Use Case: Predictive Model for Mental Health Risk

Participants will develop a predictive model using traditional Machine Learning techniques to assess mental health risk factors (such as anxiety, depression, or stress), based on structured and unstructured data. The goal is to improve early detection and provide actionable insights for mental health professionals.

1. Context

Mental health disorders are a growing concern worldwide, impacting productivity, well-being. And healthcare. Traditional detection methods rely on static questionnaires and clinical interviews, which are time-consuming and often fail to identify early warning signs.

This challenge aims to create a data-driven predictive system using classic ML algorithms to analyze multiple data sources and predict risk levels.

2. Objectives

- Data Preparation & Quality Assurance:
  You will need to prepare the data in order to not get missing values
  - Clean and normalize categorical survey responses
  - Handle missing values (e.g., "N/A", "I don't know")
  - Encode categorical and ordinal variables appropriately
- Feature Engineering & Selection
  Assign numerical values to those fields that are relevant to the following indexes/scores, and provide some kind of visualization to them such as a heatmap.
  - **Mental Health Support Index**
    (benefits, resources, anonymity, formal communication)
    - Provide the 5 pair of fields that have the most correlation between them
  - **Workplace Stigma Index**
    (fear of negative consequences, observed discrimination)
    - Provide the 5 pair of fields that have the most correlation between them
  - **Organizational Openness Score**
    (comfort discussing mental health with managers and peers)
    - Provide the 5 pair of fields that have the most correlation between them

    - After mapping all the values you seem relevant to each index you will need to save all of them in a data frame for the modeling and clustering questions.

- Modeling
  Make a and train a model for each of the following target

    - Current presence of a mental health condition
        - Identify the 10 fields that have most correlation with "Do you currently have a mental health disorder?"
        - Train the model with "Do you currently have a mental health disorder?" as a target using only the 10 fields selected

    - Likelihood of seeking professional treatment.
        - Identify the 10 fields that have most correlation with "Have you ever sought treatment for a mental health issue from a mental health professional?"
        - Train the model with "Have you ever sought treatment for a mental health issue from a mental health professional?" as a target using only the 10 fields selected

- Clustering(Worker Profiling)

    - Apply clustering techniques to identify distinct employee profiles
    - Decide or evaluate which clustering techniques are best for identifying three different groups among the employees.
    - You can support your cluster feature selection with the features previously mapped.
    - Provide the top 3 values for each cluster.

- Interpretation & Insights

    - Feature importance analysis

    - Clear explanation of model behavior

    - Visualizations understandable by non-technical audiences

3. Dataset

An anonymous survey of professionals working in the technology industry.

Dataset includes:

- Mental health status (current, past, diagnosed, treated)

- Impact on productivity and work performance
- Company policies and mental health benefits
- Workplace culture, stigma, and openness
- Demographic and job-related information

4. Deliverables

1. Reproducible code or notebook
    - Cleaned data
    - Feature engineering
    - 2 models
    - Clustering
2. Excel submissions with the results
    - Explained in the last point of this word.
3. Results document (max.8 pages), including:
    - Methodology overview
      Explain each step taken and showcase the results gotten.
    - Key findings
      Showcase the results and explain what they represent.
    - Visualizations
      Add graphs, heatmaps or mappings for the results gotten in order to support and represent the results.
    - Business or organizational recommendations
      Explain in general the results obtained and what do they indicate and what measures can be taken in order to identify potential health issues or prevent them in a work environment.

5. Evaluation

- Feature engineering 30%
- Model interpretability 50%
- Clustering 20%

6. Expected outcome

By the end of the hackathon, teams should deliver:

- An **explainable ML model**
- A **ranked list of key workplace factors**

- Clearly identified **risk profiles**
- Concrete, actionable recommendations to improve workplace mental health

7. How to submit the excel

The final output will be displayed on an excel, results must be presented in the submission_output_template.xlsx, assessing each phase, following this format:

Feature engineering

- For the feature solution you will need to provide 5 correlations for each of the indexes.
  ex:
  The first correlation of mental health support would be "*features mental health support 1a*" and "*features mental health support 1b*".
  The third correlation workplace stigma index would be "**features workplace stigma 3a**" and "*features workplace stigma 3b*"

ML training

- For the ML solution you will need to provide the top 10 correlations for each model and the f1 score you got from it.
  ex:
  For the first model, which uses as a target "*Do you currently have a mental health disorder?*" in the excel you have "*Do you currently have a mental health disorder? corr 1*" to "*Do you currently have a mental health disorder? corr 10*" and finally for the f1 score you have "*Do you currently have a mental health disorder? f1 score*" it follows the same format for the second ML model based on the target column "*Have you ever sought treatment for a mental health issue from a mental health professional?*".

Clustering

- Lastly, to indicate the top three variables with most relevance for each cluster the structure is as follows.
  ex:
  the first cluster is "*cluster 0*" and their top 3 correlations are indicated as follows, "*cluster 0 1*", "*cluster 0 2*", "*cluster 0 3*", the second cluster would be "*cluster 1*" and the third "*cluster 2*" having structured their top 3 correlations similarly to "*cluster 0*"