# Capstone Assignment 20.1: Exploratory Data Analysis and Modeling on Happiness

## Project Structure and Summary

This project explores the key factors associated with self-reported happiness using survey data from a World Value Survey(https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp) dataset. It includes:

* A Jupyter notebook with comments[Notebook] (https://github.com/msach05/Module20\_CapstoneEDA/blob/main/CapstoneProjectAnalysis-InitialReportAndDataAnalysis.ipynb)
* The datafile is included in the Data Folder
* A README document summarizing key findings (notebook link included).

## Data Cleaning and Preprocessing

1. **Missing Values:**
   * Replaced common missing value codes (-1 to -5) with NaN.
   * Dropped columns with over 20% missing values.
   * Filled remaining missing values using mode imputation.
   * Dropped any rows with remaining NaNs to ensure data consistency.
2. **Column Filtering:**
   * Selected only valid “Q” survey question columns- these were the Questions asked in survey.
   * Converted object types to numeric safely.
   * Dropped original columns if a \_reversed version was created. This was needed to ensure that all survey values are in the same order as Hapiness (increasing order)
3. **Feature Engineering:**
   * Created reversed versions of specific survey questions to align all scores positively with happiness.
   * Dropped original non-reversed columns after transformation.

## Exploratory Data Analysis (EDA)

### Correlation Analysis

* Calculated Pearson correlation of all features with Q46\_reversed (Hapiness Score - reversed happiness score to ensure higher values = higher happiness)).
* Top correlated features include:
  + Life Satisfaction (0.44)
  + Self-Reported Health (Reversed) (0.37)
  + Financial Satisfaction (0.34)
  + Freedom of Choice (0.24)
  + Feeling of Security (0.19)

### Visualizations

* **Heatmap** of top 30 features correlated with happiness.

## **Boxplots** and **Histograms** for each top correlated feature vs. happiness score.

## Modeling

### Linear Regression

* **R² Score**: 0.317
* **MSE**: 0.337
* Top Features:
  + Life Satisfaction,
  + Self-Reported Health,
  + Freedom of Choice,
  + National Pride

### Decision Tree Classifier

* **Accuracy**: 63.6%
* Performance varied across happiness levels with good recall for moderately happy class.

### Logistic Regression

* **Accuracy**: 76.4%
* Excellent precision for “Happy” class, but recall was lower for “Unhappy.”
* Influential features:
  + Self-Reported Health,
  + Financial Satisfaction,
  + Freedom of Choice
  + Marital Status
  + Feeling of security

### K-Nearest Neighbors (KNN)

* **Accuracy**: 86.6%
* Strong performance on happy class, lower recall on unhappy group.

### Ensemble Model (XGBoost + KNN, Soft Voting)

* **Accuracy**: 87.7%
* Best overall performance.
* Balanced precision and recall between happy and unhappy classes.

## Summary on Religious and Spiritual Influence on Happiness

Despite the original hypothesis of this project — that religious beliefs, spirituality, or engagement in religious practices might positively impact happiness — the current models **do not indicate a strong or consistent positive correlation between religious/spiritual factors** and happiness. Instead, health, financial satisfaction, and personal freedom were stronger indicators for happiness. (compared to the hypothesis of religion or spiritual practices)

Variables analyzed included: - Importance of God in Life - Religious Self-Identification - Attendance at Religious Services - Frequency of Prayer - Religious Membership

None of these variables emerged as top predictors of happiness in the current correlation or regression-based analyses.

## Future Directions

To further explore the influence of religious or spiritual practices, the next phase of the analysis will:

* **Group related variables** into domains like Finance, Health, Security, Social Trust, etc.
* **Control for or stabilize** these other strong factors.
* Then re-evaluate whether **religious/spiritual practices show a clearer impact** on happiness once confounding variables are accounted for.

This will include segmenting the population to analyze if spiritual or religious practices have any impact.