***Data Science Lab  
Final Report***

*Lour Hakim – 323038091  
Shahar Rashdi – 314948415*

**Mental Health in U.S. High School Students:   
A Data-Driven Analysis Based on YRBS 2023 and Socioeconomic Indicators**

**Introduction**

**Background and goals**

Adolescent mental health in the United States has become a major public health concern, with rising rates of sadness, hopelessness, and suicidal thoughts reported by high school students. These issues are influenced by a range of personal, social, and environmental factors. Yet, the complexity of how these factors interact—especially in combination with broader economic conditions—is often overlooked. Our project aims to model these complex relationships by combining student-level behavioral and demographic data with state-level economic indicators. The ultimate goal is to identify high-risk student profiles and generate interpretable insights that may support early intervention and informed policy-making.

**Data and variables**

Our project integrates three distinct datasets to investigate mental health risk factors among U.S. high school students. The primary dataset is the 2023 Youth Risk Behavior Survey (YRBS), an official CDC survey containing over 100 student-level variables related to emotional well-being, behavior, demographics, and school life. To incorporate economic context, we merged two additional sources: (1) GDP by state from the U.S. Bureau of Economic Analysis, and (2) unemployment rate and mean household income from the U.S. Census Bureau’s American Community Survey. The latter was compiled from eight state-level CSV files. Merging was performed using the sitename-to-state mapping.

Our analysis targets three mental health outcomes:

* **Q26**: “Felt sad or hopeless for two or more weeks” – binary
* **Q27**: “Seriously considered suicide” – binary
* **Q84**: “Days of poor mental health in past 30 days” – ordinal with five levels: Never, Rarely, Sometimes, Most of the time, Always

**Example Variables by Group:**

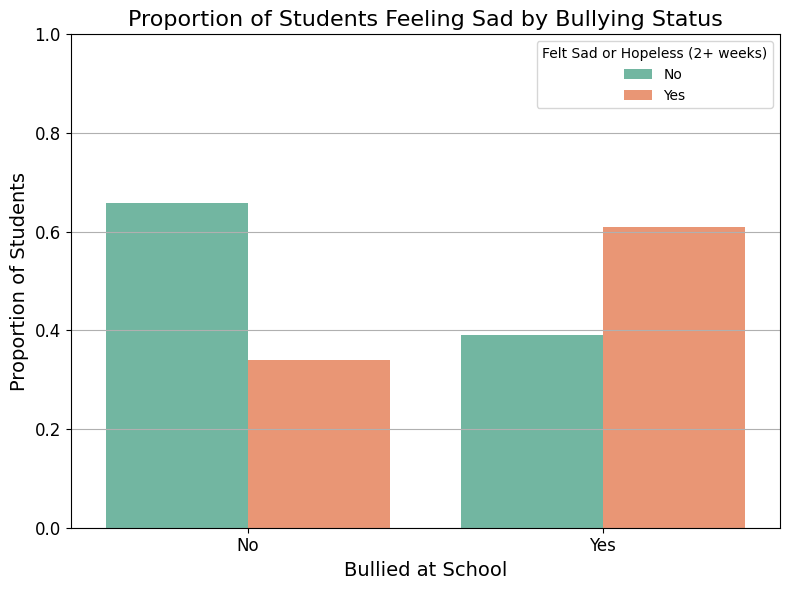
|  |  |
| --- | --- |
| **Group** | **Examples** |
| Demographic | Age, Sex, Race, Sexual & Gender Identity |
| Behavioral | Bullying, Sleep, Breakfast frequency, Screen Time, Substance use |
| Health | BMI, Physical activity |
| Academic | Self-reported grades |
| Economic | GDP, Unemployment rate, Household income (by state) |

**Data Sources:**

1. Youth Risk Behavior Survey (YRBS) – 2023  
   <https://www.cdc.gov/yrbs/data/index.html>
2. GDP by state – 2023   
   <https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas>
3. Mean income and employment rate by state – 2023  
   <https://data.census.gov/advanced>

**Important EDA results**

Exploratory analysis revealed several predictive patterns. Students who experienced bullying (Q24) had significantly higher sadness rates (Q26), with a univariate logistic regression achieving 65.9% test accuracy—well above the 57.3% constant baseline. We also identified BMI outliers using IQR thresholds and created a bmi\_outlier indicator. While this variable was not selected by the final models, the continuous BMI percentile (bmipct) consistently appeared among the most important predictors. Additionally, students from states with lower mean household income higher rates of sadness (Q26), highlighting economic disparities that informed our multi-domain modeling approach.



