**Project Proposal:** Predicting Mental Health Outcomes Using Behavioral and Demographic Data

**Student’s Name:** Abdullah Siddiqui

**Advisor:** Dr. Sridhar Ramachandran

## Kaggle Competitions Investigated

### Competition 1

**Competition Name:** Titanic: Machine Learning from Disaster

**Competition URL**: <https://www.kaggle.com/c/titanic>

**Description:** This competition is a classic binary classification problem where participants predict the passenger survival based on structured data such as age, gender, ticket class, and family size. I chose to investigate it because it is widely used as an introduction to predictive modeling, feature engineering, and ensemble methods.

**Learnings:** From the leaderboard and public kernels, I learned that feature engineering was critical. For example, extracting titles from passenger names (e.g., “Mr.”, “Miss”) or grouping ticket classes improved model accuracy. Top competitors relied heavily on tree-based algorithms like Random Forests, XGBoost, and ensemble stacking methods, showing the value of combining simple engineered features with robust models.

### Competition 2

**Competition Name:** House Prices: Advanced Regression Techniques

**Competition URL**: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

**Description:** The objective was to predict home sale prices using over 70 explanatory variables describing homes. I chose this competition because it represents a regression challenge that required careful handling of both numeric and categorical variables, much like many health datasets.

**Learnings:** Winning solutions highlighted feature preprocessing: (handling missing values, encoding categorical features), log-transforming skewed variables, and the advanced ensemble methods like LightGBM and CatBoost. Top participants also used feature importance analysis to interpret which variables mattered most, an approach I plan to apply to my thesis dataset to identify which health and lifestyle factors best predict mental health outcomes.

### Competition 3

**Competition Name:** Predict Student Performance (Knowledge Tracing)

**Competition URL**: <https://www.kaggle.com/competitions/riiid-test-answer-prediction>

**Description:** This competition focused on predicting whether a student would answer a future question correctly, based on their historical learning data. It was relevant to my interests in psychology and behavior prediction.

**Learnings:** Top solutions used deep learning models, including RNNs and Transformers, to capture temporal sequences in behavior. I learned that modeling not just the static features but also the sequences of actions can significantly improve predictive performance. This also tells me how I might use temporal health behavior data in my chosen dataset.

[If you investigated more than 3 competitions, feel free to add them here]

## Chosen Dataset

**Dataset Name:** Behavioral Risk Factor Surveillance System (BRFSS)

**Source:** Centers for Disease Control and Prevention (CDC)

**Source Link**: <https://www.cdc.gov/brfss>

**Potential Insights:** I selected this dataset because it is one of the largest, still active collected health-related surveys in the world. It contains information on health behaviors, chronic health conditions, mental health indicators, and demographic factors. I am particularly interested in exploring the relationships between lifestyle factors such as sleep, exercise, diet, smoking, and alcohol use, and reported mental health outcomes for example number of poor mental health days. This dataset provides a great opportunity to apply predictive modeling and feature engineering techniques learned from Kaggle competitions to a real-world public health problem.

## Project Scope

**Objectives:** What are the goals of the project? Connecting the project to real-world need.

* Predict mental health outcomes for example poor mental health days, using behavioral and demographic data.
* Identify which lifestyle factors contribute most strongly to mental health.
* Provide interpretable insights that could inform public health interventions.

**Deliverables:** What will be produced by the project? For example, Charts, graphs, Stats, etc.

* Cleaned and preprocessed dataset.
* Predictive models: baseline logistic regression, tree-based models, ensemble methods).
* Feature importance analysis and visualizations.
* Final report including statistical findings and interpretive insights.

**Milestones:** What are the key dates and important deadlines? You will use these for your timeline. For example, when will the data analysis start and end. When will you cross check your observations, etc.

* Week 2: Dataset exploration and cleaning.
* Week 4: Baseline models developed.
* Week 6: Advanced models tuned and evaluated.
* Week 8: Feature importance and interpretability analysis.
* Week 10: Final report and presentation.

**Tasks:** What tasks need to be completed? Think in terms of Steps …so this section can have a list of steps to complete the project.

