Week 6 Interim Project Report

Project Title

How Work–Life and Lifestyle Conditions Shape Burnout, A Cross-National Machine Learning Analysis

Team

Anoushka, Biraj, Hunter, Samridhi
Northwestern University, Class 422, Practical Machine Learning
Date: August 6, 2025

Progress Summary

Data Loading and Preparation

We began by loading and cleaning the primary dataset, which contains approximately 50,000 records with individual-level data on demographics, work hours, sleep, physical activity, occupation, country, lifestyle habits, and self-reported stress levels. Missing data analysis revealed that almost all variables were complete, except for “Severity,” which has approximately 50 percent missingness. For all further analysis, we excluded columns and rows with excessive missingn values or standardized values as needed.
All data cleaning, exploratory analysis, and visualization for this interim report were conducted using Python in Google Colab. The full EDA notebook is attached and provides additional plots, code, and summary statistics supporting the findings reported here.

Target Variable Definition

To operationalize burnout for modeling, we define a binary Burnout variable. Burnout equals one for individuals reporting “High” Stress_Level, and zero for those with “Low” or “Medium” Stress_Level. This approach is consistent with psychological research linking sustained high stress to clinical burnout. In our dataset, about one third of respondents fall into the burnout (high stress) category, and two thirds do not.

Burnout class balance:

- Not burned out (Low or Medium): 33,293
- Burned out (High): 16,707

A bar plot of the Burnout variable confirms moderate class imbalance. Both classes are large enough for robust machine learning modeling.

Exploratory Data Analysis (EDA)

We analyzed distributions and bivariate relationships for all major features.

- Sleep hours, Most respondents report six to eight hours, with no major difference between high and low burnout groups.
- Work hours, Median work week is 55 hours across groups, with no strong link to burnout in simple analysis.
- Physical activity and social media use, Both show similar distributions across burnout groups, with no meaningful separation.
- Age, The sample skews toward ages 21 to 40. No age group stands out for higher burnout rates.
- Gender, Burnout is similarly distributed among all gender groups.
- Country and Occupation, Stacked bar plots show that high stress, as our burnout definition, is consistently present in about one third of respondents for the ten most represented countries and job types. There are no significant differences in burnout rates by country or occupation.

Key Findings from EDA

No single demographic or lifestyle variable, including sleep, work hours, age, gender, physical activity, or occupation, shows strong separation between low and high burnout groups in bivariate analysis. The distribution of burnout is nearly uniform across countries and job types. This suggests that burnout in this population is a multifactorial issue, not driven by isolated features.

Next Steps and Modeling Approach

Given that bivariate analysis does not reveal strong predictors, we will focus on multivariate machine learning models that can account for interactions and combinations of factors. Our next steps include,

- Encoding categorical variables, binning where appropriate, and scaling numerical features.
- Fitting logistic regression, random forest, and XGBoost models to predict burnout.
- Using feature importance and model interpretation, including SHAP values, to identify the main drivers and interaction effects contributing to burnout risk.
- Evaluating model performance using F1 score, precision, recall, and ROC-AUC.
- If necessary, addressing any class imbalance with resampling or weighting methods.

Conclusion

Our interim analysis confirms that burnout, operationalized as high stress, is a widespread issue in this international sample. No single demographic or lifestyle predictor stands out in simple group comparisons. Moving forward, advanced modeling will be essential to uncover the complex interplay of factors leading to burnout. These findings will guide our final project work and the development of actionable recommendations for workplace and lifestyle interventions.

Reference

Full Exploratory Data Analysis and code are available in the attached Google Colab notebook.