**Modelling**

**Model Design**

We applied three complementary modeling approaches to analyze mental health risks among U.S. high school students. First, we used **Sparse Logistic Regression (SLR)** with L1 regularization, enabling automatic feature selection by shrinking irrelevant coefficients to zero. This makes the model interpretable, highlighting a small, focused set of risk factors. Second, we applied a **Random Forest (RF)** model to capture non-linear effects and variable interactions across domains, with tunable hyperparameters such as **tree depth** and **number of estimators,** and **maximum number of features considered per split.** Finally, we implemented a **Probabilistic Classifier Chain (PCC)** using logistic regression as the base learner to jointly model the binary targets q26, q27, and q84\_binary, while accounting for their interdependencies. The PCC model builds a sequence of classifiers where each one conditions on previous labels, enabling accurate multi-label predictions and meaningful feature importance aggregation. For all three models—SLR, RF, and PCC—we extracted and visualized the top 10 most influential features to support interpretation. All models were trained on the same set of preprocessed features spanning behavioral, demographic, academic, health, and economic variables. By combining SLR, RF, and PCC, we balanced interpretability, flexibility, and multi-label structure to uncover robust and actionable risk factor profiles.

**SLR Mathematical Definition:**

with objective:  
  L(β) = −Σ [yᵢ log(pᵢ) + (1 − yᵢ) log(1 − pᵢ)] + λ Σ|βⱼ|

Where:  
   
λ controls the sparsity (L1 penalty strength)

**RF Mathematical Definition:**

Where:  
 – Number of trees  
 – prediction of the m-th decision tree  
 – indicator function

**PCC Mathematical Definition:**

Where:   
 – input features  
 – multiple binary labels (e.g., q26, q27, q84\_binary)  
Each conditional probability is modeled using logistic regression:

This yields a nested linear model structure, with one logistic regression per label that conditions on the input and all previous labels.  
In all three models, the learnable parameters capture the relationships between input features and target outcomes. For Sparse Logistic Regression (SLR), these parameters are the coefficient vector β\betaβ and the intercept β0\beta\_0β0​, which quantify the influence of each predictor on the binary outcome. Random Forest (RF) learns the structure of each decision tree—specifically, the feature splits and thresholds—along with the class proportions in each terminal leaf, which determine predicted probabilities. For the Probabilistic Classifier Chain (PCC), each binary target (e.g., q26, q27, q84\_binary) is modeled using a separate logistic regression, resulting in one coefficient vector per target. These classifiers are trained sequentially, with each one incorporating predictions from previous targets as additional inputs.

**Evaluation Modes and Metrics**

We evaluated model performance at both the individual prediction level and the population level. For individual data points, we used standard metrics: 0/1 loss (reported as accuracy), log loss, false positive rate (FPR), and false negative rate (FNR). These reflect predictive precision and are especially important under class imbalance. For population-level performance, we aimed to minimize expected loss (generalization error) over the distribution that generated the training data. This was approximated by test set performance on a held-out 30% split. To ensure robust estimates, we also conducted 20-fold cross-validation on the training set. In the multi-label setting (PCC), we used Hamming loss, exact match accuracy (joint 0/1 loss across targets), and multi-output log loss based on full target vector probabilities. We also reported per-target accuracy to enable fair comparison with single-output models. Confidence intervals were computed for key metrics using the standard error from cross-validation folds.

**Metrices Mathematical Definition:**

Where:  
 – Number of samples  
 – Number of labels (targets)

**Fitting Algorithms and Process**

We applied a systematic pipeline to prepare and fit data across all models. First, we filtered the dataset to include only students who provided binary responses to q26, q27, and q84\_binary (a new variable simplifying the original 5-level q84 where levels 4–5 denote risk). All non-numeric variables were dropped, and rows with missing predictor values were excluded.

