Francesco Scolz

Attimis, 30/08/2025

Youth Stress Predictor

Machine Learning for a healthier future

Problem to solve

Creating a **predictive model** that can reliably **classify stress types** among teens (aged 14 - 21) by a series of **questions** about their life

Evaluation metric will be **Macro F1-score**, which ensures that the model's performance is evaluated equally for every class, even in imbalance situations

The **minimum score requirement** for this project will be **0.8 (80%)**

Why solve this problem

Stress among new generations is becoming an **increasing problem** over time

This **problematic** reflects on a lot of areas in life such as **school**, **sports** and even **relationships**

- With **No / Minimal stress**, people tend to be **undermotivated** and **less productive**
- **Eustress** (the “good” stress) is the **sweet spot**: motivation without ending in worse
- **Distress** (the “bad” stress) is when large amounts of this create **anxiety** and **fears**, completely overriding the potential benefits

By building a **preliminary model** that can screen potential patients, we can aim to **cut health expenses** on **useless medical visits**, while still providing a solid prevention base

⚠️ DISCLAIMER ⚠️

This is just a **probabilistic model**, and it **does not aim to substitute professional help**

Do not rely on these predictions, treating them as always true

The **information** is provided **“as is”**, and it has **no legal or medical value**

The dataset

We used an **open-source dataset** from **Kaggle**, which can be found at this link

It contains data from a **survey**, ran on **all-aged individuals**, with answers on **questions** like *“Have you recently experienced stress in your life?”* or *“Do you have trouble concentrating on your academic tasks?”*

The answers, except for gender and age, are in a range from **1 (strongly disagree)** to **5 (strongly agree)**

Explorative analysis

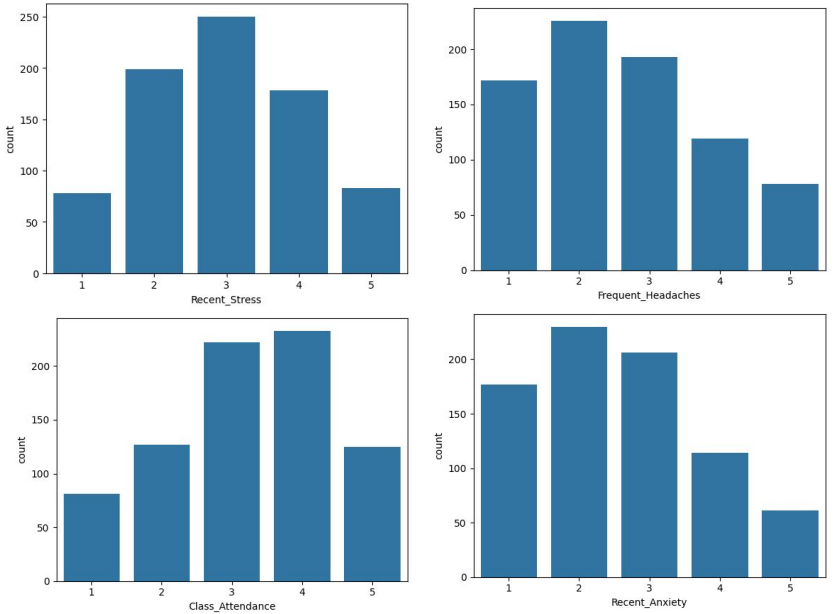
Firstly, we checked for **missing / impossible values**, which **weren't** found, so we moved to the next step

We **transformed** some columns to be **readable** from **our models** by One-Hot Encoding non-numerical values

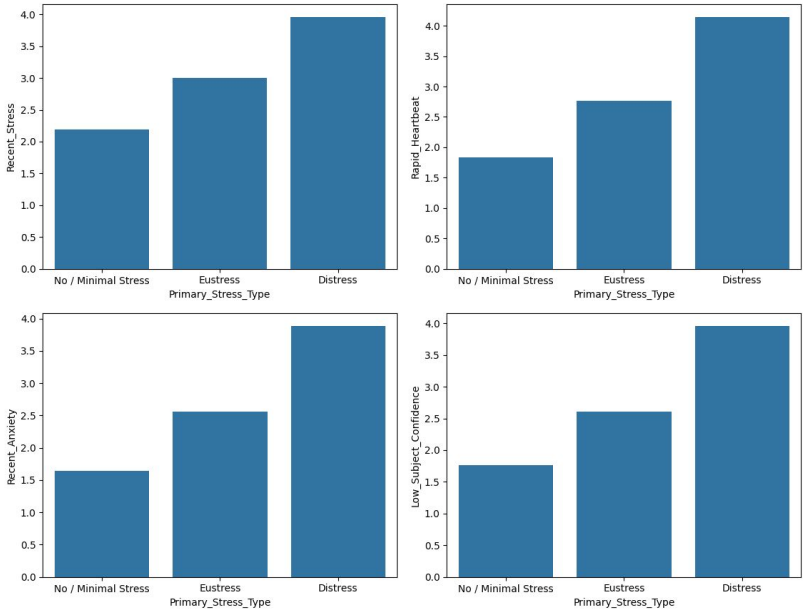
After **preprocessing** the data, we conducted an **explorative analysis** on data distributions and correlations with the target column (Primary_Stress_Type)

