## **Introduction**

The goal of this project was to create a machine learning model to predict period pain levels using survey responses. The project involved cleaning the data, selecting important features, and testing different models to find the best predictor and align with the Aura Flow project proposal key objectives.

This report explains how the model was built, how well it performed, and how its predictions compared to the actual pain levels in the test data.

## **Data Processing & Feature Selection**

### 🔹 Handling Missing Data

* Some survey responses were incomplete.
* For missing numbers, I filled them in using the median value.
* For missing words (categories), I used the most common answer in each column.
* After this step, all missing data was filled.

### 🔹 Choosing the Most Important Features

Some information in the dataset wasn’t very helpful for predicting pain levels.  
 To keep the model simple and accurate, I picked the top 4 most important features:

1. Work\_School\_Impact\_Days – How many days school or work was affected by period pain.
2. Missed\_Work\_School\_Days – Number of days missed due to period pain.
3. Concern\_Level\_About\_Bleeding – How worried someone was about their bleeding.
4. Personal\_Income – The income level of the person.

Removing extra details helped the model focus on what matters most.

### 🔹 Fixing Class Imbalance

* Some pain levels (like "No Pain") were rare in the dataset, which made it harder for the model to predict them.
* I tried adding more "No Pain" cases using SMOTE (Synthetic Minority Oversampling Technique).
* I also tested removing some of the more common pain levels.
* Neither of these approaches improved accuracy beyond 50%.
* Final choice: Adjust model settings to give more importance to underrepresented groups.

## **Model Training & Performance**

I tested three different machine learning models:

| Model | Accuracy |
| --- | --- |
| Random Forest | 35% |
| Logistic Regression | 25% |
| XGBoost (Optimized) | 50% (Best Model) |

The XGBoost model performed the best, with an accuracy of 50% on test data.

### XGBoost Results

| Pain Level | Precision | Recall | F1-Score | Support |
| --- | --- | --- | --- | --- |
| Moderate Pain | 0.50 | 0.72 | 0.58 | 46 |
| No Pain | 0.00 | 0.00 | 0.00 | 3 |
| Severe Pain | 0.57 | 0.33 | 0.42 | 24 |
| Slight Pain | 0.50 | 0.39 | 0.44 | 36 |
| Overall Accuracy | 50% |  |  |  |

## **Comparing Predictions to Actual Data**

After testing the model, I compared its predictions to the actual pain levels people reported.

### What Worked Well:

* The model was good at predicting "Moderate Pain" (72% recall).
* It correctly identified many cases of Slight and Severe pain.

### What Needs Improvement:

* The model struggled with "No Pain" cases (0% recall).
  + This is likely because I had only 3 examples in the dataset.
* Severe pain had low recall (33%), meaning the model missed some cases.

### Why Accuracy is Limited

* Not enough data (especially for "No Pain").
* Survey responses were self-reported, which can introduce bias.
* Not enough variety in features—lifestyle, medical history, or stress levels could help improve predictions.

## **Next Steps & How to Improve**

The model was able to predict pain levels with 50% accuracy, but it can still be improved.

1 Gather more data – Especially for groups like "No Pain", so the model can learn better.  
2️ Use more detailed features – Adding information on diet, hydration, exercise, or medical history might help.  
3 ️ Try deep learning models – Neural networks could find patterns that other models miss.  
4 ️ Combine multiple models – Mixing XGBoost with other models might improve accuracy.  
5️ Adjust prediction rules – Fine-tuning decision thresholds could help balance precision and recall.

## **Final Summary**

* Best Model: XGBoost (Optimized)
* Best Accuracy: 50%
* Key Features: Work\_School\_Impact\_Days, Missed\_Work\_School\_Days, Concern\_Level\_About\_Bleeding, Personal\_Income
* Main Limitation: The dataset needs more variety and balance.

This project successfully developed a predictive model for period pain levels, and future improvements could make it even better!