**DS596 Spring ’25: Final Project**

Requirements The final project will require you to develop a written proposal for a data-driven research project in biology. The final text should be around 1,500 words (in the 1,250-to-1,750-word range) and include any figures you think would be helpful. In terms of polish and style, think of this as something you might submit as a fellowship application (so include citations – not part of the word count – and all that). Your proposal should begin with a background section that introduces the biological and computational context of the problem, reviews current literature and existing approaches, and motivates the objectives and approach of the proposal.

Proposal topics could include the following:

• Development of a new computational approach

• Application of existing methods to new or under-explored data

• Analysis of multiple methods applied to common benchmark datasets

**Topic idea….**

focus on predicting depression risk using behavioral and clinical data from publicly available sources like NHANES or PHQ-9-based surveys. I plan to apply supervised machine learning models such as logistic regression and XGBoost to identify key predictive factors and evaluate their potential to support early mental health screening.

**Relevant NHANES Variables (**[link to data](https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023)**) for Depression Risk Prediction:**

Used only to define binary outcome variables for classification

* **DPQ020\_binary** – Felt down, depressed, or hopeless (binary)
  + 0 – Not at all
  + 1 – Any depressive response (several days, more than half the days, or nearly every day)
* **DPQ060\_binary** – Felt bad about yourself (binary)
  + 0 – Not at all
  + 1 – Any depressive response (several days, more than half the days, or nearly every day)

### • demographics\_nhanes

* **RIAGENDR** – Gender
* **RIDAGEYR** – Age
* **RIDRETH3** – Race/Hispanic origin (includes Asian)
* **DMDEDUC2** – Education level (age 20+)
* **DMDMARTZ** – Marital status (age 20+)
* **INDFMPIR** – Family income-to-poverty ratio

### • income\_nhanes

* **INDFMMPI** – Family monthly poverty level index
* **INDFMMPC** – Categorized monthly poverty index
* **INQ300** – Family has more than $5,000 in savings

### • sleepDisorders\_nhanes

* **SLD012** – Sleep hours on weekdays
* **SLD013** – Sleep hours on weekends

### • smokingBehavior\_nhanes

* **SMQ020** – Smoked at least 100 cigarettes in life

### • alcoholUse\_nhanes

* **ALQ111** – Ever had a drink
* **ALQ121** – Past 12-month alcohol frequency

### • physicalActivity\_nhanes

* **PAD800** – Minutes of moderate activity (per week)
* **PAD680** – Minutes of sedentary activity (per day)

### • Clinical Lab Measures

* **LBXVIDMS** – Vitamin D (ng/mL)
* **LBXGH** – Glycohemoglobin (HbA1c, %)
* **LBXHSCRP** – C-reactive protein (mg/L)