Mona Albarqi , Albatool Moathen

Project Proposal

High-Level Problem Statement: The goal of this project is to analyze factors influencing individual happiness and develop a predictive model to estimate happiness levels based on socio-demographic and lifestyle variables. This study aims to answer the question: *What are the key predictors of happiness, and how accurately can happiness levels be predicted using these predictors?*

Research Question:

- What are the primary factors influencing individual happiness?
- How well can happiness levels be predicted using demographic and behavioral data?

Background:

Happiness has been a subject of study across psychology, sociology, and economics. Previous studies have highlighted factors such as income, social relationships, health, and employment status as significant contributors. However, the interplay of these variables often varies across cultures and demographic groups. While much is known about general trends, the extent to which these predictors explain individual happiness remains uncertain.

Outcome Variable:

HAPPINESS

- **Description:** This categorical variable indicates the individual's self-reported happiness level (e.g., "Very Happy," "Pretty Happy," "Not Too Happy").
- **Summary Statistics:**
    - Proportion of each category will be calculated (e.g., frequency and percentage distribution).
    - Visualized using a bar chart to display the distribution of happiness levels.

Predictor Variables:

The dataset contains a rich set of predictors, of which the following will be included in the model:

1. **AGE**: Continuous variable (years).
2. **INCOME**: Continuous variable (annual income in currency).
3. **EDUCATION**: Discrete variable (years of education).
4. **MARITAL_STATUS**: Categorical variable (e.g., Single, Married, Divorced).
5. **GENDER**: Categorical variable (Male, Female).
6. **EMPLOYMENT_STATUS**: Categorical variable (e.g., Employed, Retired, Student).
7. **HEALTH**: Categorical variable (e.g., Excellent, Good, Poor).
8. **RELIGION**: Categorical variable (e.g., Religious, Not Religious).

9. **SOCIAL_CONTACT**: Discrete variable (number of social interactions per week).
10. **POLITICAL_VIEWS**: Categorical variable (e.g., Liberal, Moderate, Conservative).

Data Sources: The dataset is synthetic but modeled after real-world surveys such as the General Social Survey (GSS). Data will be analyzed for missingness, imputed where necessary, and verified for consistency.

Definition of Success:

A successful project will:

1. Identify significant predictors of happiness.
2. Develop a predictive model with performance metrics (e.g., accuracy, F1-score) above a baseline threshold.
3. Provide interpretable insights using variable importance and partial dependence plots.
4. Present findings in a clear, reproducible manner, with results corroborated by visualization.

Next Steps:

1. Data preprocessing: Handle missing values and perform exploratory data analysis (EDA).
2. Model development: Experiment with classification algorithms (e.g., Logistic Regression, Random Forest).
3. Evaluation: Assess model performance using cross-validation.
4. Visualization: Create plots and charts to explain key findings and illustrate model performance.
5. Reporting: Document methods, results, and conclusions in a Jupyter Notebook.
6. Presentation: Develop slides to summarize the problem, methods, results, and challenges.
7. Repository Organization: Set up a GitHub repository with clear documentation, data, and code structure.

This proposal outlines the foundational steps for the project, aiming to balance rigor with interpretability to deliver actionable insights into the determinants of happiness.