Title

BINF620 - Final Report

A Manning Smith

May 23, 2025

Contents

Abstract

AI Aknowledgement

This research utilized Claude 3.7 Sonnet, a large language model developed by Anthropic. Claude was prompted with the initial parameters of the research including the Abstract, Study Design, and Results. Claude assisted in adding visualization recommendations, and was used as a tool to debug code. As a powerful language model, Claude assisted in manuscript structuring and text synthesization, acting as an editor to written text and synthesized complex information into coherent text narratives; these narratives were not directly utilized but provided input into the written text via the author's own interpretation.

The author carefully reviewed and validated all AI contributions to ensure scientific accuracy. All analyses, interpretations, and conclusions represent the author's independent judgment and responsibility.

Background

Study Design

This study aimed to examine the relationships influencing mental health concerns among youth aged 6-17 years. The primary aim was to identify and quantify the interaction of individual, family, and social factors and the association with mental health concerns in children and adolescents, utilizing both traditional statistical approaches and advanced machine learning methodologies.

Aims

1. To determine the relative importance of different risk factors being: individual characteristics, family environment, social connections, adverse experiences, and neighborhood factors, in predicting mental health outcomes.
2. To evaluate the predictive performance of various analytical approaches, including logistic regression, regularized regression, ensemble methods, and principal component analysis-based models.
3. To examine interaction between various social factors being: after school activities, bullying, and neighborhood support in predicting mental health concerns.
4. To identify the most influential set of predictors that can effectively classify mental health risk.

Population Selection

Target Population

The study population consisted of children and adolescents aged 6-17 years participating in the 2022 NSCH. This age range was selected to capture the critical developmental period during which many mental health conditions first emerge and when social, academic, and family environmental factors become increasingly influential.

Sampleing

The NSCH employs a complex probability sampling design to produce nationally representative estimates. The survey utilizes a two-sections : - Section 1: Four questions about the presence of children in the home - Section 2: Detailed questions on the demographics and health of children

Survey results were taken as is and the following criteria was applied: - Children aged 6-17 years at the time of survey administration - Complete or sufficiently complete data on the primary outcome variable (mental health concerns)

Following data cleaning, the analytical dataset comprised over 20,000 observations, ensuring adequate statistical power for complex modeling approaches and subgroup analyses. The large sample size enabled robust estimation of interaction effects and supported the implementation of data-intensive machine learning algorithms while maintaining sufficient sample size for model validation through train-test splitting procedures.

Data and Materials

Data Source

This analysis utilized the 2022 National Survey of Children's Health (NSCH), conducted by the U.S. Census Bureau in partnership with the Health Resources and Services Administration's Maternal and Child Health Bureau. The NSCH is an annual cross-sectional survey designed to produce national and state-level estimates of child health and well-being across multiple domains.

Survey Design

The survey is primarily a mail-based data collection survey with online and telephone completion options. The survey domains, examined in this research include: - Child health - Healthcare access - Family dynamics and support systems - Neighborhood characteristics - School engagement - Adverse childhood experiences

Dataset

The final dataset was a subset from the `NSCH_2022e_Topical_CSV_CAHMI_DRCv2.csv`. The original data contained over 300 survey collection variables as well as over 200 calculated variables based on the provide codebook. In total, the codebook contains 579 variables. An alternate version of the unaltered dataset can be found on **census.gov**.

A smaller subset of this data was created for the purposes of this research and can be found in the project github available in the resources[#].

Dataset Documentation Comprehensive variable documentation was utilized to provide details regarding the data for each factor. The Methodology Report includes details such as: - Response categories and value labels - Skip patterns and logical consistency rules - Data collection methodology for each survey module

The complete data dictionary with variable definitions and descriptions was utilized via the **NSCH Codebook**. An amended version of the variable dictionary, comprised of all the variables used in the research analysis, can be found in 7**resources**.

Statistical Analysis

The study employed a comprehensive analytical approach combining traditional statistical methods with advanced machine learning techniques to examine predictors of mental health concerns among adolescents.

Reproducibility

Analysis code was version-controlled and documented to ensure reproducibility. A Random seed was set a `seed = 1776` for all stochastic procedures to enable exact replication of results.

Response Variable

The primary outcome variable was mental health concerns `MHealthConcern`, constructed as a binary indicator based on responses to survey questions regarding current mental health conditions. More specifically, the indicator to into account the variables `K2Q33B` and `K2Q32B`, indicating whether their child currently has depression or anxiety. If the response was "yes" to either condition, the resulting value was "yes".

Explanatory Variables (Risk Factors)

A total of 22 predictor variables were selected based on theoretical relevance, personal experience, and previous literature. There are 6 main factors that group the predictors: individual-level factors, social and family factors, neighborhood factors, adverse experience factors, parental mental health factors, and school engagement factors.

Individual-level factors:
- child's age `SC_AGE_YEARS`
- sex `sex_22`
- race/ethnicity `SC_RACE_R`
- physical activity level `PHYSACTIV`
- screen time usage `ScreenTime_22`
- age group category `age3_22`

Social and family environment factors:
- household composition `FAMILY_R`
- after-school activity participation `AftSchAct_22`
- event participation `EventPart_22`
- mentor availability `mentor_22`
- ability to share ideas with family `ShareIdeas_22`

Neighborhood factors:
- neighborhood safety `NbhdSafe_22`
- neighborhood support `NbhdSupp_22`
- community-level ACEs `ACE4ctCom_22`

Adverse experience factors:
- bullying `bully_22`
- victim of bullying `bullied_22`
- adverse childhood experiences count `ACEct11_22`
- discrimination experiences `ACE12`
- household adverse experiences `ACE6ctHH_22`

Parental mental health factors:
- mother's mental health status `MotherMH_22`

- father's mental health status `FatherMH_22`

School engagement factors:
- school connection measure `K8Q35`

Data Processing

Missing Data Imputation

Missing data were addressed using Multiple Imputation by Chained Equations (MICE) implemented in R. The imputation model utilized predictive mean matching (PMM) with 5 imputed datasets and 50 iterations to ensure convergence. Complete case analysis was performed on the imputed dataset to maintain statistical power while preserving the integrity of relationships between variables. This dataset was saved an utilized in python for PCA.

