1. **Data information**The dataset contains 12 columns and 1,000 entries, with attributes:

* Numerical attributes: Age, Sleep\_Hours, Work\_Hours, Physical\_Activity\_Hours
* Categorical attributes: Gender, Occupation, Country, Mental\_Health\_Condition, Severity, Consultation\_History, Stress\_Level
* **The goal of predicting mental health (binary) based on people’s occupation**

**+ Important features:** Occupation, Stress\_Level, Work\_Hours, Sleep\_Hours, Consultation\_History, Severity.

Additional Considerations

* **Occupation (Categorical)**: Since your target is to predict mental health for people in specific occupations, this is a key variable. Different occupations can have varying stress levels, working conditions, and demands, which can influence mental health.
* **Stress\_Level (Categorical)**: Stress is often directly linked to mental health issues, so this attribute is likely highly relevant.
* **Work\_Hours (Numerical)**: Excessive work hours may correlate with mental health issues, especially in high-demand occupations. This feature could help capture the intensity of the work environment.
* **Sleep\_Hours (Numerical)**: Sleep has a well-documented impact on mental health, and lack of sleep could indicate work-life imbalance or stress related to specific occupations.
* **Consultation\_History (Categorical)**: Knowing if someone has previously sought mental health consultation could indicate ongoing or past mental health issues, adding context to their current mental health condition.
* **Gender (Categorical)**: Gender can influence mental health due to various social, psychological, and workplace dynamics.
* **Physical\_Activity\_Hours (Numerical)**: Physical activity is often associated with better mental health, so this provides supportive information about general well-being.
* **Age (Numerical)**: While not directly tied to occupation, age can still influence stress and work experience levels. Mental health patterns can also vary across different age groups.
* **Country (Categorical)**: Cultural and economic contexts differ by country, and these factors can affect mental health and the experience of work in different occupations.
* **Severity (Categorical)**: This would provide insight into the intensity of any current mental health symptoms, making it a valuable predictor.

1. **Data preprocessing**
2. **Handling Missing Values**

* Check for null data

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động

Severity column has 501 missing value (Null) over 1000   
501/ 1000 \* 100% = 50.1%  
=> There is a large portion of missing data (501 rows in your case), dropping them may lead to data loss and could introduce bias

* Given that Mental\_Health\_Condition has a distribution of Yes (267) and No (234) for the missing Severity, it might be worth analyzing

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phông chữ

Mô tả được tạo tự động

So, we see the data is missing randomly so we replace missing values with the **mode** of the Severity column for the corresponding Mental\_Health\_Condition category

1. **Remove duplicated Rows** **Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, Đồ họa

   Mô tả được tạo tự động**
2. **Outlier Detection and Removal**

We check outliers of numerical attributes by applying IQR: Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, thực đơn

Mô tả được tạo tự động

As we can see data has no outliers. Scatter plot matrix Ảnh có chứa ảnh chụp màn hình, Hình chữ nhật, hình vuông, Song song

Mô tả được tạo tự động

1. **Encoding Categorical Variables**

Encoded Mental\_Health\_Condition and Consultation\_History: **Yes:1 ; No: 0**

Encoded Severity: **'None': 0, 'Low': 1, 'Medium': 2, 'High': 3**

Encoded Stress\_Level: **'Low': 0, 'Medium': 1, 'High': 2**

Encoded Gender: **'Non-binary':0, 'Prefer not to say':1,'Male':2,'Female':3**

Encoded Occupation: **'IT':0, 'Finance':1,'Healthcare':2,'Education':3,'Engineering':4,'Sales':5,'Other':6**

Encoded Country: **'USA':0,'India':1,'UK':2,'Canada':3,'Australia':4,'Germany':5,'Other':6**

**Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động** **Ảnh có chứa văn bản, ảnh chụp màn hình

Mô tả được tạo tự động**

1. **Normallization/Scaling**

If the dataset has features with different scales like "Age" (ranging from 20 to 60), the model may give more importance to features with larger numerical values. Normalization/scaling transforms the features so that they are on a similar scale, making sure no single feature dominates the others due to its larger scale.

We are scaling the numerical featuresẢnh có chứa văn bản, Phông chữ, ảnh chụp màn hình

Mô tả được tạo tự động

Plot before scaling: Ảnh có chứa tòa nhà chọc trời, tòa nhà, cửa sổ, thành phố

Mô tả được tạo tự động

Plot after scaling: Ảnh có chứa tòa nhà chọc trời, hàng, cửa sổ, tòa nhà

Mô tả được tạo tự động

1. **Correlation checking**

Checking correlations will help ensure that we are using the most relevant, non-redundant features for predicting the target variable **Mental\_Health\_Condition**

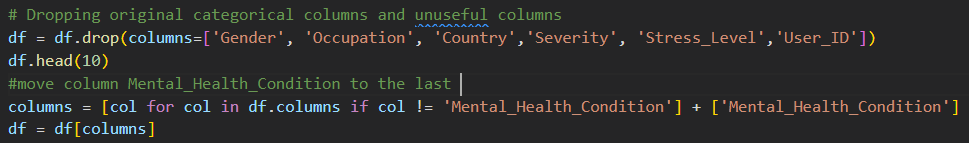
**Correlation Matrix:**

**Ảnh có chứa ảnh chụp màn hình, hình vuông, Hình chữ nhật, Nhiều màu sắc

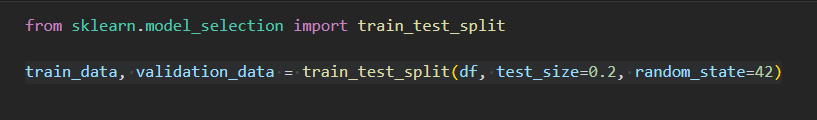
Mô tả được tạo tự động**

1. **Cleaning dataset and convert to ARFF**

* Droping original data columns and unuseful columns and nove Mental\_Health\_Condition column to the last

****

* Spliting data:
* **Validation\_data:** 20% of the original data
* **train\_data:** 80% of the original data



* Convert to ARFF file

**Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động**

* Save cleaned data to csv file: 