1. Acquire and explore the dataset.
2. Clean and preprocess variables.
3. Engineer new features based on behavioral patterns.
4. Train baseline models.
5. Apply advanced Kaggle-inspired methods: XGBoost, LightGBM.
6. Evaluate results using appropriate metrics.
7. Visualize insights and prepare final report.

**Resources:** What resources will be needed? For example, R libraries to be installed, dedicated computer if your project needs one, etc.

* Python and R: pandas, scikit-learn, LightGBM, XGBoost.
* Kaggle kernels for reference.
* CDC BRFSS documentation.

## Research Plan

**Techniques and Methods:**

* Python and R: pandas, scikit-learn, LightGBM, XGBoost.
* Kaggle kernels for reference.
* CDC BRFSS documentation.

**Application:**

These techniques will be applied to the BRFSS dataset to predict mental health outcomes and determine which behaviors most significantly impact psychological well-being. For example, I may try deriving features like combined lifestyle risk scores and analyze their predictive power.

**Hypothesis**

**Hypothesis Statement:**

Behavioral and lifestyle factors for example sleep quality, exercise frequency, alcohol use, smoking, combined with demographics can accurately predict mental health outcomes, with exercise and sleep emerging as the strongest predictors of reported poor mental health days.

## Research Milestone Timeline

Plan out your project to be completed in 10 weeks and envision what will be completed each week in some detail below. Note: submission date is the due date for this assignment and not the date you turn it in. If you turn this document before the due date, you should still plan your work by the submission date - you will simply have more time to accomplish your weekly plan.

**Week 1 from submission date of this document:**

Tasks to be completed this week: Familiarize with BRFSS dataset, review Kaggle solutions.

A description of what you plan to achieve by the end of this week.

**Week 2 from submission date of this document:**

Tasks to be completed this week: Data cleaning and variable selection.

A description of what you plan to achieve by the end of this week.

**Week 3 from submission date of this document:**

Tasks to be completed this week: Feature engineering and exploratory data analysis.

A description of what you plan to achieve by the end of this week.

**Week 4 from submission date of this document:**

Tasks to be completed this week: Develop baseline models (logistic regression, decision trees).

A description of what you plan to achieve by the end of this week.

**Week 5 from submission date of this document:**

Tasks to be completed this week: Implement gradient boosting models (XGBoost, LightGBM).

A description of what you plan to achieve by the end of this week.

**Week 6 from submission date of this document:**

Tasks to be completed this week: Hyperparameter tuning and cross-validation.

A description of what you plan to achieve by the end of this week.

**Week 7 from submission date of this document:**

Tasks to be completed this week: Interpretability analysis (SHAP, feature importance).

A description of what you plan to achieve by the end of this week.

**Week 8 from submission date of this document:**

Tasks to be completed this week: Draft initial findings, create visualizations.

A description of what you plan to achieve by the end of this week.

**Week 9 from submission date of this document:**

Tasks to be completed this week: Validate results, refine models, finalize insights.

A description of what you plan to achieve by the end of this week.

**Week 10 from submission date of this document:**

Tasks to be completed this week: Complete final report and presentation submission.

A description of **how your project will be complete by the end of this week**.

## Potential Scope Creep

**What type of Scope creep** can you envision. FYI: Scope creep refers to changes, continuous or uncontrolled growth in a project’s scope, at any point after the project begins. This can occur when the scope of a project is not properly defined, documented, or controlled. It’s important to manage and control the project scope to prevent scope creep, which can lead to projects taking longer to complete or going over budget.

* Expanding the research question beyond mental health into all health outcomes in BRFSS might be too broad.
* Attempting to use deep learning models without sufficient computational resources.
* Adding too many engineered features that complicate interpretability.  
  To manage this, I will focus only on mental health outcomes and restrict models to those that balance predictive accuracy with interpretability.

## Ethics Statement

As a student, I pledge to sincerely hold myself to the timeline and complete the project in 10 weeks. I understand the importance of time management and commitment to my work. I will strive to uphold the highest standards of academic integrity throughout this project.