# PROJECT OVERVIEW

**Problem Statement:** This project explores how lifestyle habits influence cognitive performance in humans. Using regression models, the goal is to predict cognitive scores based on factors like sleep, diet, exercise, and substance use. The analysis aims to identify which habits have the strongest impact on cognitive health.

**Dataset Summary:** The dataset contains responses from 80,000 individuals and includes a wide range of variables related to lifestyle habits and cognitive performance. Key features include:

- **Demographics**: Age, gender, education level
- **Lifestyle Factors**: Sleep duration, physical activity, diet, smoking, alcohol use, caffeine intake, screen time
- **Cognitive Measures**: Scores from cognitive performance assessments

The dataset is self-reported and intended for exploratory and predictive analysis, allowing for examination of how daily behaviors may correlate with or predict cognitive ability.

## Tools Used:

- **Python** – for data manipulation, analysis, and modeling
- **VS Code** – as the primary development environment
- **Pandas & NumPy** – for data cleaning and transformation
- **Matplotlib & Seaborn** – for exploratory data visualization
- **Scikit-learn** – for building and evaluating regression models
- **Jupyter Notebook** – for documenting the analysis process
- **Tableau** – for creating the final interactive dashboard and visual summaries
- **SHAP** – For model explainability and feature importance analysis

## Steps Taken So Far:

### Data Cleaning

The dataset was examined for common quality issues such as duplicate records and missing values; however, none were found, which allowed for a smoother analysis process. I then focused on identifying and correcting typos or inconsistencies in categorical variables to ensure clean and uniform labeling—for example, standardizing responses like “Yes” and “yes.” This step was essential for avoiding redundancy and improving the accuracy of groupings during analysis and modeling.

### Initial Feature Engineering

To enhance model performance, I performed initial feature engineering by creating new variables that could better capture underlying patterns in the data. This included combining related lifestyle factors and generating simplified or binned versions of continuous variables. I also applied transformations such as scaling or normalization where necessary to prepare the dataset for regression modeling. During this process, I began evaluating which engineered features were most likely to contribute meaningfully to prediction accuracy.

### Exploratory Data Analysis (EDA)

Initial exploratory data analysis (EDA) was conducted to understand the key patterns and distributions within the dataset. Individual variables, such as sleep hours and screen time, were visualized to explore their central tendencies and spread. Pairwise correlations between features were also examined to identify relationships between lifestyle factors and cognitive scores. This initial analysis provided valuable insights and helped to identify which factors might have the strongest influence on cognitive performance.

### Modeling (Planned/Next Steps)

Next steps include training predictive models using both basic linear regression and more complex ensemble methods like random forests. The goal is to evaluate model performance using metrics like $R^2$ and RMSE. Cross-validation will be used to ensure that the models are not overfitting the data. Additionally, interpretability techniques like SHAP will be applied to understand how individual features contribute to predictions.

### Intervention Simulation (Planned/Next Steps)

Finally, the predictive model will be used to simulate hypothetical lifestyle interventions, such as increasing sleep duration or reducing screen time, to estimate their potential impact on cognitive performance. This will help understand how changes in behavior could lead to measurable improvements in cognitive health.