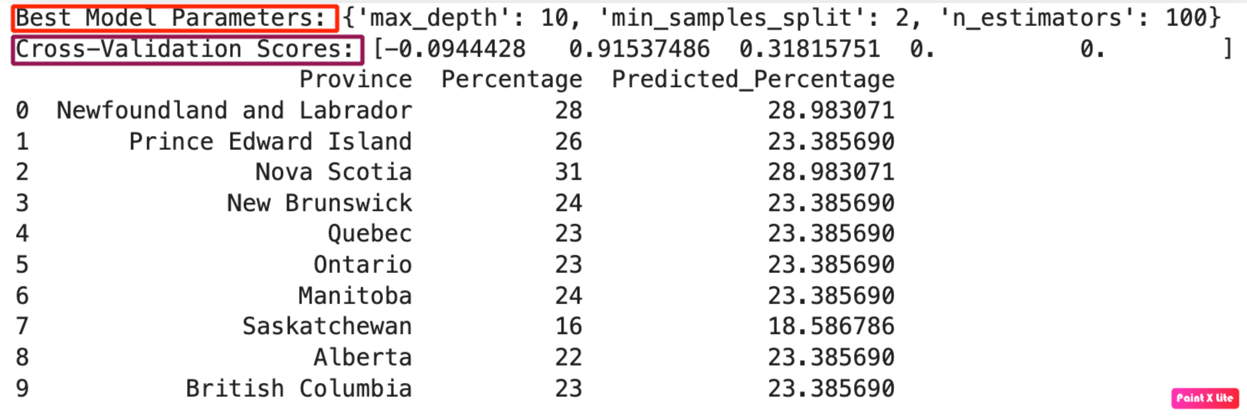
d. High-Performance Computing (HPC):

- **Job Script Execution:** Created a bash script to automate the environment setup, package installation, and script execution on an HPC cluster.
- **Parallel Processing:** Leveraged HPC resources to parallelize the Grid Search and cross-validation processes, significantly reducing computation time.
- **Output Management:** Directed output files (e.g., predictions, plots, model parameters) to a specific directory for organized storage and easy access.

4. Results

The RandomForestRegressor model’s predictions are reasonably accurate for some provinces but less accurate for others. This is reflected in both the line plot and the pie charts, where certain provinces show significant discrepancies. Provinces where the model performs well (e.g., Newfoundland and Labrador and Nova Scotia) show close alignment between actual and predicted values. In contrast, provinces with larger deviations (e.g., Prince Edward Island and Saskatchewan) suggest that the model needs some improvement.





```
Best Model Parameters: {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 100}
Cross-Validation Scores: [-0.0944428   0.91537486  0.31815751  0.          0.        ]
                    Province  Percentage  Predicted_Percentage
0  Newfoundland and Labrador          28             28.983071
1       Prince Edward Island          26             23.385690
2                Nova Scotia          31             28.983071
3              New Brunswick          24             23.385690
4                     Quebec          23             23.385690
5                    Ontario          23             23.385690
6                   Manitoba          24             23.385690
7               Saskatchewan          16             18.586786
8                    Alberta          22             23.385690
9           British Columbia          23             23.385690
```

Model Evaluation: Mean Squared Error (MSE) result: 4.377376191043089

Overall, the results suggest that demographic and medical history variables are significant predictors of PPD, and the model can be useful tool for early detection and resource allocation.

5. Discussion

The project aimed is to predict the prevalence of Post Partum Depression / Anxiety across the provinces in Canada, using RandomForestRegressor model. While the model's prediction was successful, there was a few challenges encountered during the process.

- Limited data specifically for maternal only. The datasets available are for all genders and will spend a lot of time sorting/cleaning the data. The dataset may not include all relevant factors influencing PPD, such as socioeconomic status, cultural background, or access to healthcare services.
- Significant time / time constraint in choosing the right model for the project.

Despite the challenges encountered, the project demonstrated the feasibility of using machine learning to predict PPD prevalence, with significant implications for public health and maternal mental health interventions. Addressing the limitations and challenges identified can further enhance the accuracy and applicability of the model, contributing to better support and outcomes for new mothers across Canada.

Conclusion

The RandomForestRegressor model effectively predicted the prevalence of postpartum depression (PPD) across Canadian provinces, with an average mean squared error and consistent cross-validation scores, though some discrepancies suggest the influence of unmeasured factors. The project demonstrates the potential of machine learning to predict PPD prevalence based on a range of demographic, medical, and psychological factors. The findings may aid in early identification and targeted intervention, improving maternal mental health outcomes.

In future research directions, it will be ideally if we include the following to refine the model performance and ultimately contribute to better maternal mental health care and outcomes.

- **Expand dataset**: Incorporate additional variables such as socioeconomic status, cultural background, and access to healthcare services to capture the full complexity of PPD.
- **Increase Sample Size:** Collaborate with more regions and collect longitudinal data to enhance the model's robustness.
- **Explore Other Models:** Experiment with different machine learning algorithms or ensemble methods to improve predictive accuracy and interpretability.

References

[1] “Exploring predictors and prevalence of postpartum depression among mothers” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11092128/#notes-a.l.etitle

[2] “Machine Learning Models for the Prediction of Postpartum Depression: Application and Comparison Based on a Cohort Study”
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7226048/

[3] Familiarizing the model I used: RandomForestRegressor

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor

[4] “Symptoms of postpartum anxiety and depression among women in Canada:”

https://link.springer.com/article/10.17269/s41997-020-00420-4#Sec8

Supplementary Materials

Datasets:
[1] “Mean age of mother at time of delivery (live births)”
https://www150.statcan.gc.ca/t1/tbl1/en/tv.action%3Fpid=1310041701

[2] “Maternal Mental Health Care in Canada”,
https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2019041-eng.htm

[3] “Mental Health and Access to Care Survey”
https://www150.statcan.gc.ca/n1/pub/82m0021x/82m0021x2024001-eng.htm

Github:

https://github.com/RizielC/ISP_PPD-rates-across-Canada