**DATA FRIDAY – AUGUST 2024**

**MENTAL HEALTH ANALYSIS**

**PRESENTED BY MAJOIE NGANDI**

**Comprehensive Analysis Plan for Mental Health Dataset**

**Introduction**

This analysis aims to uncover patterns and insights from the mental health dataset. The objective is to understand the demographics of affected populations, investigate potential reasons behind observed trends, and provide actionable insights for improving mental health interventions. The dataset was obtained from reliable sources through surveys and is to be evaluated.

**Questions I intend to answer**

1. **Percentage per Country of Affected Population:** What percentage of the population is affected by mental health issues(growing stress) in different countries?
2. **Student vs. Corporate Mental Health Issues:** What is the percentage of students versus corporate employees facing mental health issues? If students are more affected, what are the potential reasons?
3. **Impact of Family History:** Is having a family history of growing stress a significant factor in its prevalence?
4. **Frequency of Mental Health Issues Based on Activity:** How often do people who stay at home versus those who go out face such stress?
5. **Effectiveness of Care Options:** Are the available care options effective in managing mental health issues?
6. **Occupations with Growing Stress:** Which occupations among American men and women are showing growing stress levels?

**Technologies and Tools**

1. **Pandas and NumPy:** For data manipulation and statistical calculations.
2. **Plotly Dash:** For interactive and detailed data visualization.
3. **Scikit-learn:** For machine learning model development and evaluation.
4. **Matplotlib and Seaborn:** For static visualizations and exploratory data analysis.
5. **Jupyter Notebook:** For organizing and running the analysis interactively.

**Understanding Features:**

* **Timestamp**: A record of a particular date and time. This could be when an observation or data point was recorded regarding someone's mental health.
* **Gender**: The classification of a person as male or female.
* **Country**: The place where a person lives. This is relevant because mental health resources and cultural attitudes toward mental health can vary by country.
* **Occupation**: A person's usual or principal work or business. Occupation can influence stress levels and access to mental health resources.
* **Self-employed**: Indicates whether a person is working for themselves or has their own business, rather than working for an employer.
* **Family history**: A record of the relationships among family members along with their medical histories. Family history of mental health issues can indicate potential genetic or environmental factors.
* **Treatment**: Indicates whether a person is currently undergoing treatment for mental health issues.
* **Growing Stress**: Indicates whether a person's stress level is increasing over time.
* **Changes Habits**: This seems incomplete. It could refer to changes in behavior or habits that may indicate shifts in mental health.
* **Mental\_Health\_History**: A record of a person's mental health history, including any previous diagnoses or treatments.
* **Mood\_Swings**: Fluctuations in a person's mood, which can be indicative of various mental health conditions.
* **Coping\_Struggles**: Difficulties a person may have in coping with stressors or mental health challenges.
* **Work\_Interest**: Level of interest or engagement in work or activities, which can be affected by mental health.
* **Social\_Weakness**: Weaknesses or difficulties in social interactions or relationships, which can be impacted by mental health issues.
* **Mental\_Health\_Interview**: It seems like this could refer to a structured interview or assessment related to mental health.
* **Care\_Options**: Options available for seeking care or treatment for mental health issues, such as therapy, medication, or support groups.

**Detailed Analysis Steps**

**Step 1: Data Cleaning**

1. **Load the Dataset:** Begin by loading the dataset into a Pandas DataFrame.
2. **Remove Null Values:** Identify columns with missing values and decide on a strategy (e.g., removing rows, imputing with mean/median/mode).
3. **Remove Duplicates:** Ensure that no duplicate records are present in the dataset.
4. **Standardize Data Formats:** Ensure consistency in data formats, such as converting all text to lowercase, standardizing date formats, and categorizing numerical data appropriately.

**Step 2: Exploratory Data Analysis (EDA)**

1. **Descriptive Statistics:** Calculate mean, median, mode, standard deviation, and other statistical measures for numerical features.
2. **Distribution Plots:** Use histograms and box plots to visualize the distribution of key variables (e.g., age, stress levels).
3. **Correlation Analysis:** Use correlation matrices and heatmaps to identify relationships between different features.
4. **Categorical Data Analysis:** Analyze the distribution of categorical data such as gender, occupation, and country.

**Step 3: Visualization**

1. **Percentage per Country:** Calculate and visualize the percentage of the affected population by country.
2. **Student vs. Corporate Analysis:** Compare the prevalence of mental health issues between students and corporate employees. Investigate potential reasons for higher prevalence among students.
3. **Family History Impact:** Analyze the impact of family history on mental health issues using bar charts and histograms.
4. **Indoor vs. Outdoor Activity:** Examine the frequency of mental health issues among individuals who stay at home versus those who go out using comparative bar charts.
5. **Effectiveness of Care Options:** Assess the effectiveness of different care options available by visualizing treatment success rates.

**Step 4: Advanced Analysis**

1. **Occupation Analysis:** Identify occupations with the highest stress levels and mental health issues using bar charts and pie charts.
2. **Gender Differences:** Investigate differences in mental health issues between genders using comparative bar charts.
3. **Trend Analysis:** Analyze trends over time to see if mental health issues are increasing using line charts.

**Step 5: Machine Learning Model Development**

1. **Data Preparation:** Split the dataset into training and testing sets, ensuring a balanced distribution of classes.
2. **Feature Engineering:** Convert categorical variables into numerical values using techniques like one-hot encoding.
3. **Model Selection:** Choose appropriate machine learning algorithms (e.g., Logistic Regression, Random Forest, Support Vector Machines).
4. **Model Training:** Train the model using the training dataset and tune hyperparameters for optimal performance.
5. **Model Evaluation:** Evaluate model performance using metrics such as accuracy, precision, recall, and F1 score. Aim for at least 90% accuracy.
6. **Confusion Matrix:** Build and interpret the confusion matrix to understand model performance in detail.

**Step 6: Reporting and Documentation**

1. **Document Findings:** Compile all findings, visualizations, and insights into a comprehensive report.
2. **Visualization Dashboards:** Create interactive dashboards using Plotly Dash to allow stakeholders to explore the data and insights.
3. **Recommendations:** Provide actionable recommendations based on the analysis, such as targeted mental health interventions for specific demographics.
4. **Presentation:** Present findings on Linkedin network on Data Friday.