Categorical variables were encoded using a hybrid approach. Binary variables with values {1,2} were recoded as {1→1, 2→0}, while multiclass variables were one-hot encoded (with drop\_first=True) to avoid multicollinearity. These transformations were applied across ~80 variables (vars\_to\_dummy) to produce a fully numeric feature matrix.

For each target, we performed a stratified 70/30 train-test split based on its label distribution. For PCC, a single split stratified on q26 was reused across all three targets to maintain consistency. Numeric features—including bmipct, gdp 2023, Unemployment Rate (Percent), and Mean household income (dollars)—were standardized using StandardScaler, fit on the training set and applied to both sets.

SLR minimized an L1-regularized logistic loss using the liblinear solver. RF models used class\_weight='balanced' and were tuned using randomized search (20 iterations, 5-fold CV). PCC was implemented using a custom logistic chain model, with each classifier conditioned on prior labels.

Each model used a different fitting procedure. **SLR minimized a convex L1-regularized logistic loss function**, which was solved using a convex optimization algorithm (liblinear solver). **RF applied greedy top-down induction of decision trees**, where each tree was trained on a bootstrap sample of the training set. At each split, the algorithm considered only a **random subset of features**, controlled by the max\_features='sqrt' setting. **PCC trained one logistic regression per target in sequence**, conditioning on the input features and previously predicted labels to capture interdependencies.

**Selection, Fitting, and Evaluation Protocol**

Each target model was trained and evaluated on its own stratified 70/30 train-test split, preserving class proportions to ensure stable performance estimates. Stratification was essential due to the imbalance in mental health indicators (e.g., more students answered “no” to sadness). This setup enabled separate hold-out evaluation and unbiased generalization estimates.

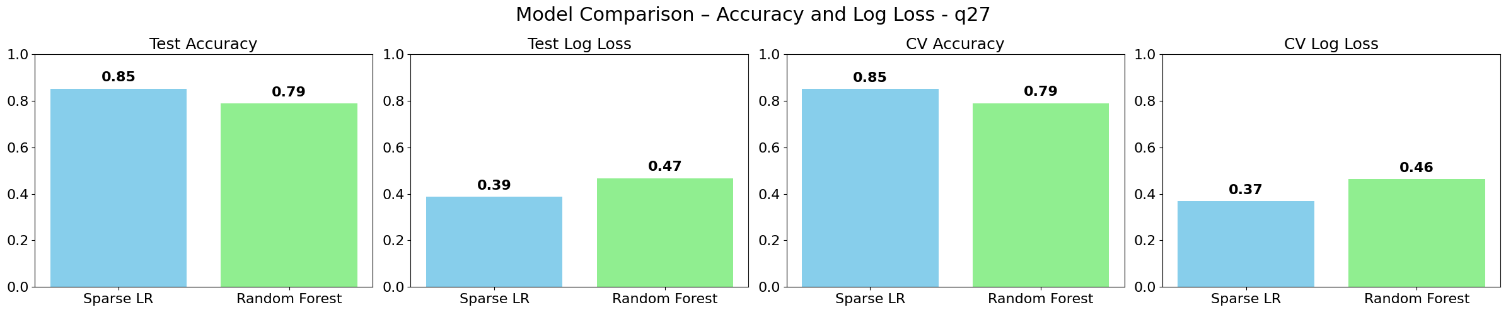
To mitigate class imbalance in RF, we used class-weighting during training to give higher importance to minority-class examples.

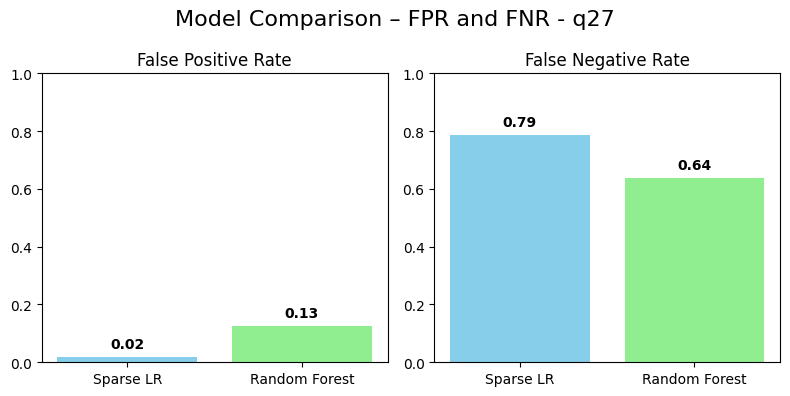
RF hyperparameters—such as number of estimators, maximum tree depth, and max\_features—were tuned using 5-fold cross-validation within the training set via a 20-iteration randomized search. No inner cross-validation was used for final evaluation: once tuning was completed, the best model was evaluated only once on the held-out test set, ensuring unbiased performance metrics. For SLR, we retained the default regularization strength (C=1.0), as L1 penalization already produced sparse and interpretable solutions. PCC was evaluated using the same train-test split as q26 and applied to all three labels.