Explorative analysis

Data distributions contained a lot of bell curves, sometimes shifted to some side (known as skewed curves), suggesting fair and honest answers through the survey



Explorative analysis



Feature correlations with the target columns, on the other hand, seemed to have a strong linear correlation. This means that as values grow, stress intensity rises too (No stress → Eustress → Distress)

With this knowledge, we can start creating fine-tuned models to our problem

Modeling

Different models were tested, but ultimately a **simple Logistic Regression** achieved the best performance, with a **Macro F1-score** of **85.38%.**

This indicates that across **all stress classes** (None, Eustress, Distress), the model achieved an **average F1-score** of **over 85%.**

The reason a **simple model** can **outperform** more **sophisticated alternatives** likely lies in the **data structure**: linear correlations are better captured by linear models such as Logistic Regression.

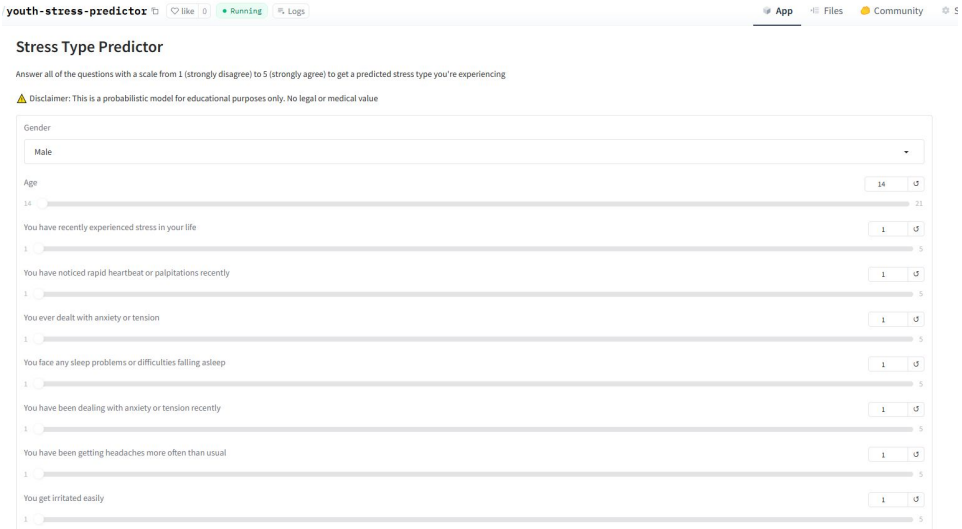
Without a **thorough exploratory data analysis**, these insights **could not have been uncovered**.

(for a full breakthrough of all the models tested, please refer to the *models_tested.xlsx* file)

Testing and demo

The model can be currently found in this **GitHub Repository** as **.pkl** file (for external implementations)

It can also be used **online as demo** on the **Hugging Face platform** at this link 👉



Conclusions

With a **Macro F1-score** of **85.38%**, this model provides a **solid baseline** for **further research**. However, limitations remain, such as the **small sample size** and **limited model testing**

Looking forward, models of this type could be deployed in **healthcare settings** to reduce **staff workload** and **associated costs**, while **continuing to be trained** alongside **medical-approved professionals**, ensuring a high-quality training experience.

Conclusions

This project **aims** to **demonstrate** what can be **accomplished with technology**, even **without** a dedicated budget. Just a **laptop**, an **internet connection**, and a **lot of dedication**

For more projects like this, feel free to **explore** my GitHub and Linkedin pages, which are full of **interesting work**

Thank you for your attention

Scolz F.