Data Scailing / Standardizing

Analytical Approaches

Univariate and bivariate analyses were conducted to examine the distribution of mental health concerns across predictor variables. Proportional differences were visualized using grouped bar charts and cross-tabulations to identify preliminary associations.

Statistical Models

Logistic Regression Multiple logistic regression models were fitted to examine associations between predictor variables and mental health outcomes. The six domain-specific models, described above, were developed to systematically examine different frameworks.

Regularized Logistic Regression To address potential multicollinearity and perform variable selection, elastic net regularization was implemented using the glmnet package. The optimal lambda parameter was selected through 5-fold cross-validation, with alpha = 0.5 to balance ridge and lasso penalties.

Interaction Effects Two-way interactions were examined between key variables, specifically: - After-school activities × sex - Physical activity × after-school activities

Machine Learning Approaches

Random Forest Classification

Random forest models were implemented using 500 trees with importance calculations enabled. The dataset was split into training (70%) and testing (30%) sets using stratified sampling to maintain outcome distribution balance. ### Gradient Boosting Machine GBM models were fitted with 500 trees, interaction depth of 3, and shrinkage parameter of 0.05. Optimal tree count was determined using out-of-bag error estimation.

Principal Component Analysis (PCA)

PCA was performed on standardized predictor variables to identify underlying latent constructs and reduce dimensionality while preserving maximum variance. The optimal number of components was determined using multiple criteria:

- Kaiser criterion (eigenvalues > 1)
- Cumulative variance explained (80% threshold)
- Scree plot examination for elbow identification

Componet Analysis

Principal component loadings were examined to understand the contribution of original variables to each component, with visualization through heatmaps to identify variable clustering patterns.

PCA Modeling

Logistic regression and random forest models were fitted using the selected principal components as predictors to compare performance with original variable models.

Model Evaluation

Performance Metrics

Model performance was evaluated using various metrics for binary classification:

- Area Under the ROC Curve (AUC): Primary metric for model discrimination ability
- Accuracy: Overall correct classification rate
- Sensitivity (Recall): True positive rate
- Specificity: True negative rate
- Precision: Positive predictive value

Model Comparison

Models were compared using Area Under the ROC Curve (AUC) as the primary performance metric, with additional evaluation using accuracy, sensitivity, specificity, and precision. ROC curve analysis was employed to assess model discrimination ability and compare performance across different analytical approaches.

Variable Importance

Analysis Environments

All analyses were conducted using R version 4.x and Python 3.x. Key R packages included: mice (imputation), glmnet (regularized regression), randomForest, gbm, caret (model training), and pROC (ROC analysis). Python analyses utilized scikit-learn, pandas, and numpy libraries.

Statistical significance was set at $= 0.05$ for all tests, with 95% confidence intervals reported for odds ratios and effect estimates.

The full list of packages and version can be found in the **supporting code and resources** section.

Results

Discussion

Conclusion

References

[[#]]. Fredricks, J. A., & Eccles, J. S. (2006). Is extracurricular participation associated with beneficial outcomes? Concurrent and longitudinal relations. Developmental Psychology, 42(4), 698–713. https://doi.org/10.1037/0012-1649.42.4.698

[[#]]. Loades, M. E., Chatburn, E., Higson-Sweeney, N., Reynolds, S., Shafran, R., Brigden, A., Linney, C., McManus, M. N., Borwick, C., & Crawley, E. (2020). Rapid Systematic Review: The Impact of Social Isolation and Loneliness on the Mental Health of Children and Adolescents in the Context of COVID-19. Journal of the American Academy of Child and Adolescent Psychiatry, 59(11), 1218–1239.e3. https://doi.org/10.1016/j.jaac.2020.05.009

[[#]].

Supportting Code & Resources

Github Project

The complete project results and supporting code can be found **linked here at github**.

Data Files

Scripts

The end to end script can be **referenced here at the github project**.

R Packages and Versions

- RColorBrewer | Version 1.1.3
- ggplot2 | Version 3.5.2
- readxl | Version 1.4.5
- tibble | Version 3.2.1
- dplyr | Version 1.1.4
- tidyverse | Version 2.0.0
- writexl | Version 1.5.4
- knitr | Version 1.50
- png | Version 0.1-8
- tinytex | Version 0.57
- imager | Version 1.0.3
- bookdown | Version 0.43
- ROCR | Version 1.0-11
- randomForest | Version 4.7-1.2
- gridExtra | Version 2.3
- caret | Version 7.0-1
- mlbench | Version 2.1-6
- kableExtra | Version 1.4.0
- neuralnet | Version 1.44.2
- naivebayes | Version 1.0.0
- tidyr | Version 1.3.1
- fastDummies | Version 1.7.5
- mice | Version 3.17.0
- corrplot | version 0.95
- car | Version 3.1-3
- glmnet | version 4.1-8
- MASS | version 7.3-65
- pROC | version 1.18.5
- gbm | Version 2.2.2

Python Packages

The complete list of required python packages can be installed with the same version via the `requirements.txt` in the data folder.

AI Acknowledgement

Tables & Figures