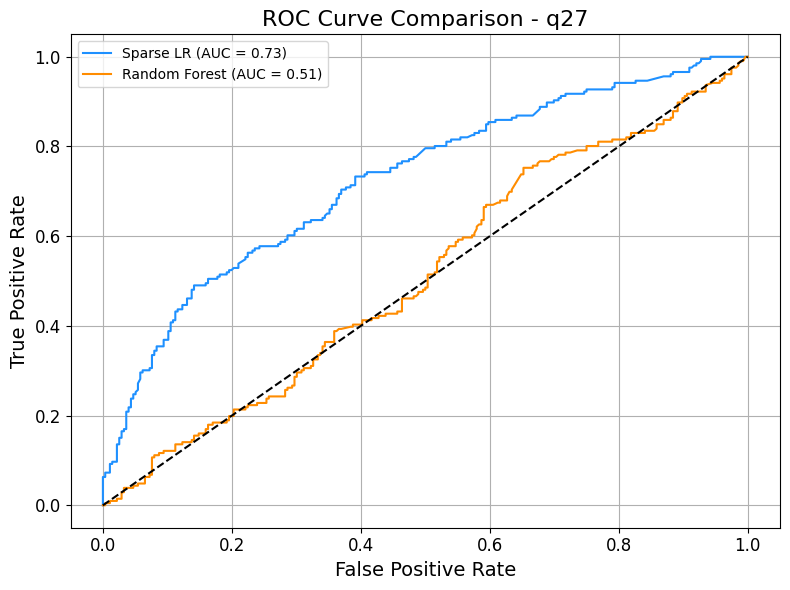
Training time per model was not explicitly constrained but remained within seconds to minutes due to moderate feature dimensionality.

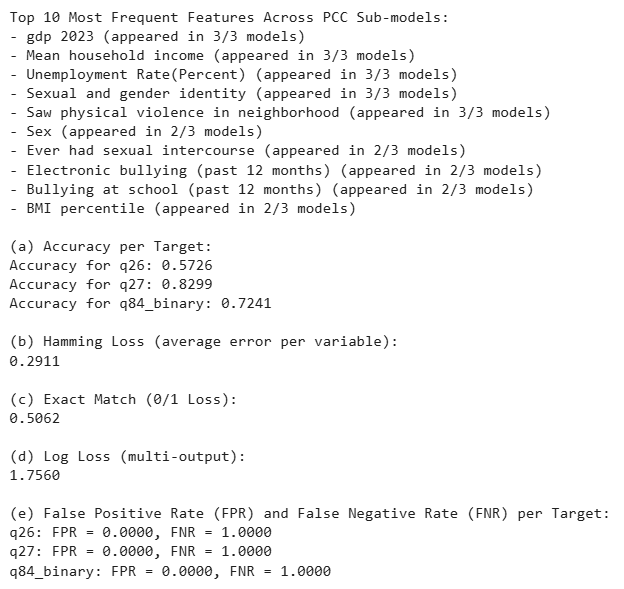
**Results**

**Model Evaluation**









We evaluated Sparse Logistic Regression (SLR), Random Forest (RF), and Probabilistic Classifier Chains (PCC) across three binary targets using cross-validation and test performance metrics with quantified uncertainty.

**SLR consistently outperformed RF** across most dimensions. It achieved the lowest log loss (e.g., 0.39 for q27), higher cross-validated accuracy (e.g., 0.85 for q27 vs. 0.79 in RF), and superior AUC scores (0.73 for all targets vs. 0.71, 0.51, and 0.56 in RF). However, **RF had lower False Negative Rates (FNR)** for q26 (0.34 vs. 0.42) and q84 (0.45 vs. 0.72), a meaningful trade-off since minimizing missed at-risk students is critical. SLR’s False Positive Rate was consistently lower, notably 0.02 for q27.

**PCC**, while conceptually strong in modeling interdependencies, struggled in practice — achieving 0% true positives across all targets (FNR = 1.0), making it unreliable in its current form. However, it had a reasonable Hamming Loss (0.2911), a 0.5062 exact match rate, and identified meaningful features repeated across sub-models. Its multi-output log loss (1.7560) remains informative for future improvement.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

**To assess critical errors**, we analyzed test cases with the highest log loss for q27. All were false negatives with high model confidence, suggesting overconfident misclassification. SHAP analysis revealed that these students had profiles typical of low-risk groups (e.g., cisgender males, no bullying exposure). This highlights blind spots in the model and suggests the need for additional variables capturing less visible or unmeasured risk factors.

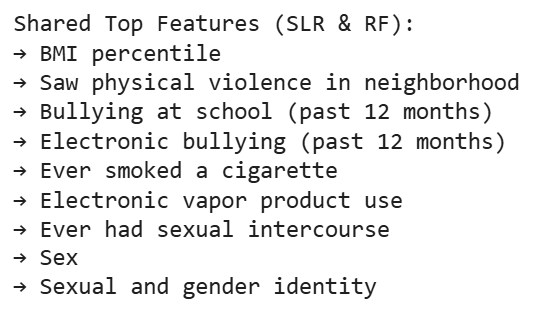
**Model Interpretation**

Across all three target variables, several predictors consistently emerged as influential in both Sparse Logistic Regression (SLR) and Random Forest (RF), offering robust interpretability. For example, **'Sex'** and **'Sexual and gender identity'** appeared as top predictors in all models. In SLR for q26, being female (positive coefficient) was strongly associated with increased reporting of persistent sadness (q26), while Sexual and gender identity had a negative coefficient, indicating that LGBTQ+students are at higher risk of experiencing sadness, suicidal thoughts, and hopelessness. This is a meaningful and expected finding that aligns with well-documented disparities in adolescent mental health outcomes.

**Electronic vapor product use**, **bullying exposure**, and **ever had sexual intercourse** also showed strong positive associations across targets, aligning with prior evidence linking risky behaviors and victimization to mental health risks. Socioeconomic factors like **income**, **GDP**, and **unemployment** were prominent in the PCC model but had smaller direct effects in SLR, suggesting they may influence outcomes indirectly or through interactions.

BMI percentile was a consistent but moderate contributor, while variables like **forced sexual intercourse** appeared uniquely important for q27 and q84, raising serious ethical and intervention concerns. Overall, model coefficients align with domain knowledge, though the RF model lacks directionality, reinforcing the value of SLR for hypothesis generation.

|  |  |
| --- | --- |
|  |  |



**Conclusion**

Our first research question—**how accurately mental health risk can be predicted**—can be partially answered. The Sparse Logistic Regression model achieved reasonable performance across all targets, with AUC values around 0.73, test accuracies up to 85%, and improved log loss compared to Random Forest. However, high false negative rates, especially for q27 and q84, highlight limitations in capturing all high-risk students. Thus, while our models are informative, they are not yet reliable enough for critical interventions.

The second research question—**how and which behavioral and socioeconomic factors are most predictive**—was answered more conclusively. Variables like sex, sexual and gender identity, electronic vapor product use, and bullying experiences consistently ranked among the most important predictors. Economic factors such as household income and unemployment rate showed weaker but still notable influence, often correlating with increased sadness or hopelessness.

To improve prediction, future work should consider more expressive models like probabilistic classifier chains with a carefully designed joint loss function. Incorporating temporal information or school-level context may also refine associations.

**Follow-up questions** could explore how protective factors like school connectedness or parental support buffer mental health risk, and whether predictions can be personalized across different student subgroups (e.g., LGBTQ+ or economically disadvantaged students). For example, how does school support mediate risk for LGBTQ+ students reporting bullying? Could linking behavioral data with clinical histories enhance precision? These directions would build on our findings and inform targeted prevention.